# Does Positive Reinforcement Work?: A Quasi-Experimental Study of the Effects of Positive Feedback on Reddit

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

Koustuv Saha
ksaha2@illinois.edu
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

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

## Abstract

Social media platform design often incorporates explicit signals of positive feedback. Some moderators provide positive feedback with the goal of positive reinforcement, but are often unsure of their ability to actually influence user behavior. Despite its widespread use and theory touting positive feedback as crucial for user motivation, its effect on recipients is relatively unknown. This paper examines how positive feedback impacts Reddit users and evaluates its differential effects to understand who benefits most from receiving positive feedback. Through a causal inference study of 11M posts across 4 months, we find that users who received positive feedback made more frequent (2% per day) and higher quality (57% higher score; 2% fewer removals per day) posts compared to a set of matched control users. Our findings highlight the need for platforms, communities, and moderators to expand their perspective on moderation and complement punitive approaches with positive reinforcement strategies to foster desirable behavior online.

## CCS Concepts

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

## Keywords

Online Communities; Online Moderation; Feedback Mechanisms; Causal Inference

**ACM Reference Format:**
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 *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3706598.3713830

## 1 Introduction

"We heard you... awards are back!"– Reddit Product Team[1]

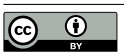
In September 2023, Reddit removed several positive feedback mechanisms, including awards and Reddit gold, from its interface[2] in an effort to create better ways to empower communities to reward content. Although *gilding*—i.e., the ability to "gild" posts by donating a Reddit Premium subscription—is no longer a feature on Reddit, the platform brought back other forms of awards in May 2024[1] based on feedback from its users. These recent changes reflect both the platform's interest in supporting positive feedback and the users' desire to continue providing such feedback.

Positive feedback is a key feature of positive reinforcement, a long-studied behavior modification principle in the field of behavioral psychology shown to be effective in offline settings to encourage desired behavior through positive stimuli [30]. In the online context, recent HCI research has revealed that some moderators of online communities consider positive reinforcement a part of their role [55]. In particular, the system of rewards utilized in positive reinforcement involves the use of the available signals on the platform. Importantly, most social media platforms include some form of positive feedback (e.g., Reddit upvotes, Facebook likes, etc.). In fact, the presence of positive feedback as an affordance is associated with a community consisting of higher-quality content [70]. Furthermore, HCI researchers have established that positive feedback increases a user's motivation to contribute to a community [54]. As a result, positive feedback may be vital both for users to want to participate and for platforms that rely on contributions to be successful. Despite the ubiquity of positive feedback across platforms and the established theories motivating its potential to have meaningful impacts offline and online, there is a lack of empirical evidence supporting the effectiveness of receiving positive feedback on improving user-level outcomes. Learning whether positive feedback can measurably change important outcomes can alter our perspective on moderation and motivate the need for platforms to support proactive approaches to moderation centered around reinforcement in conjunction with preexisting punitive approaches.

We ground our work in the taxonomy of moderation developed by Grimmelmann [34]. The taxonomy highlights openness through community participation in moderation as an important goal that encourages democracy. Grimmelmann [34] also mentions two key moderation actions, organization and norm-setting, to help shape the content in feeds and create shared norms between community-members, respectively.

Distributed moderation has been theorized to be effective [34] and has been found effective at enforcing norms in practice [60]. While Reddit is ostensibly moderated by centralized moderation teams, the use of widespread positive feedback inherently creates a democratic system of distributed moderation that enables users to voice opinions on what content should be encouraged, aligning with the moderation goal of openness [34]. We need to understand

[1] https://reddit.com/r/reddit/comments/1css0ws/we_heard_you_awards_are_back/
[2] https://reddit.com/r/reddit/comments/14ytp7s/reworking_awarding_changes_to_awards_coins_and/




*CHI '25, Yokohama, Japan*

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713830

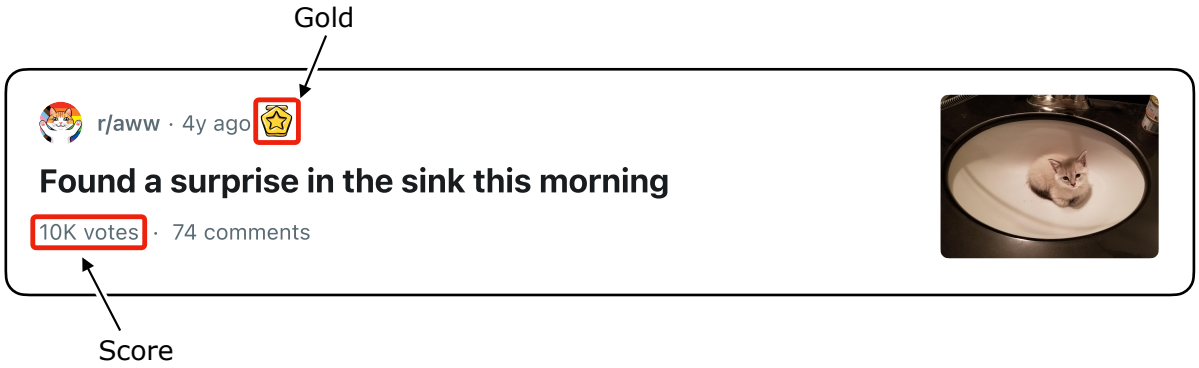



**Figure 1: This is an example post from r/aww that received gold, represented by a small gold icon above the title. The post's score is reported under its title. This post's score was in the top 25th percentile for r/aww in the month it was posted.**

the cause-and-effect relationships between these positive-feedback-based distributed moderation practices and the other goals of moderation (i.e., reorganizing communities and facilitating norm-setting) to inform design solutions that can leverage the power of positive feedback to promote healthier communities.

To explore distributed moderation through positive feedback, we focus on Reddit gold and upvotes, two forms of positive feedback available to all community members. Figure 1 provides an example post that received both forms of feedback. Gold was a beloved [103] but sparsely-used, paid feature while upvotes are free to give out and used in all sub-communities on Reddit. Prior research confirms that moderators utilize both gold and upvotes as mechanisms for positive reinforcement [55], however both forms of feedback are given anonymously, so we cannot evaluate their efficacy as moderator-specific feedback and instead explore whether they are capable of accomplishing the goals of moderation when given by the community more generally (i.e., distributed).

## 1.1 Research Questions

The posting quality, posting frequency, and norm-adherence of users are important to communities [107], platforms [54], and moderators [55] alike. Thus, we ask whether positive reinforcement through distributed moderation is capable of influencing these important user-level outcomes. More specifically, we pose the following research questions:

**RQ1:** What is the effect of positive feedback on recipients?
- (a) How does receiving gold or being highly upvoted impact users' future behavior in a community?
- (b) How long do the effects of positive feedback on user behavior last?

**RQ2:** How do the effects of positive feedback vary across different types of recipients?
- (a) What types of users experience stronger effects of receiving positive feedback?
- (b) What types of communities experience stronger average effects of receiving positive feedback?

## 1.2 Summary of Contributions

*1.2.1 Methods.* With these research questions in mind, we employ causal inference methods to estimate the causal effects of receiving positive feedback (i.e., our treatments). In particular, we draw on the potential outcomes framework [39, 80] to conduct stratified matching followed by difference-in-difference analyses. To do so, we first form a group of treatment posts which received a large quantity of upvotes and/or received gold. We then form a control group of similar posts from the same communities that did not receive such forms of positive feedback. With these two groups, we perform stratified matching to match users with similar propensities of being treated. Then, we conduct a difference-in-differences analysis to understand the effect of receiving positive feedback on a user's future behavior in their community and compare pre- and post-treatment behavior to understand how long these effects persist.

We then explore the differential effects of our treatments through regressions on the individual treatment effects to identify which types of users and communities experience stronger effects of receiving positive feedback.

*1.2.2 Findings.* We find that users who receive gold or high score on a post receive more positive feedback on future posts compared to if they had never received the positive feedback. Highly-upvoted users also experience an increase in their posting rate and a decrease in their rate of removals. Additionally, we observe that users who received stronger treatment (i.e., more gold or upvotes) experience larger increases in future positive feedback received. Importantly, Reddit newcomers experience larger increases in posting frequency and larger decreases in removal rates after being highly upvoted.

*1.2.3 Implications.* This work contributes to HCI research by quantifying the power user interface elements have on influencing user behavior. Specifically, some researchers have developed their own systems to encourage certain user behaviors [20, 38, 109] and others have explored the effect of existing interface features, such as badges [5] and reputation systems [3, 77], on users. We build on this work by highlighting upvotes, a signal nearly universally available in some form across social media platforms, and by gaining a deeper understanding of how users across platforms may benefit from the use of positive feedback. We also advance moderation research

by expanding our understanding of how community-driven moderation can function in conjunction with positive reinforcement strategies. Based on our findings, we make design recommendations for platforms and HCI researchers to facilitate the use of positive feedback signals for encouraging desirable behavior and easing norm acquisition. Furthermore, we identify specific populations (e.g., newcomers, users who have not recently experienced positive feedback) that may gain the most from receiving positive feedback. We also highlight the benefits of using positive feedback for platforms and moderators and call on designers to create explicit interface features to support this form of proactive moderation. Finally, we highlight ways moderators can work with the current system of positive feedback to guide users towards norm-abiding and high-quality content.

## 2 Background

This section provides an overview of relevant background literature related to behavioral psychology, distributed moderation theory, the effect of interface signals, and gift-giving in online communities.

### 2.1 Principles in Behavioral Psychology

Early research in behavioral psychology involved exploring the effect of positive consequences on encouraging behavior, firstly through Thorndike's law of effect [33]. This principle states that people repeat behavior with satisfying consequences and avoid behavior with unpleasant consequences. Thorndike's principles have been studied in experimental setups [8], but subsequent research building on the law of effect has been more thoroughly validated.

More specifically, building on the law of effect, B. F. Skinner developed the idea of positive and negative reinforcement and punishment in his work on operant conditioning [30, 92–94]. Prior moderation research has often focused on the use of punitive techniques such as content removals, bans, and quarantines, but our work centers around *positive reinforcement*, the introduction of a positive stimulus to encourage behavior, to evaluate whether it can encourage users online to repeat the behavior that led to the positive outcome. The use of positive reinforcement has been validated in offline settings like workplaces [7, 13, 73, 104], education [6, 12, 62], parenting [25, 97], and athletic training [29, 105, 108].

One specific type of reinforcing feedback that has been studied is praise. Kazdin [47] find that praising desired behavior increases the likelihood that the behavior will be repeated. On Reddit, praise happens through upvoting and awarding signals. Prior work found that some moderators specifically provide positive feedback like gold and upvotes to reinforce behavior and others specifically aimed to praise the author [55].

Additional research has identified another type of reinforcement: vicarious. The idea of vicarious reinforcement is that bystanders of positive reinforcement may be vicariously influenced to repeat the behavior being reinforced [10, 46]. Jhaver et al. [42] explore the effect of exposure to post-removal explanations on bystanders, essentially vicarious punishment. Though they found that the exposure did not encourage bystanders to learn the norms of the community, prior work in in social computing shows that vicarious reinforcement through highlighting examples of norm-adhering behavior in a community may help with further norm-adherence [50]. Seering et al. [91] showed that users on Twitch imitated observed behaviors, especially from high status users. Furthermore, some users believe providing positive feedback can increase a post's visibility in the feed [55], effectively highlighting the content for the rest of the community. This exposure to posts approved of by the community may enable community norm acquisition by newcomers and other bystanders through vicarious reinforcement.

Miltenberger [66] further explored the factors that may impact the effectiveness of positive reinforcement strategies in practice and presents three notable factors: contingency, satiation, and immediacy. Contingency states that a positive stimulus will have the strongest effect when it is only presented in response to the behavior being reinforced and at no other time. Satiation is the idea that a positive stimulus will have a diminished effect on the recipient if they have been too exposed to the stimulus. Wang and Diakopoulos [102] demonstrate the idea of satiation in their work which reveals that receiving a New York Times Pick badge, a form of positive reinforcement, has the largest effect on first-time recipients' future comment quality and diminishing effects for more frequent recipients. Finally, the concept of immediacy indicates that the reinforcing stimulus will be more effective when received more quickly. Choi et al. [22] explored the idea of immediacy in the context of YouTube creator hearts and found that videos received more engagement when creator hearts are given sooner after the video was published.

*Our work builds on the theory of operant conditioning and prior work demonstrating the effectiveness of positive reinforcement in offline settings to understand community-driven moderation on Reddit.*

### 2.2 Distributed Moderation

Reddit's moderation system exhibits qualities of both centralized and distributed moderation systems. There are some platform-level moderation structures (e.g., admins[3]), but sub-communities create their own sets of rules [75] and most moderation falls to volunteer moderation teams within each community. While these volunteers themselves are Reddit users, they still form centralized moderation structures within communities. These moderation teams are crucial to user safety, but the non-moderating users play a similarly important role in moderation. Reddit enables users to give various forms of feedback to contributions, such as upvotes, downvotes, and awards. These signals are then taken into consideration for the sorting of community feeds. In this way, users have the collective power to highlight content they approve of and hide content they do not. In practice, this is an example of distributed moderation, a system through which members of a community are responsible for making moderation decisions.

There is extensive research on Reddit moderation, including identifying abusive behavior [4, 9, 19, 36, 98], evaluating the efficacy of moderation strategies [16, 17, 41, 65, 95], and characterizing moderator experiences [49, 88]. However, prior work does not often explore the power of Reddit's distributed moderation through user feedback, though some Reddit moderators themselves have stated that users should utilize upvotes and downvotes on the platform to self-regulate, demonstrating support for distributed moderation from active participants in Reddit's centralized moderation [55].

[3] https://redditinc.com/policies/content-policy

Prior work into distributed moderation often focuses on the platform Slashdot. Lampe and Johnston [58] specifically highlight the effect of positive and negative feedback given to an author's first contribution on their future participation on Slashdot. Additionally, Lampe and Resnick [59] and Lampe et al. [60] show the power distributed moderation on Slashdot has to identify high and low quality contributions, while also reporting some of the challenges associated with the approach. Jiang et al. [43] and Grimmelmann [34] similarly identify trade-offs between various moderation practices, including comparisons of centralized and distributed moderation. Jiang et al. [43] deem distributed moderation a more democratic process as it better reflects community-wide opinions.

In this work, we build on prior work presenting the advantages of distributed moderation and demonstrating the support for such an approach, even from members of centralized moderation teams, to understand its effect in practice. *We focus specifically on forms of positive feedback to advance our understanding of whether distributed moderation on Reddit is an effective form of community-driven positive reinforcement.*

## 2.3 Guiding User Behavior Through Interface Signals

Platform designers have the power to determine what users can do in their communities by choosing to include or exclude various interface elements. For example, some platforms utilize badges [5], gamification mechanisms [37, 71, 87], or reputation systems [3, 77, 78], which have all been shown to encourage participation. Prior work has also shown that users can be encouraged to make higher-quality contributions through the use of specific highlighting mechanisms [102]. Similarly, posts on Reddit receive one upvote by default, an instance of example-setting that may encourage other users to follow suit. Upvotes and other explicit signals of positive feedback on Reddit are examples of formal feedback, which may be more effective at encouraging norm awareness and abidance than informal forms of feedback [50]. Finally, other research has shown that prominently displaying community rules and guidelines encourages adherence, especially from newcomers [65]. This practice is used across Reddit and Twitch.[4]

Social computing researchers have developed new tools and signals on platforms to encourage certain behaviors. Im et al. [38], for example, introduced a signal to the Twitter interface to inform users about an account's history with toxicity and misinformation. Chang et al. [20] introduce an intervention for Reddit users engaging with inflammatory content to encourage more reflection about whether they should contribute to uncivil conversations. Similarly, Yen et al. [109] build a tool to visualize the amount of negativity in live-streaming chatrooms to encourage viewers to proactively moderate the space by contributing prosocially.

Specific user behaviors can also be encouraged through more high-level considerations. For example, platform designers determine whether users are able to remain anonymous in their contributions which can influence users' willingness to contribute certain types of content. More specifically, prior work has found that anonymity leads to more sensitive self-disclosures [24, 63, 72] and more thoughtful feedback [23]. However, in some communities, this may come at the cost of lower-quality contributions [69] or more disruptions [26].

Similarly, platforms can consider social translucence [27, 31], the concept of emphasizing the visibility of social information in platform design, when making decisions. For example, X (formerly Twitter) has recently changed the visibility of users' liked posts so others can no longer see who liked another user's post, nor can users see the posts other users have liked.[5] Similarly, Instagram now provides the option to make the like count on posts private.[6] Ostensibly these changes are motivated by user privacy or maintaining user well-being, but there is disagreement in prior work regarding whether like counts on Instagram are responsible for negative effects on user well-being [101]. Additionally, there are negative side effects of these adjustments on social translucence because they mask vital network information in a way that may hinder interactions between users.

However, certain interface features can enable users to influence their communities. Reddit upvotes, for example, are a widespread mechanism used across the platform and are controlled entirely by Reddit users. There has been prior research in understanding the role of upvotes on Reddit [56, 67], including in spaces like advice-giving [61] and political communities [14, 70]. Prior work has shown that interfaces that allow upvotes typically have higher-quality content [70]. Additionally, Carman et al. [14] manipulated the number of upvotes given to posts in political and non-political subreddits and found that posts with artificially-inflated upvote counts experienced increases in their final upvote and reply counts, demonstrating the power of upvoting to increase engagement and positive feedback. *We extend this research to understand the specific influence positive feedback has on the recipients themselves.*

## 2.4 Gift-Giving and Reciprocity

Social computing literature highlights gift-giving as a mechanism that builds a sense of reciprocity from gift-givers. Essentially, users expect to receive similar treatment from their communities as they give [53]. Other work shows that receiving a gift encourages users to give similar gifts in the future [52], highlighting reciprocity for the recipient. Importantly, gift-giving is a means to build reputation which social computing literature touts as a clear signal of desired behaviors [57], an incentive for good behavior [77], and a tool to build trust and engagement among users [78].

In Reddit communities, positive feedback mechanisms are staples of gift-giving, and if they abide by the principles of reciprocity, recipients of positive feedback may be motivated to provide similar feedback in the future. This highlights the importance of understanding user-level outcomes of gift-giving through positive feedback. *Our work explores the user-level outcomes in response gift-giving in the form of Reddit gold and upvotes.*

# 3 Study Design and Data

This section describes our data-collection process and the causal-inference approach we adopted drawing on the potential outcomes framework [79, 80] used in prior social computing research [17,

[4] https://safety.twitch.tv/s/article/Chat-Tools

[5] https://x.com/XEng/status/1800634371906380067

[6] https://about.instagram.com/blog/announcements/giving-people-more-control

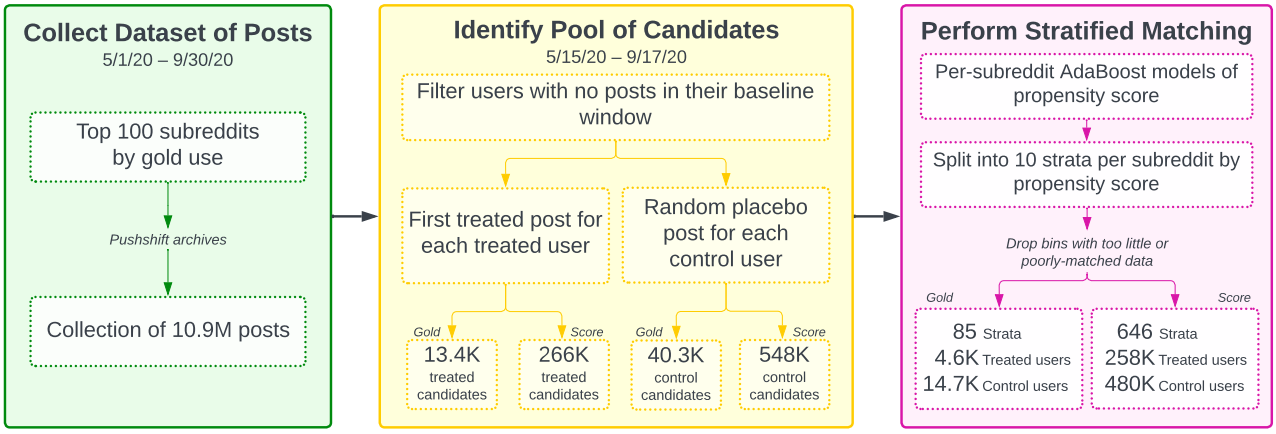



**Figure 2: Pipeline from data collection to forming matched treated and control samples used in our causal inference methodology.**

42, 82]. Our goal is to examine the causal effects of receiving positive feedback (i.e., the treatment) compared to what would have happened if the treatment was not administered (i.e., the *counterfactual*). However, it is not possible to find a post that simultaneously did and did not receive positive feedback, so a *true* counterfactual is missing or unobserved. The potential outcomes framework allows us to estimate the missing counterfactual for each treated post based on the outcomes of similar, observed posts that were not treated (i.e., did not receive positive feedback). We estimate the counterfactual by identifying a treatment group (posts that receive positive feedback) and a similar control group (posts that did not receive positive feedback). We identify similar posts from each group to create matches, allowing us to compare the trajectory of similar users who differ only in whether they received our treatment (i.e., positive feedback). The full pipeline of this process is visualized in Figure 2 along with descriptive statistics about the dataset.

## 3.1 Treatments: Posts That Received Positive Feedback

There are many forms of feedback available on Reddit. We select two to focus on in this work: *gold* and *score*. Figure 1 visualizes both forms of feedback in a real Reddit post.

Gold was an award-like signal consisting of gold icons that were awarded to posts. Gold and other awards were discontinued by Reddit in 2023,[7] however gold was considered beloved by some users on Reddit [103] and many users were disappointed that they were not among the features brought back to the platform in May 2024. Thus, we believe there is still value in exploring how gold affected the recipients to understand whether Reddit removed a mechanism that had positive impacts on user behavior and experience.

Because gold was a paid feature, it is a potentially strong signal of quality given the rarity with which it was given out. However, the real-world cost of gilding means that gold may not be entirely representative of all opinions or values from a community. To account for this, we select score (i.e., aggregated upvotes and downvotes) as our second treatment. Upvotes are the most widespread signal of approval on Reddit, utilized by every community. To contrast gold, upvotes are free to give out, therefore there is no limit on how many a user can hand out, neither explicitly imposed by the platform nor implicitly imposed by cost. We note that some moderators believe that upvotes are not necessarily a reliable indicator of quality, specifically in the context of communities like r/AskHistorians that value input from experts [43]. However, these moderators are not claiming that upvotes cannot reliably inform us what a community likes. Thus, we recognize this limitation of our work and do not claim that the posts treated with upvotes contain accurate information or even content that moderators themselves would like to encourage. Furthermore, we rely on the concept of contingency [66] by assuming that receiving the treatments is contingent upon posting something deserving of reinforcement within the community. Thus, we trust community signals of approval and do not evaluate how deserving the treated posts are. This mimics approaches of prior work which focus only on the outcomes of treatment and do not consider the precursors leading to the treatment being applied [41, 42].

## 3.2 Identify Pool of Candidates

Following approaches in prior work [17, 42, 48], we use these two forms of treatment to identify treated and control candidates. To begin, we use Pushshift [11] archives of Reddit data from May 1, 2020 to September 30, 2020 (our study period). This research did not involve any participation from users and all data was publicly available on Reddit, thus our work did not require approval from an institutional review board. We followed best practices to maintain anonymity in our data (e.g., looking at data only in aggregate) and ensure user privacy.

Since upvotes are used universally and gold was not utilized by all communities, we construct a sample from the 100 subreddits which awarded the most gold in the study period. We form distinct pools of treatment and control posts, $C_{\text{gold}}$ and $C_{\text{score}}$, for each treatment, made within our *treatment window*: May 15, 2020 to September 17, 2020. This smaller window allows us to study user behavior in the 14 days before and after treatment. Posts which received at least one gold are considered treated candidates in $C_{\text{gold}}$. Posts that receive no gold are added to $C_{\text{gold}}$ as control candidates.

[7] https://www.reddit.com/r/reddit/comments/14ytp7s/reworking_awarding_changes_to_awards_coins_and/

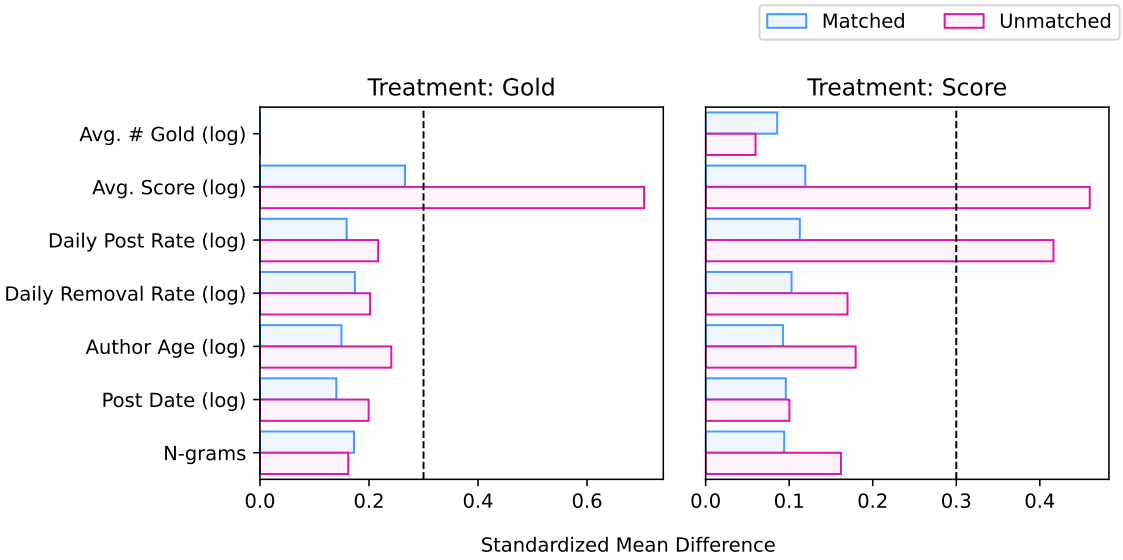



**Figure 3: Standardized mean differences (SMD) for each covariate averaged across all strata. We report both SMDs for matched data and unmatched data for comparison. The figure shows that matching improved the SMD for most covariates and that the SMD for all matched data is below the threshold 0.3, indicating high-quality matches.**

To construct $C_{\text{score}}$, we need to identify some threshold above which a post is considered to be highly upvoted. To accomplish this, we split the distribution of scores for all posts in each subreddit into 4 quantiles per month. The highest quantile consists of the highest-scored posts which are added to $C_{\text{score}}$ as treated candidates. Posts in the bottom two quantiles comprise the control candidates in $C_{\text{score}}$. We drop all posts in the second highest quantile to provide a buffer between control scores and treated scores, ensuring that our treatment is strongly defined.

For each treated and control candidate in $C_{\text{gold}}$ and $C_{\text{score}}$, we ensure that the author of the post has made at least one post in the 14 days prior, none of which received the respective treatment. This provides us with baseline activity levels and avoids users receiving multiple treatments in a short span of time. In this setup, it is possible for one user to appear in both candidate pools multiple times, a fluid interpretation of our treatments since the start of our study period is arbitrary so we do not know which users receive treatment before the start of our analysis.

For both treatments, we sample 3 times as many control users from each subreddit as there are treated users to balance the groups while still providing sufficient control candidates for matching. Figure 2 reports summary statistics about the final pools of candidates, $C_{\text{gold}}$ and $C_{\text{score}}$.

For each candidate in $C_{\text{gold}}$ and $C_{\text{score}}$, we dynamically define three events of interest:

- **Baseline window:** the posting activity of the post's author in the 14 days prior to treatment or placebo.
- **Treatment/placebo time:** the creation time of the post that received the treatment or placebo.
- **Observation window**: the posting activity of the post's author in the 14 days after treatment/placebo time.

We note that our analysis is limited by the fact that the precise time at which positive feedback is given is not publicly accessible, so we do not have confirmation that the recipients are aware that the treatment has been applied. We use the creation time of the treated post as a proxy for treatment time which may raise issues with immediacy [66], in that authors are likely not treated immediately after contributing and they may not realize they received the treatment until even later. We could have further restricted the treatment pool to authors who also made comments on their gilded post, however it is not guaranteed that the gold was given prior to their comment. Thus, we consider our analysis to be a conservative estimate of the effect of positive feedback on users.

## 3.3 Stratified Matching

We want to minimize the differences in pre-treatment author and environmental characteristics between the control and treatment groups so observed effects can be attributed to the treatments. For each treated user, our goal is to identify similar control users from the same community based on treatment or placebo post content and pre-treatment posting behavior and content quality.

We perform within-subreddit stratified matching on propensity scores in each candidate pool. Stratified propensity score matching has been used in prior work [48, 68, 84, 86, 100, 110] to minimize biases of propensity score matching identified by King and Nielsen [51]. This approach enables us to balance for the bias-variance trade-off of either too biased (one-to-one-matching) or too variant (unmatched) data comparisons, and help isolate and examine the effects of treatment within each stratum [42, 83]. For each subreddit, we train an AdaBoost classifier which outputs a propensity score given the following covariates (log-scaled to achieve more normal distributions) as independent variables:

- Daily posting rate in the baseline window.
- Average score on posts in the baseline period.
- Treatment/placebo time.
- Author's age (i.e., days since account creation) at treatment/placebo time.
- Daily removal rate in the baseline period.
- Frequency of 100 bigrams from the text contained in the treatment/placebo post (i.e., titles and post bodies). We use the 100 most common bigrams appearing in all of the posts in $C_{\text{gold}}$ or $C_{\text{score}}$.

When considering score as our treatment, we also include the following covariate:

- Average number of gold received per post in the baseline period.

This variable does not apply to authors of posts in $C_{\text{gold}}$ because, by definition, they must not have been gilded in their baseline period.

The resulting propensity scores reflect how likely a user is to belong to the treatment group based on the covariates. These covariates allow us to match treatment and control users that have similar "quality" posts in the pre-treatment window, so the authors should be similarly good writers before the treatment occurs. After matching, any effects observed after the treatment is applied can be attributed to the treatment itself. Within each subreddit, we bin the propensity scores into 10 strata using quantiles, filter out the strata with fewer than 10 users, and ensure that each stratum contains both treated and control users.

Each resulting stratum represents a group of users with similar likelihood of being treated, thus the strata are a set of many-to-many matches for each subreddit such that every treated user in a stratum is matched with all control users in the stratum. We assign each stratum a unique identifier to be used in subsequent analysis. We refer to the stratum identifier of author $a$ as $s_a$.

To measure the quality of our matches, we compute the Standardized Mean Difference (SMD) between the treatment and control samples of each stratum for each covariate. We use an upper bound of 0.3 to indicate good match quality overall [44, 48]. We drop all strata with a mean SMD over all covariates that exceeds 0.3 to ensure that we retain only well-matched treatment and control posts and obtain an unbiased estimate of the treatment effects. Figure 3 visualizes the SMD values for each of our covariates averaged across the remaining strata (i.e., *matched* samples) and across the original pool of candidates (i.e., *unmatched* samples) to show that matching improved the SMDs for our covariates. We note that the SMDs for two of the covariates were slightly lower before matching, however the mean matched SMDs are still below our threshold of 0.3. We report descriptive statistics about the final matches in Figure 2.

### 3.4 Outcomes

With this dataset of well-matched treatment and control posts, we examine the effects of receiving positive feedback on four major user-level themes: reception from the community, content, engagement, and norm-adherence. To measure community reception, we look at the effect of our treatments on each user's future *average number of gold* and future *average score* per day. These outcomes capture whether receiving positive feedback encourages users to post more positive-feedback-worthy content in the future. To understand the effect of positive feedback on content, we measure the *text difference* between the treatment or placebo post and the recipient's future contributions. This difference is measured by computing sentence embeddings [76] for the text in each post's title and body and measuring the cosine distance to the text of their treated or placebo post. For engagement, prior work describes positive feedback as a mechanism capable of increasing user motivation to participate Kraut et al. [54]. We use *daily posting rate* as a proxy for user motivation to participate. Finally, to capture the impact of treatment on norm-adherence, we utilize moderation decisions by each community's centralized moderation team through each user's *daily removal rate*.

For each post in our matched sample, we collect all other posts by the same author in the baseline and observation windows. This gives us a set of 309K posts for our gold treatment and 5.4M posts for our score treatment. We calculate daily values for each of our five outcomes, taking care to exclude the treated or placebo post to not skew any averages. This results in 129K daily aggregates for the gold treatment and 2.6M daily aggregates for the score treatment.

## 4 RQ1: What is the effect of positive feedback on recipients?

Using our matched treatment and control samples, we perform analyses to understand the causal impact of receiving positive feedback on user behavior and how long those effects last.

### 4.1 RQ1a: How does receiving gold or being highly upvoted impact users' future behavior in a community?

To understand how the treatments affect users (RQ1a), we conduct a Difference-in-Differences (DiD) analysis similar to analyses done in prior work [2, 17, 35]. The DiD analysis compares the difference in treated user's behavior before and after receiving treatment against the difference the control group experiences over time. Like any causal inference study using observational data, there are limitations to our approach [106]. Our matching process accounts for measured confounds in our data (i.e., the covariates) and the DiD analysis captures time-invariant unmeasured confounds, for example, users in small communities experiencing a surge in subscribers may receive higher score on their posts across the board. It is possible, however, that external events impact how a community reacts to content in ways not captured by our observational data, leaving some time-variant unmeasured confounds unaccounted for. Future work can employ true experiments to account for such confounds and extend our understanding of the impact of positive feedback.

To estimate causal effects through DiD analysis, the data must satisfy the *parallel trends assumption* [1]. This means the difference between the outcomes for the treatment and control groups must be constant over time, prior to treatment. Similar to prior work [64, 81], we visually inspect our data in Figure 4 and conclude that the parallel trends assumption is satisfied in all but two cases, daily post rate and daily removal rate for the gold outcome, which we exclude from the DiD analysis. This ensures the internal validity of the DiD models for the other outcomes and any observed effects [1].

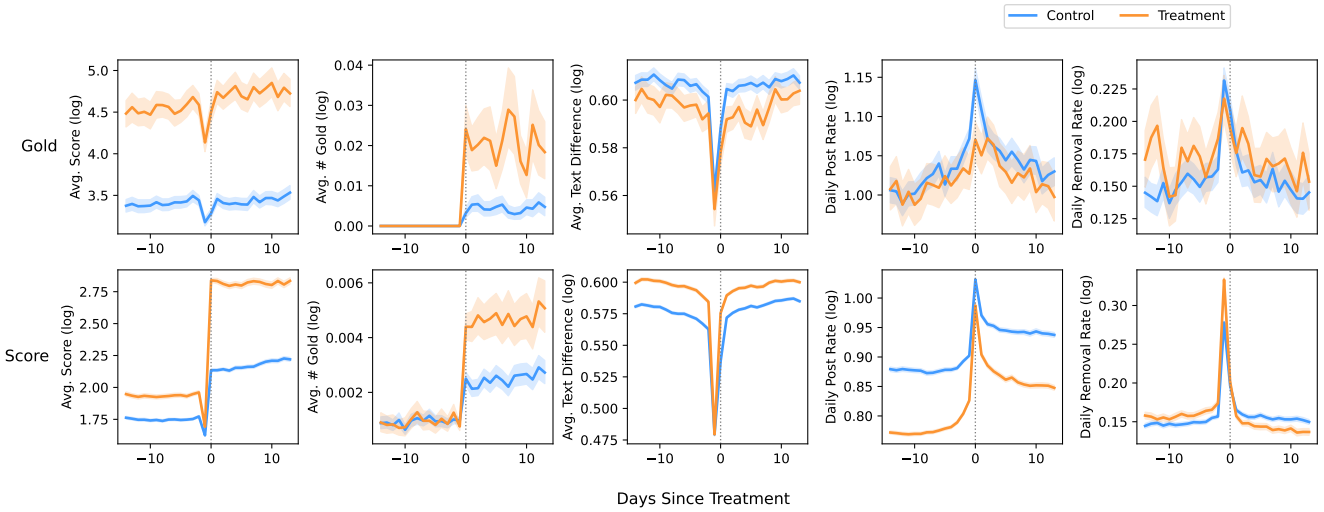



**Figure 4: Visualization of each outcome over the baseline and observation windows (excluding the treatment/placebo posts). These plots are used to validate the parallel trends assumption needed for difference-in-differences analysis.**

We use our daily aggregates in the baseline and observation windows for each user in the following regression:

$$y_{a,d} \sim a_t * \delta(d \geq 0) + d + s_a$$

where $y_{a,d}$ is the value of outcome $y$ for author $a$ at $d$ days after treatment/placebo, $a_t$ indicates whether author $a$ was treated, $d$ represents the number of days since treatment/placebo, $\delta(d \geq 0)$ indicates whether the aggregated data occurred after treatment, and $s_a$ is the stratum identifier of the author (see Section 3.3). We conduct one regression for each treatment and each outcome.

*4.1.1 Findings.* The results of our DiD analysis are reported in Table 1. The increase column is calculated from the coefficient on the $a_t * \delta(d \geq 0)$ term which is the interaction between time and treatment group. This indicates the effect of treatment on treated users in the observation window compared to if they had not received the treatment, i.e., the counterfactual. To reduce the likelihood of spurious significant relationships, we perform a Bonferroni correction by multiplying the $p$-values of each coefficient by the product of the number of regression tests (8) and the number of independent variables per regression (4). To capture the effect size, we also report the Cohen's $d$ measured between the difference in each dependent variable before and after treatment for each treated user compared to the same difference before and after placebo for each control user.

We find that both forms of treatment lead to large positive increases in treated users' score on future posts compared to if they had not received the treatment. More specifically, receiving gold and being highly upvoted correlate with a user receiving 19.28% and 57.11% higher score than they would have, respectively. Both treatments have much more modest, but still significant, effects on the average number of gold received on future posts. The gold treatment encourages 1.63% more gold in the future while the score treatment encourages an increase of less than 1% more gold.

**Table 1: Difference-in-differences regression analysis results. We report the percent increase observed in the dependent variable for each treated user compared to if they had never received the treatment. Regressions with Bonferroni-corrected $p$-values are reported with asterisks ($^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$). Cohen's $d$ reports the effect size of treatment on the dependent variables. The table shows significant increases in score and gold as a result of either treatments, and both increases in text difference and daily post rate and a decrease in daily removal rate as a result of the score treatment.**

| Treatment | DV | Increase | Cohen's $d$ |
|---|---|---|---|
| Gold | Avg. Score | 19.28%*** | 0.16 |
| | Avg. # Gold | 1.63%*** | 0.27 |
| | Avg. Text Difference | -0.23% | -0.05 |
| Score | Avg. Score | 57.11%*** | 0.14 |
| | Avg. # Gold | 0.2%*** | 0.04 |
| | Avg. Text Difference | 0.34%*** | 0.01 |
| | Daily Post Rate | 2.24%*** | 0.04 |
| | Daily Removal Rate | -2.58%*** | -0.02 |

Interestingly, users treated with score post content 0.34% more distinct from their treated post than they would have if they had not received the treatment. They also post 2.24% more per day than they would have and experience a negative effect on removal rate. This demonstrates that being highly upvoted might encourage users to post more frequent positive contributions while learning high-level community norms and thus violating them less often.

Prior work on the platform Slashdot suggests that less popular posts may be more likely to be removed by volunteer moderators in a distributed moderation system [60]. Therefore the decrease in

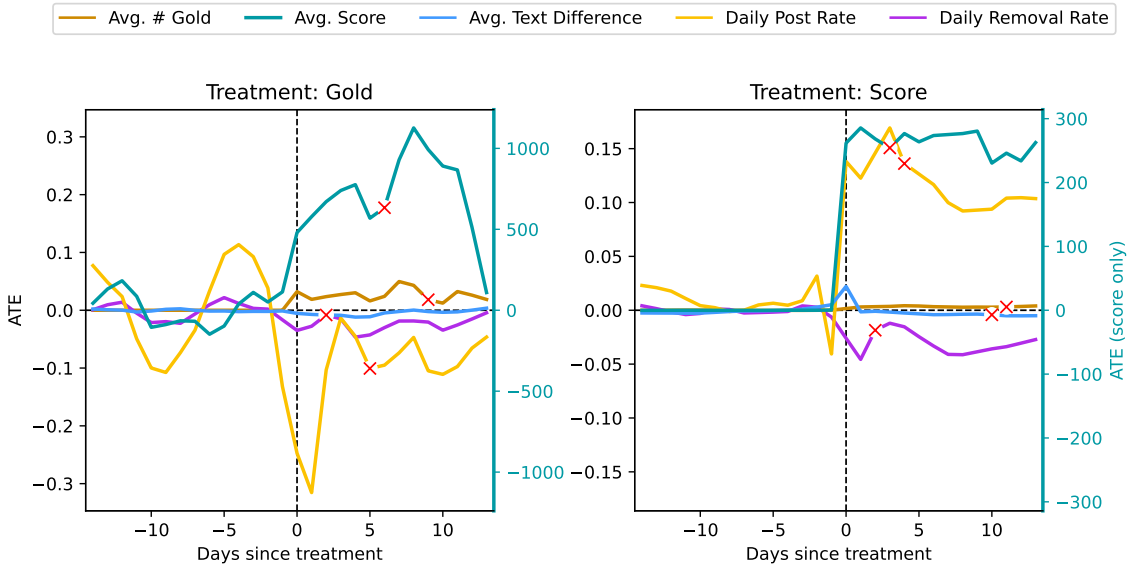



**Figure 5: Average treatment effects (ATE) of each treatment on the five outcomes. The rightmost $y$-axes correspond to the ATE of the score outcome which has a much larger range. The "x" markers indicate the days at which each outcome measure reaches saturation (i.e., decreases or stays the same for two consecutive days). The figure demonstrates the lasting effect of the score treatment on each outcome compared to the gold treatment, where the ATEs mostly return to zero by the end of our observation window.**

removal rate after treatment may be attributable to the concurrent increase in score and not an indication of norm acquisition. However, other research on Reddit failed to confirm this trend of less harsh moderation decisions pertaining to higher-scored posts [99].

Overall, receiving positive feedback of either kind encourages users to contribute posts that are better-received by their communities, and in some cases, also promotes user engagement and norm acquisition. Given the increase in future positive feedback and the decrease in norm-violations, we say that being highly upvoted encourages users to post higher-quality content according to both the community and the centralized moderation team.

## 4.2 RQ1b: How long do the effects of positive feedback on user behavior last?

To determine how long the effect of each treatment lasts (RQ1b), we compute the Average Treatment Effect (ATE) for each of our five outcomes on each day in our study period. ATE is computed as the difference between the treated users' change in outcome and the control users' change in outcome before and after treatment. We visualize the ATE values in Figure 5 averaged across each day in the baseline and observation windows. For each outcome, we also identify the day at which it reaches *saturation*, meaning the ATE for the outcome stops increasing for two consecutive days after treatment. Formally, our saturation point is defined as the day $d$ such that $ATE(d) \leq ATE(d-1) \leq ATE(d-2)$. For the removal rate outcome, we reverse the definition of saturation point since we are interested in seeing decreases in norm violations. In other words, the saturation point is the first day $d$ for which $ATE(d) \geq ATE(d-1) \geq ATE(d-2)$. These points are denoted by red "x" symbols in Figure 5.

We observe that for the gold treatment, the ATE values for all five outcomes return to nearly zero by the end of the observation window, indicating minimal differences between treated and control users. For the users treated with score, however, the ATE of their score and post rate in the observation window remain well above zero at the end of the two-week observation window, even though both measures reach their saturation point by day 5. This indicates that the effect of being highly-upvoted on future upvotes and engagement may last longer than the effect of being gilded. The elevated scores in the observation window may also serve as additional treatments, creating a snowball effect.

*4.2.1 Rebound.* We extend this analysis by exploring how often the outcome variables return roughly to their baseline values. Each user's *baseline region* for an outcome measure $\mathcal{M}$ is defined as the range $[0.9 * \mathcal{M}_{\text{baseline}}, 1.1 * \mathcal{M}_{\text{baseline}}]$. For each treatment-outcome pair, we find all users who return to their baseline region for two consecutive posts (i.e., rebound). Table 2 summarizes the number of users who rebound for each outcome. The percentages in Table 2 only consider users who post twice in their observation windows to satisfy the definition of rebound.

We notice that most treated users do not rebound on the score outcome within 14 days, reflecting the trend shown in Figure 5 that score is elevated throughout the entire observation window. Daily post rate follows this same trend regardless of the treatment. However, for the average number of gold, nearly all users in both treatments rebound to their baseline values within two weeks. This is understandable given the rarity of gilding in the dataset.

**Table 2: Summary table describing the number of treated users that rebound. By definition, users must post at least twice in the 2-week post-treatment window in order to rebound, which is true for 2,599 (60.2%) of users treated with gold and 90,993 (35.27%) users treated with score. The table shows that users almost always rebound on their average number of gold, rarely rebound on their average score and daily post rate, and sometimes rebound on their text difference and daily removal rate.**

| Treatment | Outcome | # Rebound | # Never Rebound |
|---|---|---|---|
| Gold | Avg. Score | 74 (1.72%) | 2,525 (58.53%) |
| | Avg. # Gold | 2,547 (59.04%) | 52 (1.21%) |
| | Avg. Text Difference | 1,768 (40.98%) | 831 (19.26%) |
| | Daily Post Rate | 159 (3.69%) | 2,440 (56.56%) |
| | Daily Removal Rate | 1,282 (29.72%) | 1,317 (30.53%) |
| Score | Avg. Score | 3,640 (1.41%) | 87,353 (33.87%) |
| | Avg. # Gold | 90,139 (34.95%) | 854 (0.33%) |
| | Avg. Text Difference | 44,509 (17.26%) | 46,484 (18.02%) |
| | Daily Post Rate | 468 (0.18%) | 90,525 (35.10%) |
| | Daily Removal Rate | 52,314 (20.28%) | 38,679 (15.00%) |

Finally, we observe that the percentage of users who rebound to baseline text difference and removal rates after either treatment is much more evenly split. This suggests that the treatments have a lasting impact on the content of some users' future posts and teach some users about the norms more effectively than others. Our subsequent RQ2 analysis explores this trend in more depth.

## 5 RQ2: How do the effects of positive feedback vary across different types of recipients?

Our RQ1 analysis raised questions about differential effects of treatment: do some users or communities benefit from treatment more than others? We conduct two subsequent regression analyses to explore this question.

### 5.1 RQ2a: What types of users experience stronger effects of receiving positive feedback?

We compute Individual Treatment Effects (ITE) of each of our treatments on all five outcomes to evaluate user-level differential effects. ITE is used in prior work to compute how much a single treated user is affected by the treatment [85]. In our work specifically, ITE is computed as the difference between a treated user's change in behavior after treatment and the average change in all control users' behavior from the same stratum after placebo. More precisely, for an individual treated user $u_{t,s}$ from stratum $s$, we compute the ITE for each outcome as follows:

$$ITE_{\text{outcome}}(u_{t,s}) = \Delta u_{t,s}(\text{outcome}) - \text{mean}(\Delta u_{c,s}(\text{outcome}) : u_{c,s} \in \mathcal{D}_{control,s})$$

where $\mathcal{D}_{control,s}$ is the set of all control users in stratum $s$. Users who do not post in the observation window are excluded from this analysis since they do not have post-treatment data.

We then perform one linear regression analysis for each treatment-outcome pair using the outcome's ITE as the dependent variable. Baseline values for our five outcomes are independent variables along with the author's age, karma (i.e., the sum of scores on all posts) in the month prior to the baseline window, stratum, and the strength of treatment (i.e., number of gold or score on treated post). We take the log of each numerical independent variable.

*5.1.1 Findings.* Table 3 reports the coefficients from this regression. The bold coefficients are those we discuss in this section. All significance tests are Bonferroni-corrected based on the number of independent variables and regressions run. For the gold treatment, each $p$-value is multiplied by 40 (5 regressions, 8 IVs) and for score, each $p$-value is multiplied by 45 (5 regressions, 9 IVs). This regression analysis demonstrates that the strength of either treatment may lead to stronger effects. For example, posts treated with more gold are positively correlated with the treated user experiencing stronger effects on their future number of gold. Similarly, users who are treated with higher score have more extreme effects on their future positive feedback than users with more modest treatments. With this score treatment, however, we see that being treated with lower score also correlates with a larger effect on removal rate.

We also see trends aligned with the ideas of satiation and deprivation introduced by Miltenberger [66]. Specifically, our regression shows that users who have higher baseline average score experienced diminished effects of either treatment on their future score. Similarly, users with a higher average number of gold in their baseline window see less dramatic effects of the score treatment on their future number of gold. Both these trends reflect the idea that users are more strongly impacted by stimuli (e.g., gold or upvotes) if they have been deprived of that stimulus.

Additionally, the age of a user's account seems to affect the intensity of the impact of the score treatment on some of the outcomes. Specifically, newer accounts/authors experience the largest effects of being highly upvoted on their daily posting frequency and daily removal rates. Based on our results in Table 1, this means larger increases in post rate and larger decreases in norm violation rates for newer authors. Thus, Reddit newcomers may be especially motivated by score to both continue contributing to a community and also learn more about the community's norms.

Younger recipients of the score treatment also experience smaller increases in text difference, meaning newcomers post content in the

**Table 3: Individual Treatment Effects (ITE) regression results. Cells that are grayed out indicate independent variables not present in the regression. Regressions with Bonferroni-corrected $p$-values ($^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$). Coefficients with $p > 0.05$ are excluded. Bold coefficients are discussed in Section 5.1.1.**

| Treatment → | Gold | | | | | Score | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| IVs ↓ ITE → | Score | # Gold | Text Diff. | Post Rate | Rem. Rate | Score | # Gold | Text Diff. | Post Rate | Rem. Rate |
| Author Age | | | | | | | | **4.6E-04**$^{***}$ | **-9.7E-03**$^{***}$ | **-2.6E-03**$^{***}$ |
| Baseline Avg. Score | **-5.5E-05**$^{***}$ | 1.8E-06$^{***}$ | | | | **-1.4E-02**$^{***}$ | | | -3.1E-04$^{***}$ | 1.4E-03$^{***}$ |
| Baseline Text Diff. | | | -5.5E-01$^{***}$ | | 2.0E-01$^{**}$ | -1.5E-01$^{***}$ | -5.9E-03$^{***}$ | -8.0E-01$^{***}$ | -3.2E-02$^{***}$ | 2.5E-01$^{***}$ |
| Baseline Post Rate | | | | -2.3E-02$^{***}$ | | -8.8E-01$^{***}$ | -4.2E-03$^{***}$ | -7.2E-03$^{***}$ | -2.0E-01$^{***}$ | 3.2E-01$^{***}$ |
| Baseline Rem. Rate | | | | -7.1E-02$^{*}$ | -2.6E-01$^{***}$ | 2.2E+00$^{***}$ | | 4.4E-02$^{***}$ | -3.1E-01$^{***}$ | -2.4E+00$^{***}$ |
| Karma Last Month | 1.1E-01$^{***}$ | | | | | 1.8E-02$^{***}$ | | 5.4E-04$^{***}$ | 1.0E-03$^{*}$ | |
| Gold Received | | **3.0E-02**$^{**}$ | | | | | | | | |
| Score Received | | | | | | **9.7E-02**$^{***}$ | **5.7E-04**$^{***}$ | | | **-3.0E-03**$^{***}$ |
| Baseline Avg. # Gold | | | | | | | **-5.0E-01**$^{***}$ | | | |
| R-squared | 0.111 | 0.052 | 0.395 | 0.109 | 0.080 | 0.108 | 0.183 | 0.579 | 0.058 | 0.250 |

future more similar to the post that received the reward compared to more established Reddit users.

## 5.2 RQ2b: What types of communities experience stronger average effects of receiving positive feedback?

We follow up this user-level analysis with a community-level analysis involving Community Treatment Effects (CTE) for each treatment and outcome pairing. We define the CTE of an outcome for a Reddit community $r$ as follows:

$$CTE_{\text{outcome}}(r) = \text{mean}(\Delta u_{t,r}(\text{outcome}) : u_{t,r} \in \mathcal{D}_{treated,r}) - \text{mean}(\Delta u_{c,r}(\text{outcome}) : u_{c,r} \in \mathcal{D}_{control,r})$$

where $\mathcal{D}_{treated,r}$ and $\mathcal{D}_{control,r}$ represent the sets of all treated and control users in subreddit $r$ respectively.

Similar to the previous ITE analysis, we use our CTE values as dependent variables for ten linear regression analyses with subreddit-level monthly activity statistics from our entire study period as independent variables. This includes post rate, average number of gold, average score, the number of unique authors, and the number of subscribers.

*5.2.1 Findings.* After Bonferroni correction, these regressions yielded only two significant coefficients revealing that subreddits that gild more often see larger effects of the score treatment on their future number of gold and text difference. The lack of significant results indicates that the treatments have similar effects regardless of the community in which they occur.

# 6 Discussion

Our work contributes a computational approach and a causal inference framework through which researchers can examine the differential effects of treatment outcomes. We outline a method of identifying treatment and control posts, performing a difference-in-differences analysis, and determining user- and community-level factors that influence the strength of the treatment effect. This methodology can be adapted with different treatments, outcomes, and platforms to fit researchers' needs. Additionally, while we focused on two types of positive feedback available on Reddit, upvotes in particular are almost universally applicable across social media platforms and have analogous features on Facebook, Instagram, X, Stack Overflow, and YouTube, among others. This ubiquity of upvotes enhances the generalizability of our work. Researchers interested in other platforms can use these findings to inform research into analogous features, and other forms of formal feedback, and extend this work with analyses of positive feedback signals not available on Reddit.

Our findings demonstrate the power Reddit affords community members through signals of positive feedback on the interface. We also revealed the differential effects of positive feedback, specifically that different forms of feedback have different effects on different types of recipients. These results have important implications for content moderation and motivate the need to emphasize positive feedback in HCI research and in platform design.

## 6.1 Prioritize highlighting contributions from users recently deprived of positive feedback.

A key differential effect we observed relates to the principle of deprivation in positive reinforcement theory. The principle states that a stimulus has larger impacts on subjects deprived of it [66]. Our findings confirm this theory in the context of positive feedback on Reddit, highlighting that users who had not recently received gold experienced stronger effects of being gilded on their future gold received. Similarly, users who had not recently been highly-upvoted experienced larger increases in their future score when experiencing the score treatment.

*Design recommendations. We recommend that designers and HCI researchers consider the principles of satiation and deprivation to best utilize valuable community resources like time and effort.* Specifically, we suggest that they enable moderators and community members to maximize the potential effects of their positive feedback by building on work highlighting users' history of toxicity [38] and instead providing signals of how deprived or satiated an author is, or by prioritizing content in social media feeds contributed by authors deprived of positive feedback.

## 6.2 Provide mechanisms to bring attention to high-quality newcomer contributions.

As moderators are largely responsible for sifting through norm-violating content, they stand to benefit the most from users learning to contribute norm-abiding content. However, the idea of norm

acquisition can be challenging, especially for newcomers to a community [54], and is often a barrier to entry when users want to contribute to a community in good faith [45, 74, 89]. Some platforms (e.g., Wikipedia) have developed systems to assist newcomers in learning to make quality contributions, but research has shown that some of these methods may be hurting newcomers more than they are helping [90]. Additionally, while some norms are consistent platform-wide, there are other norms that may only apply to a handful of subreddits [18]. Thus, even established Reddit users who participate in many communities can face difficulties trying to comply with the rules of other communities [40].

We showed that positive feedback in the form of upvotes results in the recipient posting fewer norm-violating posts than they would have without the treatment, implying that recipients are learning community norms. Additionally, newer accounts see stronger effects on their norm-adherence after treatment compared to more established Reddit users indicating that platform-level newcomers benefit most from positive feedback in the context of norm acquisition. Newcomers also more closely match the text of their future posts to the one that was rewarded, suggesting the power of positive feedback to shape the actual content of newcomer contributions.

*Design recommendations.* These findings indicate that *positive feedback as a means to support norm acquisition is imperative for the success of moderation teams.* Specifically, we know that exposure to harmful content is concerning to volunteer moderators [96]. Our work shows that users who receive positive feedback are less likely to be removed and thus not exposing moderators to potentially harmful, norm-violating content as frequently as they would have. Therefore, the use of positive feedback may have positive effects on moderator well-being.

To enable this norm acquisition, we call on designers and HCI researchers to explore explicit ways for moderation teams to encourage established community-members to review newcomer contributions and provide positive feedback when applicable. For example, this may be a highlighting mechanism similar to YouTube creator hearts which have been shown to increase positive engagement for hearted comments [22]. Designers may also consider building off a prior intervention that nudges users to preemptively reflect on their participation [20] and build a system to nudge users to specifically engage with newcomer content. Prior work on visualizing conversational metrics [21] can be extended to develop tools that guide moderator attention towards desirable behavior. Finally, platforms can incorporate a dedicated feed of newcomer contributions within communities to enable easy discovery of newcomer content.

## 6.3 Platforms need positive feedback to sustain user motivation to contribute.

Kraut et al. [54] state that "to be successful, online communities need the people who participate in them to contribute the resources on which the group's existence is built." In the context of Reddit, those resources include memes, links to news articles, or other subreddit-specific contributions. Without users making such contributions, platforms would decline in popularity and lose out on profit. Thus, platforms must maintain user motivation.

User motivation can be intrinsic or extrinsic [54]. In the context of our research, intrinsically-motivated users contribute high-quality content because they enjoy doing so. Kraut et al. [54] highlight performance-based positive feedback as an especially effective method of enhancing users' intrinsic motivation. Therefore, users who receive community-wide approval through positive feedback may experience increased intrinsic motivation to participate.

Extrinsically-motivated users may be more interested in making high-quality posts as a means to increase their reputation (i.e., karma) in the community. Rewards such as gold and upvotes are extrinsic motivators in themselves, which can help these types of users feel more motivated to participate [54].

*Implications for platforms and moderators.* Regardless of whether our treated users' motivation was intrinsic or extrinsic, it was enhanced by positive feedback [54]. Our findings confirm that users who contribute posts that get highly-upvoted increase their posting rate, indicating a potential increase in motivation to participate. Consequently, we assert that *positive feedback is crucial to platform success given its significant role in supporting user motivation.*

## 6.4 Centralized and distributed moderation systems can support communities through positive feedback.

Reddit moderation consists of a centralized moderation team and a community of users implicitly engaging in distributed moderation practices. Based on our findings, both facets of Reddit's moderation system can help improve the quality of content and norm-adherence in communities.

*Users already have the power to shape their communities through distributed moderation.* The sorting algorithms of feeds on platforms like Reddit are often opaque to users and referred to as "black boxes" since platforms do not disclose the metrics they use to generate these feeds [15]. However, users may have hypotheses about how the sorting works [28]. For example, surveyed moderators reported providing various forms of positive feedback to boost the visibility of content, implying that signals like score and gold affect a post's position in an algorithmically-curated feed [55]. However, our research has shown that positive feedback mechanisms can go beyond influencing the sorting algorithm and actually impact the quality of content generated by users in the future.

Because gold and upvotes both encourage users to contribute posts that receive higher score and more gold, we conclude that both forms of positive feedback encourage users to contribute posts that will be better received by the community. This indicates that the community overall is more satisfied with contributions made by recipients of positive feedback. This reveals the power communities already have to proactively encourage a better pool of content through positive reinforcement. Thus, though moderators have additional weight and power to make decisions and enforce norms in a community, they are not the only ones capable of effecting positive change.

Though this work focuses on the capacity of community-members to facilitate distributed moderation through existing interface signals, centralized moderation teams on Reddit can learn from our analysis to inform how they moderate. Prior work found that

YouTube comments receiving a "heart" from the creator of a video resulted in increased positive feedback from the community as a whole [22]. Thus, when moderators demonstrate their approval of a contribution through highlighting, the community at large may reward it with positive feedback mechanisms that we found to positively impact user-level outcomes.

*Implications for online moderation.* We found that users already have the power to shape their communities through distributed moderation. When members of a community collectively reward a contribution, we see a significant positive impact of the distributed effort. Thus, *platforms and researchers should support community-driven moderation through positive feedback to give communities the agency to proactively curate their content.*

Moreover, members of centralized moderation teams can curate posts to be rewarded through distributed moderation. Since Reddit does not explicitly provide moderator-specific feedback mechanisms, *we suggest moderation teams focus on influencing the visibility of posts which can impact the amount of positive feedback they receive.* Moderators can pin high-quality posts or periodically consolidate posts in a moderator-curated list in the subreddit sidebar, both of which moderators have reported doing [55].

## 6.5 Limitations and Future Work

Our research opens many avenues for the exploration of positive reinforcement and quantifying desirable content in online spaces. We summarize the limitations of our work and propose extensions to address them.

*6.5.1 Build models of desirability.* Our analysis does not consider whether the posts that received the positive feedback are deserving of such approval. Definitions of desirability can vary greatly [32, 56], thus future work should explore what is considered desirable within and across online communities, and develop context-sensitive approaches to model and identify desirable behavior online.

*6.5.2 Extend understanding of content-specific reinforcement.* Our analysis showed that recipients of positive feedback receive even more positive feedback on subsequent posts, but that the content of those posts is more distinct from the rewarded post than it would have been without the treatment. This suggests that receiving positive feedback helps users apply community norms beyond making redundant contributions. However, this effect was small and it is possible that working with posts, which do not always have much text, does not allow us to capture content-level reinforcement. We suggest researchers extend this work by exploring the effects of positive feedback on reinforcing specific types of content within both posts and comments.

*6.5.3 Explore the effects of positive feedback on bystanders.* While the effects studied in this research were all specific to the recipients of positive feedback, there may also be effects on users who interact with rewarded content through vicarious reinforcement. Jhaver et al. [42] show that bystanders witnessing post removal explanations experience effects on their future behavior. Future work is needed to understand whether bystanders of positive feedback are similarly affected and whether users can learn community norms by witnessing such positive interventions.

*6.5.4 Compare community feedback to moderator-specific forms of feedback.* Our choice of treatments is limited in that some moderators do not consider upvotes a signal of quality [43]. This can be problematic in communities that require expertise for participation, since users giving out upvotes may not be equipped to evaluate how much a post deserves positive feedback. In such communities, moderators may be better judges of quality. Prior work shows that moderators are utilizing moderator-specific signals as a means for positive reinforcement [55], thus future work should study how these forms of feedback affect the recipients and whether positive feedback from a position of authority has more or less capacity to reinforce user-level outcomes than anonymous forms of feedback.

## 7 Conclusion

In this work, we examined the effects of two forms of positive feedback on Reddit users and study whether positive reinforcement can be employed through distributed moderation practices. We found that both receiving gold on a post and being highly-upvoted by the community encourage users to post content in the future that receives even more positive feedback than they would have otherwise. Through a user-specific analysis, we also find that receiving larger amounts of positive feedback is associated with larger effects on future positive feedback though users already satiated by recent positive feedback will not see as strong effects. Importantly, we also found that newcomers experience stronger effects of being highly-upvoted with respect to their future rate of contribution and norm acquisition. Based on our findings, we contribute design recommendations and discuss the implications of our work with respect to moderators, communities, and platforms beyond Reddit.

## Acknowledgments

We thank the members of the Social Computing Lab (SCUBA) at the University of Illinois Urbana-Champaign for their feedback and input on this work. We also thank the anonymous reviewers for their thoughtful and valuable reviews.

## References

[1] 2016. Difference-in-Difference Estimation | Columbia Public Health. https://www.publichealth.columbia.edu/research/population-health-methods/difference-difference-estimation

[2] Alberto Abadie. 2005. Semiparametric difference-in-differences estimators. *The review of economic studies* 72, 1 (2005), 1–19. Publisher: Wiley-Blackwell.

[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] Hind Almerekhi, Supervised by Bernard J. Jansen, and co-supervised by Haewoon Kwak. 2020. Investigating Toxicity Across Multiple Reddit Communities, Users, and Moderators. In *Companion Proceedings of the Web Conference 2020 (WWW '20)*. Association for Computing Machinery, New York, NY, USA, 294–298. doi:10.1145/3366424.3382091

[5] 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

[6] 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.

[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] Vivek R. Athalye, Fernando J. Santos, Jose M. Carmena, and Rui M. Costa. 2018. Evidence for a neural law of effect. *Science* 359, 6379 (March 2018), 1024–1029. doi:10.1126/science.aao6058

[9] Sunyam Bagga, Andrew Piper, and Derek Ruths. 2021. "Are you kidding me?": Detecting Unpalatable Questions on Reddit. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 2083–2099. doi:10.18653/v1/2021.eacl-main.179

[10] Albert Bandura and Peter G. Barab. 1971. Conditions governing nonreinforced imitation. *Developmental Psychology* 5, 2 (1971), 244–255. doi:10.1037/h0031499 Place: US Publisher: American Psychological Association.

[11] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 830–839.

[12] L. Adlai Boyd, William S. Keilbaugh, and Saul Axelrod. 1981. The direct and indirect effects of positive reinforcement on on-task behavior. *Behavior Therapy* 12, 1 (Jan. 1981), 80–92. doi:10.1016/S0005-7894(81)80108-6

[13] 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.

[14] Mark Carman, Mark Koerber, Jiuyong Li, Kim-Kwang Raymond Choo, and Helen Ashman. 2018. Manipulating Visibility of Political and Apolitical Threads on Reddit via Score Boosting. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. 184–190. doi:10.1109/TrustCom/BigDataSE.2018.00037

[15] Jackie Chan, Charlotte Lambert, Fred 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. http://www.eshwarchandrasekharan.com/uploads/3/8/0/4/38043045/icwsm2024-community-resilience.pdf

[16] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit. In *ACM Transactions on Computer-Human Interaction*, Vol. 29. 1–26. doi:10.1145/3490499

[17] 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

[18] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 2. 1–25. doi:10.1145/3274301

[19] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, Denver Colorado USA, 3175–3187. doi:10.1145/3025453.3026018

[20] Jonathan P. Chang, Charlotte Schluger, and Cristian Danescu-Niculescu-Mizil. 2022. Thread With Caution: Proactively Helping Users Assess and Deescalate Tension in Their Online Discussions. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 6. 1–37. doi:10.1145/3555603

[21] Frederick Choi, Tanvi Bajpai, Sowmya Pratipati, and Eshwar Chandrasekharan. 2023. ConvEx: A Visual Conversation Exploration System for Discord Moderators. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 7. 262:1–262:30. doi:10.1145/3610053

[22] Frederick Choi, Charlotte Lambert, Vinay Koshy, Sowmya Pratipati, Tue Do, and Eshwar Chandrasekharan. 2025. Creator Hearts: Investigating the Impact Positive Signals from YouTube Creators in Shaping Comment Section Behavior. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, Yokohoma, Japan. doi:10.1145/3706598.3713521

[23] Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 71–80. doi:10.1609/icwsm.v8i1.14526 Number: 1.

[24] Cathlin V. Clark-Gordon, Nicholas D. Bowman, Alan K. Goodboy, and Alyssa Wright. 2019. Anonymity and Online Self-Disclosure: A Meta-Analysis. *Communication Reports* 32, 2 (May 2019), 98–111. doi:10.1080/08934215.2019.1607516

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

[26] Judith S Donath. 1998. Identity and deception in the virtual community. (1998).

[27] Thomas Erickson and Wendy A. Kellogg. 2000. Social translucence: an approach to designing systems that support social processes. *ACM Transactions on Computer-Human Interaction* 7, 1 (March 2000), 59–83. doi:10.1145/344949.345004

[28] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I always assumed that I wasn't really that close to [her]": Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 153–162. doi:10.1145/2702123.2702556

[29] Saqib Fayyaz, Sabahat Afsheen, and Adeel Khan. 2021. View of Impact of Positive Reinforcement Theory on Weightlifter's Performance. *International Journal of Physical Education and Sports Sciences* 5 (2021). https://journals.uol.edu.pk/the-sky/article/view/1004/778

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

[31] Eric Gilbert. 2012. Designing social translucence over social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Austin Texas USA, 2731–2740. doi:10.1145/2207676.2208670

[32] 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).

[33] P.O. Gray. 2010. *Psychology*. Worth. https://books.google.com/books?id=wiABQgAACAAJ

[34] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42. Publisher: HeinOnline.

[35] Nir Grinberg, P. Alex Dow, Lada A. Adamic, and Mor Naaman. 2016. Changes in Engagement Before and After Posting to Facebook. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 564–574. doi:10.1145/2858036.2858501

[36] Hussam Habib, Maaz Bin Musa, Fareed Zaffar, and Rishab Nithyanand. 2019. To Act or React: Investigating Proactive Strategies For Online Community Moderation. (June 2019). http://arxiv.org/abs/1906.11932 arXiv:1906.11932 [cs].

[37] 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.

[38] 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

[39] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[40] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3. 1–33. doi:10.1145/3359294

[41] 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

[42] Shagun Jhaver, Himanshu Rathi, and Koustuv Saha. 2024. Bystanders of Online Moderation: Examining the Effects of Witnessing Post-Removal Explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–9. doi:10.1145/3613904.3642204

[43] Jialun Aaron Jiang, Peipei Nie, Jed R. Brubaker, and Casey Fiesler. 2023. A Trade-off-centered Framework of Content Moderation. *ACM Transactions on Computer-Human Interaction* 30, 1 (Feb. 2023), 1–34. doi:10.1145/3534929

[44] Vigdis By Kampenes, Tore Dybå, Jo E. Hannay, and Dag I. K. Sjøberg. 2007. A Systematic Review of Effect Size in Software Engineering Experiments. *Information and Software Technology* 49, 11 (Nov. 2007), 1073–1086. doi:10.1016/j.infsof.2007.02.015

[45] Yoshihisa Kashima, Samuel Wilson, Dean Lusher, Leonie J. Pearson, and Craig Pearson. 2013. The acquisition of perceived descriptive norms as social category learning in social networks. *Social Networks* 35, 4 (Oct. 2013), 711–719. doi:10.1016/j.socnet.2013.06.002

[46] Alan E. Kazdin. 1973. The effect of vicarious reinforcement on attentive behavior in the classroom. *Journal of Applied Behavior Analysis* 6, 1 (1973), 71–78. doi:10.1901/jaba.1973.6-71

[47] Alan E. Kazdin. 1978. *History of behavior modification: Experimental foundations of contemporary research*. University Park Press, Baltimore, MD, US. Pages: xi, 468.

[48] 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.

[49] Charles Kiene and Benjamin Mako Hill. 2020. Who Uses Bots? A Statistical Analysis of Bot Usage in Moderation Teams. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3334480.3382960

[50] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building successful online communities: Evidence-based social design* (2012). Publisher: MIT Press, Cambridge, MA, USA.

[51] Gary King and Richard Nielsen. 2019. Why Propensity Scores Should Not Be Used for Matching. *Political Analysis* 27, 4 (Oct. 2019), 435–454. doi:10.1017/pan.2019.11

[52] René F. Kizilcec, Eytan Bakshy, Dean Eckles, and Moira Burke. 2018. Social Influence and Reciprocity in Online Gift Giving. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–11. doi:10.1145/3173574.3173700

[53] Peter Kollock et al. 1999. The Economies of Online Cooperation: Gifts and Public Goods in Cyberspace. *Communities in cyberspace* 239 (1999).

[54] Robert E. Kraut, Paul Resnick, Sara Kiesler, Moira Burke, Yan Chen, Niki Kittur, Joseph Konstan, Yuqing Ren, and John Riedl. 2011. *Building Successful Online Communities: Evidence-Based Social Design*. The MIT Press. http://www.jstor.org/stable/j.ctt5hhgvw

[55] Charlotte Lambert, Fred 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* CSCW (2024).

[56] Charlotte Lambert, Yoshee Jain, Koustuv Saha, and Eshwar Chandrasekharan. 2024. Investigating How Gilds Were Employed on Reddit. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*. 624–628.

[57] Cliff Lampe. 2012. The role of reputation systems in managing online communities. *H. Masum, M. Tovey, eds* (2012), 77–88.

[58] Cliff Lampe and Erik Johnston. 2005. Follow the (slash) dot: effects of feedback on new members in an online community. In *Proceedings of the 2005 ACM International Conference on Supporting Group Work (GROUP '05)*. Association for Computing Machinery, New York, NY, USA, 11–20. doi:10.1145/1099203.1099206

[59] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 543–550.

[60] Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. 2014. Crowdsourcing Civility: A Natural Experiment Examining the Effects of Distributed Moderation in Online Forums. *Government Information Quarterly* 31, 2 (2014), 317–326. doi:10.1016/j.giq.2013.11.005

[61] Rickey Lu. 2024. Audience design and pragmatic conceptions of moves and upvotes during advice-giving on Reddit. *Journal of Pragmatics* 219 (Jan. 2024), 30–47. doi:10.1016/j.pragma.2023.11.006

[62] Richard S. Lysakowski and Herbert J. Walberg. 1981. Classroom Reinforcement and Learning: A Quantitative Synthesis. *The Journal of Educational Research* 75, 2 (Nov. 1981), 69–77. doi:10.1080/00220671.1981.10885359

[63] Xiao Ma, Jeff Hancock, and Mor Naaman. 2016. Anonymity, Intimacy and Self-Disclosure in Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 3857–3869. doi:10.1145/2858036.2858414

[64] Michelle Marcus and Pedro H. C. Sant'Anna. 2021. The Role of Parallel Trends in Event Study Settings: An Application to Environmental Economics. *Journal of the Association of Environmental and Resource Economists* 8, 2 (2021), 235–275. doi:10.1086/711509 _eprint: https://doi.org/10.1086/711509.

[65] J. Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. In *Proceedings of the National Academy of Sciences*, Vol. 116. 9785–9789. doi:10.1073/pnas.1813486116

[66] Raymond G. Miltenberger. 2016. *Behavior modification: Principles and procedures, 6th ed.* Cengage Learning, Boston, MA, US. Pages: xxii, 665.

[67] Donn Morrison and Conor Hayes. 2013. Here, have an upvote: communication behaviour and karma on Reddit. In *GI-Jahrestagung*. https://api.semanticscholar.org/CorpusID:45454415

[68] Alexandra Olteanu, Onur Varol, and Emre Kiciman. 2017. Distilling the Outcomes of Personal Experiences: A Propensity-scored Analysis of Social Media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 370–386. doi:10.1145/2998181.2998353

[69] Eli Omernick and Sara Owsley Sood. 2013. The Impact of Anonymity in Online Communities. In *2013 International Conference on Social Computing*. 526–535. doi:10.1109/SocialCom.2013.80

[70] Orestis Papakyriakopoulos, Severin Engelmann, and Amy Winecoff. 2023. Upvotes? Downvotes? No Votes? Understanding the relationship between reaction mechanisms and political discourse on Reddit. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–28. doi:10.1145/3544548.3580644

[71] 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. Publisher: Elsevier.

[72] Sai Teja Peddinti, Keith W. Ross, and Justin Cappos. 2014. "On the internet, nobody knows you're a dog": a twitter case study of anonymity in social networks. In *Proceedings of the second ACM conference on Online social networks*. ACM, Dublin Ireland, 83–94. doi:10.1145/2660460.2660467

[73] 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.

[74] Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. 2020. Quick, Community-Specific Learning: How Distinctive Toxicity Norms Are Maintained in Political Subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 557–568. doi:10.1609/icwsm.v14i1.7323

[75] Harita Reddy and Eshwar Chandrasekharan. 2023. Evolution of Rules in Reddit Communities. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 278–282.

[76] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[77] Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. 2000. Reputation systems. *Commun. ACM* 43, 12 (2000), 45–48. Publisher: ACM New York, NY, USA.

[78] Paul Resnick, Richard Zeckhauser, John Swanson, and Kate Lockwood. 2006. The value of reputation on eBay: A controlled experiment. *Experimental economics* 9 (2006), 79–101. Publisher: Springer.

[79] Paul R Rosenbaum and Donald B Rubin. 1983. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)* 45, 2 (1983), 212–218. Publisher: Wiley Online Library.

[80] 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: Taylor & Francis _eprint: https://doi.org/10.1198/016214504000001880.

[81] Andrew M Ryan, Evangelos Kontopantelis, Ariel Linden, and James F Burgess. 2019. Now trending: Coping with non-parallel trends in difference-in-differences analysis. *Statistical Methods in Medical Research* 28, 12 (Dec. 2019), 3697–3711. doi:10.1177/0962280218814570 Publisher: SAGE Publications Ltd STM.

[82] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and Psychological Effects of Hateful Speech in Online College Communities. In *Proceedings of the 10th ACM Conference on Web Science*. ACM, Boston Massachusetts USA, 255–264. doi:10.1145/3292522.3326032

[83] Koustuv Saha, Yozen Liu, Nicholas Vincent, Farhan Asif Chowdhury, Leonardo Neves, Neil Shah, and Maarten W Bos. 2021. Advertiming matters: Examining user ad consumption for effective ad allocations on social media. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

[84] 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

[85] Koustuv Saha, Benjamin Sugar, John Torous, Bruno Abrahao, Emre Kıcıman, and Munmun De Choudhury. 2019. A Social Media Study on the Effects of Psychiatric Medication Use. *Proceedings of the International AAAI Conference on Web and Social Media* 13 (July 2019), 440–451. doi:10.1609/icwsm.v13i01.3242

[86] Koustuv Saha, Ingmar Weber, and Munmun De Choudhury. 2018. A Social Media Based Examination of the Effects of Counseling Recommendations after Student Deaths on College Campuses. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (June 2018). doi:10.1609/icwsm.v12i1.15016

[87] 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.

[88] Brennan Schaffner, Arjun Nitin Bhagoji, Siyuan Cheng, Jacqueline Mei, Jay L Shen, Grace Wang, Marshini Chetty, Nick Feamster, Genevieve Lakier, and Chenhao Tan. 2024. "Community Guidelines Make this the Best Party on the Internet": An In-Depth Study of Online Platforms' Content Moderation Policies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3613904.3642333

[89] Rebecca L. Schaumberg and Samuel E. Skowronek. 2022. Shame Broadcasts Social Norms: The Positive Social Effects of Shame on Norm Acquisition and Normative Behavior. *Psychological Science* 33, 8 (Aug. 2022), 1257–1277. doi:10.1177/09567976221075303 Publisher: SAGE Publications Inc.

[90] Jodi Schneider, Bluma S. Gelley, and Aaron Halfaker. 2014. Accept, decline, postpone: How newcomer productivity is reduced in English Wikipedia by pre-publication review. In *Proceedings of The International Symposium on Open Collaboration*. ACM, Berlin Germany, 1–10. doi:10.1145/2641580.2641614

[91] 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

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

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

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

[95] Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3. 1–21. doi:10.1145/3359265

[96] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. doi:10.1145/3411764.3445092

[97] Laurence Steinberg, Susie D Lamborn, Sanford M Dornbusch, and Nancy Darling. 1992. Impact of parenting practices on adolescent achievement: Authoritative parenting, school involvement, and encouragement to succeed. *Child development* 63, 5 (1992), 1266–1281. Publisher: Wiley Online Library.

[98] Rafal Urbaniak, Patrycja Tempska, Maria Dowgiałło, Michał Ptaszyński, Marcin Fortuna, Michał Marcińczuk, Jan Piesiewicz, Gniewosz Leliwa, Kamil Soliwoda, Ida Dziublewska, Nataliya Sulzhytskaya, Aleksandra Karnicka, Paweł Skrzek, Paula Karbowska, Maciej Brochocki, and Michał Wroczyński. 2022. Namespotting: Username toxicity and actual toxic behavior on Reddit. *Computers in Human Behavior* 136 (Nov. 2022), 107371. doi:10.1016/j.chb.2022.107371

[99] Chris Vargo, Gina Masullo, and Tobias Hopp. 2024. Deciding to Delete Posts on Reddit: What Factors Influence Content Removal. https://hdl.handle.net/10125/106696

[100] Gaurav Verma, Ankur Bhardwaj, Talayeh Aledavood, Munmun De Choudhury, and Srijan Kumar. 2022. Examining the impact of sharing COVID-19 misinformation online on mental health. *Scientific Reports* 12, 1 (May 2022), 8045. doi:10.1038/s41598-022-11488-y

[101] Elaine Wallace and Isabel Buil. 2021. Hiding Instagram Likes: Effects on negative affect and loneliness. *Personality and Individual Differences* 170 (Feb. 2021), 110509. doi:10.1016/j.paid.2020.110509

[102] 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

[103] Jess Weatherbed. 2024. Reddit brings back its old award system — 'we messed up' - The Verge. https://www.theverge.com/2024/5/17/24158848/reddit-brings-back-award-system-gold-coins-messed-up

[104] Leong Teen Wei and Rashad Yazdanifard. 2014. The impact of Positive Reinforcement on Employees' Performance in Organizations. *American Journal of Industrial and Business Management* 04, 01 (2014), 9–12. doi:10.4236/ajibm.2014.41002

[105] Robert Weinberg, Howard Garland, Lawrence Bruya, and Allen Jackson. 1990. Effect of Goal Difficulty and Positive Reinforcement on Endurance Performance. *Journal of Sport and Exercise Psychology* 12, 2 (June 1990), 144–156. doi:10.1123/jsep.12.2.144

[106] Galen Weld, Peter West, Maria Glenski, David Arbour, Ryan A. Rossi, and Tim Althoff. 2022. Adjusting for Confounders with Text: Challenges and an Empirical Evaluation Framework for Causal Inference. *Proceedings of the International AAAI Conference on Web and Social Media* 16 (May 2022), 1109–1120. doi:10.1609/icwsm.v16i1.19362

[107] Galen Weld, Amy X. Zhang, and Tim Althoff. 2024. Making Online Communities 'Better': A Taxonomy of Community Values on Reddit. *Proceedings of the International AAAI Conference on Web and Social Media* 18 (May 2024), 1611–1633. doi:10.1609/icwsm.v18i1.31413

[108] Diane M. Wiese, Maureen R. Weiss, and David P. Yukelson. 1991. Sport Psychology in the Training Room: A Survey of Athletic Trainers. *The Sport Psychologist* 5, 1 (March 1991), 15–24. doi:10.1123/tsp.5.1.15

[109] Ryan Yen, Li Feng, Brinda Mehra, Ching Christie Pang, Siying Hu, and Zhicong Lu. 2023. StoryChat: Designing a Narrative-Based Viewer Participation Tool for Live Streaming Chatrooms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3544548.3580912

[110] Yunhao Yuan, Koustuv Saha, Barbara Keller, Erkki Tapio Isometsä, and Talayeh Aledavood. 2023. Mental Health Coping Stories on Social Media: A Causal-Inference Study of Papageno Effect. In *Proceedings of the ACM Web Conference 2023*. ACM, Austin TX USA, 2677–2685. doi:10.1145/3543507.3583350