CORRECTED VERSION OF RECORD; SEE LAST PAGE OF ARTICLE

https://pubsonline.informs.org/journal/mksc

MARKETING SCIENCE
Articles in Advance, pp. 1–25
ISSN 0732-2399 (print), ISSN 1526-548X (online)

How Do Content Producers Respond to Engagement on Social Media Platforms?

Simha Mummalaneni,[a,*] Hema Yoganarasimhan,[b] Varad Pathak[c]

[a] William & Mary, Williamsburg, Virginia 23187; [b] University of Washington, Seattle, Washington 98195; [c] Independent Contributor, Menlo Park, California 94025

*Corresponding author

Contact: nmummalaneni@wm.edu, https://orcid.org/0000-0003-3364-2088 (SM); hemay@uw.edu, https://orcid.org/0000-0003-0703-5196 (HY); pathak.varad@gmail.com (VP)

Received: April 26, 2023
Revised: August 12, 2024; September 24, 2025
Accepted: December 18, 2025
Published Online in Articles in Advance: February 26, 2026, and updated March 2, 2026

https://doi.org/10.1287/mksc.2023.0178

Abstract. When creating new content, many social media users hope to receive engagement from other users. This research examines how receiving that engagement affects different users' subsequent behavior on the platform. We address this question through a field experiment on Twitter in which some users' posts were purposefully shown more often to other users, which (on average) increased the amount of engagement they received. We estimate a doubly robust instrumental variable model that allows us to estimate individual-level treatment effects, and we find substantial heterogeneity across users in terms of how they respond to additional engagement: most users do not significantly change their behavior, but some users respond by substantially increasing their time spent on the platform, posting more content, and engaging more with other users' content. Users who respond most strongly are systematically different than the rest of the user base on observable pre-experiment user metrics, thereby providing substantive insights about which users value engagement very highly. Our results demonstrate how social media platforms can increase content creation, content consumption, and overall usage of the platform by focusing on this group of users and targeting them with interventions that are intended to increase the amount of engagement they receive.

History: Tat Chan served as the senior editor for this article.
Supplemental Material: The online appendices and data files are available at https://doi.org/10.1287/mksc.2023.0178.

Keywords: two-sided platforms • social media • customer engagement • causal machine learning

1. Introduction

Social media has become an integral part of our daily lives, with billions of people using various platforms to connect, share information, and stay informed. According to recent statistics, as of January 2023, there are approximately 4.6 billion active social media users worldwide (Kemp 2023). A unique aspect of social media platforms is that users are both content creators and consumers in this setting, and the platform serves as an intermediary. This is in contrast to other digital media platforms like Spotify or Netflix, where the content is created or licensed by the platform and users mainly function as consumers of the content. The long-run success of social media platforms thus depends on their ability to encourage users to use the platform more, both to consume more content and to produce more content. There are a couple of ways in which the platform can do this directly: by paying users directly for their content[1] or by giving them attention or recognition (Huang and Narayanan 2020, Burtch et al. 2022, Lu et al. 2023). Another natural channel that can potentially motivate users to increase usage of the platform is if they receive peer engagement on their content (e.g., favorites, likes, or retweets) provided by other consumers on the platform (Restivo and van de Rijt 2014, Eckles et al. 2016, Gallus et al. 2020), which is the focus of this study.

In this paper, we study how content producers respond to increased engagement on social media platforms and how the platform can leverage this information to improve platform usage, content production, and content consumption.

We combine data from a large-scale field experiment on Twitter with a doubly robust instrumental variable (DRIV) approach to address these questions. The first component of our approach consists of a large-scale field experiment on Twitter consisting of 4.9 million users. Unlike direct payment or recognition, engagement is an intervention that is not directly under the platform's control, because engagements are given by consumers and not the platform. Therefore, we adopt an encouragement design (Messing 2013), where we exogenously shift the opportunity to get incremental engagements. Twitter users in the experiment were

randomly assigned into either a control group or a "boost" condition, which artificially increased the relevance scores of their tweets for a two-week period. As a result, boosted users' content was shown more often and more prominently than it otherwise would have been during this time, which in turn led to increased engagement on their content. We also collect data on a rich set of user-level pretreatment variables based on their demographics and usage in the two weeks prior to the experiment. Finally, we track their post-treatment behavior for a two-week period on a variety of metrics: minutes active on the platform, number of days active on the platform, content produced (tweets composed), and the engagement given to other users' content.

The simplest way to analyze the experiment's data are to compare time spent on the platform (which is a metric directly relevant to ad revenue) between users who were assigned to the boost condition versus the control condition. We find that this intent-to-treat (ITT) estimate is positive and significant, that is, being boosted has a positive effect on the time spent on the platform. However, ITT estimates in two-sided markets suffer from two fundamental drawbacks. First, they lack counterfactual validity; that is, these estimates do not represent what would happen if we applied the boost condition to all the producers in the system. Indeed, if all the producers' items were equally boosted, the resulting rank ordering would be the same as the baseline case where all producers are in the control condition. Second, the ITT estimate does not represent the incremental benefit of a producer receiving one additional engagement.

Therefore, we next consider a two-stage least squares (2SLS) model where we use the experiment bucket (control or boost) as an instrument that exogenously varies the actual treatment (number of engagements) but does not directly influence the outcome. This estimation procedure gives us a positive average treatment effect (ATE), which implies that one incremental engagement leads to an increase in time spent (minutes active) over the two-week post-treatment period. However, 2SLS estimates are consistent only under two conditions: (1) constant treatment effect and (2) no heterogeneity in treatment intensity (Syrgkanis et al. 2019). The first requires that the treatment effect is constant across all users. This is unlikely to be true in our setting since social media users are likely to be highly heterogeneous in how much they value incremental engagement. The second condition fails because the treatment intensity (the number of additional engagements received) is not equal across all users who were assigned to the boost condition. For example, people who tweet a lot will get more engagement than those who tweet less because the former will have more items that get boosted. A final challenge, similar to the ITT estimate, the ATE (even if consistent) is not particularly useful from a managerial perspective because any targeted intervention from the platform's side requires it to know which users would be the most responsive to incremental engagement; that is, it needs individual-level estimates of the causal impact of incremental engagement.

To address these challenges, we employ a DRIV method that allows for both heterogeneous treatment intensity and heterogeneous treatment effects in the estimation and provides individual-level conditional average treatment effect (CATE) estimates (Syrgkanis et al. 2019). This approach has all the standard properties of double machine learning methods (Chernozhukov et al. 2018). Furthermore, the CATE value is allowed to be a flexible function of user-specific pretreatment variables learned from the data.

Estimating heterogeneous treatment effects means that we can examine how much different kinds of users increase their activity in response to receiving additional engagement on their content. We find that most users on the platform demonstrate a weak response, because the average user increases their time spent on the platform by about 8.4 seconds in response to one incremental engagement. However, there is a long right tail for this metric—The top 1% of users each increases their time on the platform by more than 2.12 minutes (roughly 19 times larger than the median). Users who respond most strongly are systematically different than the rest of the user base on observable pre-experiment user metrics, thereby providing substantive insights about which users value engagement very highly. The kinds of users who respond to engagement by significantly improving their time spent on the platform tend to be connected with others (more followers and also following more accounts), have older accounts, but do not use the platform as regularly.

Next, we examine the source of this incremental time spent: whether it comes from increased content production, increased content consumption, or both. To that end, we estimate the DRIV models on two separate outcome variables: (1) the number of tweets composed, which is a pure production-focused measure, and (2) the number of favorites given, a pure consumption-focused metric. Overall, we find that incremental engagement encourages users on both fronts: They produce more content and engage more with others' content, but the relative effect size is larger for the production outcome. This is understandable because they have just received positive engagement with the content they have recently produced.

Furthermore, we examine the impact of engagement on two other outcome variables of interest to the platform. First, we consider monetizable active days, which is the number of days a user is active on the platform. This metric captures the regularity of platform usage. We find that receiving incremental engagement causes

users to increase their monetizable active days, but the improvements are smaller in magnitude and in percentage terms compared with the results for minutes active. This result is partially driven by the fact that most users are already active every day, and therefore they cannot be improved on this metric. Second, we consider the total engagements given to other users as the outcome variable of interest. This metric captures the spillover effects of giving an incremental engagement to a focal user. Again, we find a positive effect on this metric, which suggests a "virtuous cycle," that is, a user who receives engagement may go on to provide engagement for a second user, who in turn may provide engagement for a third user, and so on.

Finally, we examine the returns to using our approach to identify which producers to target for interventions that lead to increased engagement. We focus on quantifying the total gain in activity for the platform if it were to target users based on four possible criteria: users in the top 1% of CATE values for minutes active, users in the top 1% of CATE values for engagement given, top 1% of users based on user activity (in the pretreatment period), and bottom 1% of users based on user activity (in the pretreatment period). The first two criteria are based on our DRIV models and are intended to focus on groups of users who are more responsive to incremental engagement, whereas the latter two criteria are commonly used activity heuristics that allow the platform to focus on different subsets of their user base.

We find that targeting users with high CATE values is more effective than targeting users based on heuristics that use activity-level thresholds; that is, the two former approaches yield bigger improvements in total minutes active. Between the two CATE-based approaches, we find substantial benefits to targeting users in the top 1% of CATE values for minutes active rather than engagement given. This pattern holds true if we consider the improvement among the users who are targeted, but also if we consider the improvement among the entire platform due to targeted users providing positive spillovers to others. Overall, this suggests that there is substantial value in using our CATE estimates (that are flexible functions of all the pretreatment variables) compared with using heuristic thresholds that are commonly used in the industry.

In summary, our paper provides a few key contributions to the literature on social media and user behavior in two-sided platforms. First, from a methodological perspective, our framework (that combines boosting experiments and doubly robust instrumental variable estimation) is quite general and be used by a wide variety of social media platforms for similar purposes. From a substantive perspective, we show that, when users receive increased engagement with their social media content, they subsequently change their activity on the same platform on a variety of metrics: They spend more time on the platform, they create more content, they engage more with other users' content, and they increase the regularity of their platform usage. Importantly, there is significant heterogeneity in user responsiveness on all these metrics, with a long right tail of users who exhibit high responsiveness. From a managerial perspective, we show that the firm can achieve substantial improvements in platform usage by targeting users based on the estimates from our approach compared with standard baselines.

2. Related Literature

Our paper relates to the literature on usage patterns in social media platforms. Papers in this stream usually try to quantify users' incentives to create content and interact with others' content. Early research on this topic focused on the effect of social ties on content generation. Using observational data, Shriver et al. (2013) show that a user's social ties have a positive effect on their propensity to create content and vice versa. They use an instrumental variable approach to control for the fact that ties and posts are often codetermined/endogenous. Toubia and Stephen (2013) examine Twitter users' incentives to create content using data from a field experiment where they randomly add fake followers to some accounts. Using this exogenous variation in follows, they show that, although users get both intrinsic and image-related utility from posting, the latter plays a bigger role. Ahn et al. (2016) develop a forward-looking structural model of user-generated content production and consumption and consider counterfactuals where the platform can sponsor content. Guo et al. (2023) show that the amount of information in the early content posted on a platform can have a negative impact on the quantity of future knowledge content but a positive effect on the diversity of the content.

A separate stream of research has focused on the effect of visible interventions from the platform such as featuring or publicly recognizing users' content. Huang and Narayanan (2020) and Burtch et al. (2022) show that such increased attention and recognition can have a positive effect on users' subsequent content production. On the other hand, Lu et al. (2023) find that recognition (a digital badge awarded by the platform) leads to increased content generation but reduces content consumption immediately after receiving the award, although the longer-term effects are positive on both outcomes.[2] In contrast to these papers, we focus on how users respond to peer engagement on their content (e.g., favorites, retweets, and replies), which is substantively different from the recognition by the platform. Further, because the platform cannot directly provide engagement, there are a set of additional methodological challenges both in terms of estimation and intervention that we need to address (see Section 4.1 for details).

A related stream of work focuses on peer feedback and peer recognition (as opposed to platform recognition), and the results here are mixed. Restivo and van de Rijt (2014) and Gallus et al. (2020) directly manipulate the peer feedback received by contributors on forums and find that there is no significant effect on post-treatment user activity. Given the relatively small size of their samples, this inconclusive result could be because the true effects are small but positive on average, or it could be because users have heterogeneous effects that yield a near-zero average treatment effect. More closely related to our work, Eckles et al. (2016) adopt an encouragement design (similar to ours) and find that receiving feedback or recognition from one's peers has a positive average downstream effect. Our research contributes to this literature in a few ways. First, we are able to speak to this debate and show that the average treatment effect of peer engagement is positive (in our setting). Second, in addition to content generation, we examine multiple dimensions of user behavior: production, consumption, time spent, regularity of usage, and engagements given. Third, unlike the previous papers, we are able to quantify user-level treatment effects for peer engagement, thereby showing that there is heterogeneity in the value that users derive from peer engagement. These treatment effects can be used by the platform to identify and target users who are the most responsive to engagement through interventions in the user interface or in a two-sided recommendation system that balances both producer and consumer engagement. Finally, from a methodological perspective, our estimation task has significant additional challenges compared with the earlier literature since we need to account for endogenous treatment intensity and individual-level heterogeneity in our analysis (as discussed in Section 4.2).

More broadly, our paper relates to the growing marketing literature on the customization of digital products and promotions using machine learning methods (see Rafieian and Yoganarasimhan (2023) for a detailed overview). The main difference between this literature and our paper is that our experimental data has an intent to treat structure, which makes the estimation of personalized treatment effects more challenging. Finally, our paper also contributes to the literature on how to design optimal recommendation systems in both computer science and marketing (Falk 2019, Yoganarasimhan 2020, Liu et al. 2021). Specifically, our producer-level treatment effects can serve as inputs into a recommendation system that incorporates both consumer-specific engagement scores and also individual producer-level utilities. Thus, it can be used to balance the platform's goals of increasing activity among both content producers and content consumers.

3. Setting

On social media platforms, users both create content and consume (read, or watch) content produced by other users. Typically, when consuming content, users on social media platforms do not see all the content that is available on the platform. Instead, the platform serves as the intermediary by helping consumers to find content that they might enjoy and also helping producers to find an audience for their content. To accomplish this, social media platforms like Twitter have a recommendation system that ranks items that could potentially be shown to the consumer.[3] These recommendation systems are usually focused on maximizing consumer utility. Platforms typically cannot measure consumer utility directly, so many platforms instead use consumer engagement as their measure for a positive consumer outcome. We consider a situation in which the platform observes consumer engagement metrics such as favorites or sharing behavior (e.g., retweets and replies). At a given time, there are I different items that could be shown to consumers. In our setting, for each specific item $i \in I$ and consumer c, the platform calculates the following relevance score:

$$s_{ic} = \mathbb{P}(c \text{ engages with } i | i \text{ is shown to } c). \quad (1)$$

Higher values of s_{ic} imply that the consumer is more likely to engage with a particular item, which therefore suggests that it should be shown to them. Typically, the platform estimates s_{ic} in real time for each $\{i,c\}$ combination using a broad scope of information including the item's popularity, the item's genre or topic, the item producer's prior popularity, the consumer's prior engagement history, whether the item is currently very popular, and so on. The prediction problem itself is typically estimated using standard supervised machine learning methods (e.g., boosted trees). The specific variables, methods, and algorithms used to make this prediction are beyond the scope of this paper—These are typically idiosyncratic to the specific platform, and we treat them as fixed for the purpose of our research.[4] Once the platform has predicted s_{ic} for each $\{i,c\}$, it shows the items I to the consumer c in descending order based on their relevance scores. On Twitter, this was known as the "Top Tweets" option on the user's Home timeline at the time of the experiment.

3.1. Field Experiment

To understand how receiving engagement as a producer affects the user's subsequent activity on the platform, the platform would like to understand how much each producer p would increase their activity a_p if the number of engagements received (denoted by e_p) went up by one:

$$\theta_p = \mathbb{E}[a_p | e_p + 1] - \mathbb{E}[a_p | e_p]. \quad (2)$$

The θ_p term can be interpreted as each producer's incremental utility from receiving one additional

engagement on the platform (over a baseline engagement of e_p). The platform typically has information about the producer's activity and how many engagements each producer has received. However, predicting θ_p is difficult with existing archival data because the number of engagements e_p is endogenous. One concern is the possibility of reverse causation. For example, producers who are highly active and produce quality content may be spending a lot of time on the platform and at the same time may also receive a lot of engagement on their content. In such cases, it is the quality and activity of the producer that is driving the downstream consumer engagement rather than vice versa. Another possibility is that a high level of activity (a_p) and a high level of engagement (e_p) could be jointly caused by a common driver such as an important current event or discussions of topics that are currently popular.

Given these endogeneity issues, a cleaner approach would be for the platform to run an experiment and estimate θ_p from the experimental results. For instance, the platform could randomize the number of engagements that each producer receives and then observe how that affects subsequent user-level activity. However, the engagements are decided by consumers and are not directly under the platform's control, so the only way to randomize them directly would be by lying or making up false engagement numbers. We treat this as being an unacceptable option for the platform.

To sidestep these concerns, we adopt a peer encouragement experimental approach that exogenously varies how often and how prominently each producer gets shown to consumers. Similar designs have previously been used by Messing (2013) and Eckles et al. (2016). In our setting, the main difference is that the boost condition increases the probability of being shown at the producer level rather than employing interventions at the consumer or edge level. See Angrist et al. (1996) and Bradlow (1998) for a more general discussion of encouragement designs in randomized experiments.

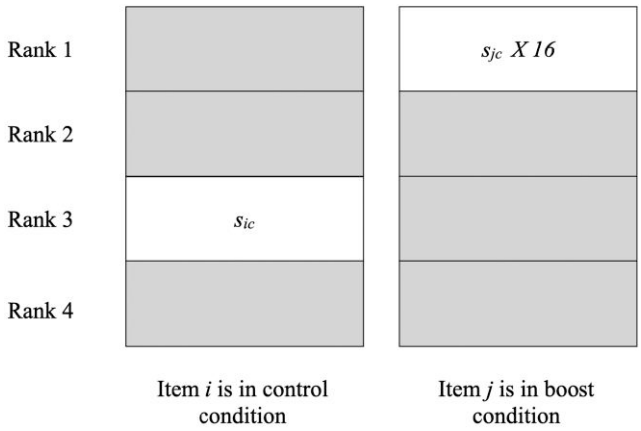
We conduct a large-scale field experiment on Twitter consisting of approximately 4.9 million users over a two-week period in 2021. Each user was randomly assigned to one of two different buckets: 33.33% were assigned to a boosted group and the remaining 66.6% were assigned to the control group. Tweets by users (producers) in the boosted group were boosted by a factor of 16 (i.e., their scores s_{ic} were multiplied by 16) in the recommendation system, which means their items rise in the rankings, and therefore be seen more often by consumers and seen at higher positions.[5] Tweets of users in the control group were not boosted (i.e., their scores s_{ic} remain untouched), and as a result, they do not rise in the rankings. Figure 1 presents a pictorial depiction of the experiment design. Producers do not directly observe how often their items are shown to consumers[6]; instead, they only observe the number of engagements they receive. Note that this experimental design allows us to treat the bucket (boost versus control) as the exogenous instrument (Z) because the bucket exogenously shifts the intensity of treatment (amount of engagement received), but has no effect on the outcome directly. In the rest of the paper, we use the term "boosted group" instead of "treatment group" to refer to users who were boosted because treatment in our context refers to the engagement received, and users in both the boosted and control groups receive engagements (although the magnitude of engagements differ across the two groups). Therefore, we do not use the term "treated group" to avoid confusion around the definition of treatment.

Data for the analysis come from the three time periods, as shown in Figure 2 and discussed below:

- Pretreatment period: August 19–September 1, 2021. This is the two-week period before the experiment. We use it to generate user features (X_p) that can be used both to model the heterogeneity in getting engagement (during the treatment period) as well for estimating heterogeneous treatment effects.
- Treatment period: September 2–September 17, 2021. The two-week period during which users' tweets are boosted by 16 times if they are in the boost condition. The control group's tweets do not get any additional boost factor.
- Post-treatment period: September 18–October 1, 2021. This is the period when we measure the impact of the treatment on producers' activity.

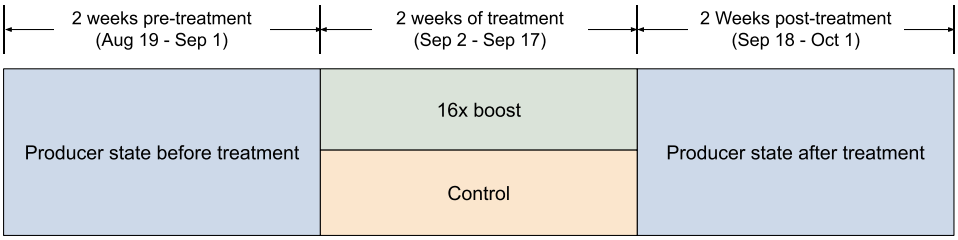
Before we proceed to data and analysis, we make note of two points about the experiment design. First, users were added to the experiment only if they

Figure 1. Design of the Boosting Experiment



Notes. In this example, users i and j are in the control and boost conditions, respectively. The items created by user i do not receive any boost to their scores. In contrast, the scores of all the items created by user j are multiplied by 16, and as a result, j's items are likely to be shown at higher ranks (and hence be more visible to consumers).

Figure 2. (Color online) Timeline for the Experiment



tweeted at least once during the experiment period (September 2–17). For example, if a user tweeted for the first time on September 15, then this user was considered eligible to be added to the experiment—and if assigned to the boost condition, they saw all their posts boosted by the 16 times factor for the rest of the experiment. Essentially, this implies users who were not active during the experiment period are not in the experiment (in either the control or boost conditions).[7] Second, in principle, our experiment design can give rise to stable unit treatment value assumption (SUTVA) violations. That is, when we boost one set of users, it can have a negative effect on the ranking of users in the control group. This type of SUTVA violation (or interference problem) makes a simple boost versus control comparison flawed. However, because the fraction of users boosted was very small (approximately 0.8% of the overall population of active users on Twitter during this period), these interference effects are minimal in our context. That said, our main analysis only requires that the buckets (boost and control) be exogenously assigned, and it allows the actual engagement to be endogenous to the producer and/or the experiment. Thus, even if there were interference effects that influenced the ranking of control users, our estimates on the incremental impact of an additional engagement on the downstream activity would still be valid.

3.2. Summary Statistics

In this section, we present the summary statistics of the pretreatment variables or producer features (X_p), the realized treatment variable, that is, engagement received by the producer (e_p), and the outcome variable, which is the activity of the producer in the post-treatment period (a_p).

Table 1 presents the summary statistics of the producer features X_p by the experimental bucket (or instrument Z). These features represent the cumulative activity and engagement of the user during the two-week pretreatment period, as well as some user-specific features. Notice that almost all the variables have large standard deviations. This is because the distribution of these variables has a heavy right tail; that is, there are some very heavy users whose high levels of usage and activity mask the lower activity levels of the vast majority of users. Therefore, the median is a better measure of the central tendency of the data in this case, and we will use that in the rest of the discussion.

First, we see that the median user in the experiment spent 291 minutes on the platform in the two-week pretreatment period. During this time, they composed around six tweets and received over three favorites, zero retweets, and one reply. Further, the median user made one follow, gave 20 favorites, and sent one retweet and one reply in this two-week period. We also have two metrics that capture the regularity with which consumers use the platform: active days and monetizable active days. The former is the number of days a user logged into the platform over the two-week pretreatment period. The latter is the number of days a user logged in and provided nonzero ad revenue in the same period.[8] The median for these metrics is 14, which suggests that the population consists of fairly active and regular users. Next, we discuss the user features. The median user in the experiment has been with the platform for 3.64 years, has 107 followers, and is following 189 other accounts.

In addition, we observe two categorical user-level variables that capture producer heterogeneity: user state and country. The former is a variable created by Twitter that is a summary measure of the recent usage and activity of the user. These user states can take eight possible values, and Table 2 shows the distribution of values for the boosted and control groups at the start of the experiment. We see that more than 44% of the users are heavy tweeters, whereas 26% are heavy users who are non-tweeters (who predominantly consume others' content rather than producing content themselves). The rest of the six states are less prevalent and form the remaining 30% of the users in the experiment. Next, Table 3 shows the distribution of users' countries in the control and boosted groups. We see that the distribution of users' countries of origin broadly follows the standard distribution of Twitter usage/adoption across the globe. Users from Japan and United States make up around 37% of the experiment, whereas the rest of the countries have a relatively small presence.

Next, we present the summary statistics of the treatment variable (e_p) (the total engagement received

Table 1. Summary Statistics of Pretreatment Producer Features (X_p) for the Control and the Boosted Groups in the 14-Day Period Before the Experiment

Variable	Mean	Standard deviation	25th percentile	Median	75th percentile
Control group					
Tweets composed	41.51	343.56	1	6	25
Favorites received	202.89	9,277.49	0	3	25
Retweets received	36.54	2,419.55	0	0	1
Replies received	21.45	447.46	0	1	6
Follows received	17.44	341.14	0	1	4
Follows made	10.78	57.42	0	1	5
Favorites given	155.04	495.80	2	20	100
Retweets sent	31.45	184.89	0	1	10
Replies sent	21.08	110.52	0	1	9
Quote retweets sent	3.59	22.6	0	0	1
Minutes active	622.34	914.86	74	291	804
Active days	11.73	3.9	11	14	14
Monetizable active days	11.31	4.31	1	14	14
Active followers	606.25	13,611.71	19	70	218
Total followers	1,114.48	50,241.14	24	99	330
Following	442.86	2,030.24	64	181	435
Indicator push notification enabled	0.73	0.44	0	0	1
Account age (days)	1,723.56	1,430.48	430.32	1,328.87	2,995.28
Boosted group					
Tweets composed	41.28	164.03	1	7	25
Favorites received	200.89	9,430.95	0	4	25
Retweets received	34.52	149.19	0	0	1
Replies received	21.62	381.79	0	1	6
Follows received	17.77	651.91	0	1	4
Follows made	10.82	57.81	0	1	5
Favorites given	155.46	498	2	20	102
Retweets sent	31.71	190.05	0	1	10
Replies sent	21.24	112.94	0	1	9
Quote retweets sent	36.08	21.57	0	0	1
Minutes active	623.28	935.65	74	292	805
Active days	11.73	3.89	1	14	14
Monetizable active days	11.32	4.30	1	14	14
Active followers	624.36	16,004.53	19	70	218
Total followers	1,173.45	57,376.16	24	99	330
Following	441.14	1,711.09	64	181	435
Indicator push notification enabled	0.73	0.44	0	1	1
Account age (days)	1,724.44	1,430.12	431.52	1,328.83	2,997.89

during the treatment period) in Table 4. We define the total engagement received as the sum of replies, retweets, and favorites received during the treatment period.[9] We see that producers in the boosted condition received a lot more engagement than those in the control condition, which suggests that the experiment successfully yielded exogenous variation in the likelihood of receiving engagement.

4. Preliminary Analysis

We now present some preliminary analysis based on the experimental data. First, we calculate an ITT treatment

Table 2. Distribution of User States at the Time of Entry into the Experiment for the Control and Boosted Groups

User state	Control	Percent of control	Boosted	Percent of boosted
Heavy tweeter	1,449,126	44.60%	724,457	44.67%
Heavy non-tweeter	845,820	26.03%	422,060	26.02%
Medium tweeter	353,954	10.89%	177,518	10.94%
Medium non-tweeter	233,844	7.20%	116,338	7.17%
Light	148,151	4.56%	73,575	4.5%
Very light	110,452	3.40%	54,762	3.38%
Near zero	31,160	0.96%	15,526	0.96%
New	24,589	0.76%	12,064	0.74%
Total	3,249,367	100%	1,622,227	100%

Table 3. Distribution of Users' Country of Origin in the Control and Boosted Groups

Country	Control	Percent of control	Boosted	Percent of boosted
Japan	660,755	20.33%	328,835	20.27%
United States	563,974	17.36%	281,969	17.38%
Brazil	243,027	7.48%	121,906	7.51%
Philippines	146,040	4.49%	72,366	7.17%
United Kingdom	134,194	4.13%	67,072	4.46%
Indonesia	126,689	3.90%	63,205	3.90%
Turkey	117,751	3.62%	58,694	3.62%
South Africa	104,386	3.21%	52,397	3.23%
Mexico	94,168	2.90%	46,771	2.88%
Argentina	8,130	2.50%	40,529	2.50%
Rest of the world	2,199,114	67.68%	1,133,744	69.89%
Total	3,249,367	100%	1,622,227	100%

effect by comparing average outcomes among producers who were assigned to the boost versus control conditions. Next, we discuss the challenges in interpreting and using the ITT estimates in counterfactual policy design. Then, we use the bucket assignment as an instrument and estimate a 2SLS model. This yields a preliminary estimate of the ATE.

4.1. ITT Estimates

The first model in Table 5 shows the effect of the experiment condition on the intensity of treatment received. We see that being in the boosted condition does lead to a significant increase in engagement received. On average, boosted producers received more than 455 more engagements than producers in the control condition during the two-week treatment period.

Next, we use the bucket assignment to estimate a simple ITT treatment effect on the main outcome variable of interest: the number of minutes the user spent on the platform in the two-week period following the experiment. This variable is most closely aligned with the platform's monetization goals: The longer a user spends on the platform, the higher the ad revenue from the user. This ITT regression is the simplest analysis of the experimental data and compares the post-treatment activity levels for producers in the boosted versus control groups. Formally, because we are randomly assigning producers to the two groups, the ITT treatment effect (α_1) can be estimated using the following model:

$$a_p = \alpha_0 + \alpha_1(\text{boost condition})_p + \varepsilon_p. \quad (3)$$

The estimates from this analysis are shown in Model 2 of Table 5. We find that being in the boosted condition has a positive effect on users' post-treatment activity. Users in the boosted condition spend approximately six more minutes on the platform (i.e., ≈1% more time) than those in the control condition. This indicates that, on average, giving producers more engagement has a positive effect on their future time spent on the platform.[10]

Although this 1% improvement estimate shows that being boosted has a positive impact on producers' overall activity in the post-treatment period, it does not provide us with the causal impact of giving more engagement on producers' future usage. Indeed, these estimates lack counterfactual validity and cannot be mapped to marginal treatment effects. We discuss these challenges in detail below.

4.1.1. Counterfactual Validity and Policy Design. A fundamental challenge with the ITT estimates is their lack of counterfactual validity; that is, these estimates do not represent what would happen if we applied the boost condition to all the producers in the system. Indeed, if all the producers' items were equally boosted (i.e., their scores were multiplied by 16 or any positive number), then the end result would be that no one is boosted: the resulting rank ordering would be the same as the baseline case where all producers are in the control condition. The ITT estimates simply give us the incremental effect on a producer's activity when there is a small portion of the producers being boosted. As the fraction of producers getting the boost treatment increases, the ITT estimates would also change (decrease). Therefore, the current ITT estimates are not a meaningful predictor of what the incremental change in activity would be under different counterfactual conditions. We refer readers to Ha-Thuc et al. (2020) for a

Table 4. Summary Statistics of Treatment or Engagement Received (e_p) for the Control and Boosted Groups in the 14-Day Period During the Experiment

Group	Mean	Standard deviation	25th percentile	Median	75th percentile
Control	268.53	8,985.18	1	8	40
Boosted	724.17	24,465.11	1	8	46

Table 5. Effect of Instrument Z (Control or Boost) on Treatment Received During the Experiment (e_p) and Post-treatment Activity (a_p)

	Dependent variable	
	Treatment received (1)	*Producer activity* (2)
Boost condition	455.642***	5.976***
	(19.845)	(0.840)
Constant	268.530***	589.461***
	(4.985)	(0.482)
R^2	0.000	0.000
No. of observations	4,871,594	4,871,594

Note. Robust standard errors are shown in parentheses.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

detailed discussion on the lack of counterfactual validity of boosting experiments. Given these issues, we need to estimate a metric that has counterfactual validity, that is, a metric that is invariant to the treatment assigned to other producers and can be directly used in policy design. As we will see in the next section, metrics that do have counterfactual validity are the average treatment effect and the conditional average treatment effect.

4.1.2. Marginal Treatment Effects.

The ITT estimate does not represent the incremental benefit (marginal treatment effect) of a producer receiving one additional engagement. Instead, it describes the effect of receiving, however, many additional engagements producers received as a result of being assigned to the boost condition. This distinction arises because the ITT estimate is an average treatment effect among people who were put into the boost condition, but in order to design a counterfactual recommendation system/targeting policy, we instead need a marginal treatment effect.

4.2. ATE of Engagement Using Instrumental Variables Approach

One way to improve on the ITT estimate is by estimating marginal treatment effects with a 2SLS model. Note that randomly assigning a producer to the boosted group serves as an external source of variation that increases how often they will be shown to consumers. For each of their specific items, this also has an effect on the number of engagements that the producer is going to receive. Recall that producers do not observe whether they are in a boosted group or a control group in our experiment, and do not observe how many impressions their tweets got. Therefore, the experiment condition serves as an instrumental variable—It exogenously varies how many engagements a producer receives but only through affecting how often their items are shown. Thus, we can use the boost versus control assignment as the instrument (Z) and the total engagement received during the experiment period as the treatment variable.

Our goal is to use the instrumental variable Z to model how each individual producer's activity a_p depends on their received consumer engagement e_p. For each producer, Z_p is one if the producer was randomly assigned to the boosted condition in our experiment and is zero if they were assigned to the control condition. The first stage of our 2SLS estimator is a linear regression where the endogenous engagement variable e_p is regressed on the instrument Z. In the second stage, the outcome variable a_p is regressed on the predicted engagement values $\hat{e}_p$ generated from the first stage.

$$\text{1st stage}: e_p = \gamma_0 + \gamma_1 Z_p + \epsilon$$
$$\hat{e}_p = \hat{\gamma}_0 + \hat{\gamma}_1 Z_p$$
$$\text{2nd stage}: a_p = \eta_0 + \eta_1 \hat{e}_p + \varepsilon$$

The results from this 2SLS model are shown in Table 6. We find that every incremental engagement that producers receive over the two-week treatment period increases producers' activity in the subsequent two-week post-treatment period by 0.0131 minutes.

Note that an implicit assumption in the 2SLS model is the following: The only channel through which the instrument affects post-treatment activity is the number of incremental engagements received during the experiment. However, in practice, the experiment conditions can also affect other channels that may, in turn, impact post-treatment activity. In particular, users who are boosted may also receive more followers during the experiment, and the increase in the number of followers may increase their postexperiment activity. We consider a series of analyses to test if this is a serious issue in our setting. Overall, we find that even if users receive more followers as a result of boosting, there is no systematic empirical evidence to suggest that this biases the estimate of incremental engagement. Please see Online Appendix C for detailed theoretical and empirical evidence.[11] Nevertheless, we caveat our findings and note that there is no way to completely rule out this hypothesis without an additional instrument (e.g., another bucket in the experiment). As such, if

Table 6. 2SLS Estimates of the Effect of Incremental Engagement (e_p) on Post-treatment Activity (a_p), Where the Instrument Is the Treatment vs. Control Bucket (Z)

Dependent variable: *Producer activity*	Coefficients and standard errors
Engagement received	0.0131***
	(0.0019)
Constant	585.94***
	(0.8755)
No. of observations	4,871,594

Note. Robust standard errors are shown in parentheses.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

there is sufficient reason to believe that such alternative pathways could meaningfully impact post-treatment activity, the firm should control for them in the analysis.

Finally, the 2SLS model requires two conditions for consistency: (1) constant treatment effect and (2) no heterogeneity in treatment intensity (Angrist and Imbens 1995, Syrgkanis et al. 2019). Unfortunately, neither of these two assumptions is true in our setting. In addition, average treatment effects (even if consistent) cannot help with policy design. We discuss these three challenges in detail below.

4.2.1. Treatment Effect Heterogeneity. Producers may differ significantly in terms of how much they value to additional consumer engagements. We expect users with different pretreatment variables to be differentially responsive to an incremental engagement, both in terms of individual attributes (e.g., followers, followings, and account age), as well as behavioral usage features (e.g., minutes active, engagement given, and engagement received). For example, new users may be more responsive to additional engagement compared with older users whose usage patterns may be more persistent. Thus, the estimated ATE may mask significant heterogeneity in the individual-level treatment effects.

One approach to generating individual-level treatment effects would be to use the orthogonal instrumental variable (OrthoIV) approach proposed by Chernozhukov et al. (2018). This method uses double machine learning to generate improved estimates over the standard 2SLS approach. However, it does not appropriately deal with the heterogeneous treatment intensity issues we described earlier, so it is not the best fit for this particular context. In Online Appendix D, we provide a summary of the OrthoIV method, as well as the results that it yields.

4.2.2. Heterogeneous Treatment Intensity and Compliance Issues. In there is heterogeneity in the treatment effects, the presence of heterogeneous treatment intensity can further invalidate the 2SLS estimates (see Syrgkanis et al. (2019) for a detailed discussion.

Notice that the field experiment described in Section 3.1 does not directly vary the number of times a specific producer gets shown. Instead, producers in the boost condition have their item scores multiplied by a constant value, thereby improving their chances of being shown to each consumer on the platform. The impact of being put in the boost condition can vary tremendously between different producers, as a function of producer attributes. For example, a producer who creates a lot of items will receive a stronger treatment "intensity" (i.e., more engagements) than a producer who writes fewer tweets because the former will have more items that get boosted. Similarly, a producer who has lots of followers is likely to receive more engagements because there is a larger base of prospective consumers who can engage with their items (upon seeing them).

This heterogeneous treatment intensity problem is known in the statistics literature as a "partial compliance" or "heterogeneous compliance" issue (Angrist and Imbens 1995, Dawid 2003). In our context, we find that there is significant heterogeneity in the intensity of treatment received by producers as a function of other producer-level observables. For example, we see that the treatment intensity varies with user state, for example, "heavy tweeters" and "medium tweeters" are more likely to receive more engagement (see Table A1 in the Online Appendix A). Thus, our analysis needs to account for this heterogeneity, this is, we cannot assume that all individuals who were assigned to the boost condition received the same level of treatment.

4.2.3. Relevance to Recommendation Policy Design. A final challenge is that the ATE (even if consistent) is not particularly useful from a policy design perspective. If the goal is to use this analysis to design counterfactual targeting policies, then heterogeneous (ideally individual-specific) treatment effects are necessary because the platform needs to identify which producers should be prioritized relative to others. Thus, just like the ITT estimates, the ATE estimates cannot be used for counterfactual policy design, because a constant estimate across all producers implies that all producers would be equally prioritized, which is equivalent to no one being prioritized.

Given these issues and those discussed in Section 4.1, we cannot use the ITT, 2SLS, or OrthoIV estimates for the purposes of identfying which producers have the strongest increase in activity if they receive additional engagement on their content.

5. Problem Definition

We now define the firm's problem more precisely and use the data from our field experiment to derive consistent estimates of a different estimand that does not suffer from the drawbacks discussed earlier. Specifically, our goal is to estimate the heterogeneous marginal treatment effect for each producer p as follows:

$$\theta(X_p) = \mathbb{E}[a_p | e_p + 1, X_p] - \mathbb{E}[a_p | e_p, X_p]. \quad (4)$$

This treatment effect represents how much each user would increase their activity on the platform if they received one additional engagement. In practice, the platform would need to estimate CATE values in order to summarize the treatment effect for users with features X_p.

Focusing on CATE values is useful for three reasons. First, understanding the heterogeneous causal impact of engagement on subsequent user behavior is an interesting question from a scientific perspective. As discussed in Section 2, no existing work has estimated this metric in the context of user response to engagement on social media settings; as such, this exercise can give researchers and managers insight into the magnitude and variation of CATE across users in a real large-scale social media platform. Second, a unique aspect of our setting is the rich set of user features (that consist of both past usage behavior and demographic data), which can be used to capture the heterogeneity in CATE. From a substantive perspective, knowing which types of users are more likely to respond to engagement is valuable because these findings can help us understand factors that influence producer behavior in social media platforms. Furthermore, managers can use these findings to develop a profile of users who are the most responsive to engagement and target them.

Finally, from a counterfactual policy perspective, the platform's problem is to identify and target producers who will have the strongest positive response after receiving engagement. Because CATE estimates have counterfactual validity, they can be used to develop a variety of potential interventions that prioritize responsive producers. For example, after estimating CATE, the platform can boost the producers with the highest CATE values (e.g., those in the top 1%) by a constant factor in its recommendation system, thereby providing a targeted version of the intervention we assigned at random in our experiment. A second option would be to prominently feature the content of the most responsive users. Alternately, instead of using CATE thresholds to target the responsive users, it can directly use these CATE estimates in a two-sided recommendation system that balances consumer utility/engagement with producer utility/engagement. To account for the fact that consumers and producers are both affected by the recommendation system, the platform can instead use a recommendation system that accounts for both of these constituencies. Recall that we previously defined the producer-specific CATE or response to receiving additional engagement as $\theta(X_p)$ (see Equation (4)). Thus, given an estimate of $\theta(X_p)$, the platform can calculate a weighted average of producer and consumer utility to rank items. For item i, consumer c, and producer p, we can then use the following score (σ_{icp}) to rank order items for consumer c:

$$\begin{aligned}\sigma_{icp} &= s_{ic}[1 + \lambda\theta(X_p)] \\ &= \mathbb{P}(\text{engagement}_{ic})[1 + \lambda\mathbb{E}(\text{incremental producerutility}_p|\text{engagement}_{ic})].\end{aligned}$$

Note that the relevance score s_{ic} and the individual-specific CATE value $\theta(X_p)$ can both be updated on a regular basis in order to generate up-to-date recommendations. In addition, the term λ can be chosen to meet the platform's goals: If they value consumer utility very highly, then λ should be low so that the weighted term σ_{icp} is mostly based on the probability of a consumer engaging with a particular piece of content. On the other hand, if the platform is heavily focused on generating additional engagement for the producers who are most responsive to that, then λ should be high.

One approach that may initially seem promising would be to estimate a heterogeneous ITT effect of the boost condition, $\alpha_1(16X\text{Boost}, X_p)$, and use that to develop targeting policies that provide a 16 times boost to a small set of users. However, the ITT effect $\alpha_1(16X\text{Boost}, X_p)$ is a function of engagement received, which in turn is a function of the experiment itself (the type and size of users boosted), and is therefore unlikely to be counterfactually valid. We refer readers to Online Appendix E for a more detailed explanation of this issue.

6. Estimating Heterogeneous Treatment Effects

A naive approach to estimating CATE values is to slice the data along different pretreatment variables and estimate ATEs within those slices of the data. However, such an approach is both practically infeasible and conceptually problematic for two reasons. First, we have a large number of pretreatment variables (X_p variables), which are highly correlated, and it is not obvious which variables we should use to slice the data. Second, slicing the data manually and exploring whether the treatment intensity and ATEs vary across different subslices is subject to p-hacking concerns, and hence not recommended (Athey and Imbens 2016). To avoid these problems, the recent practice in the literature has been to use machine learning methods that learn the heterogeneity in treatment intensity and treatment effects using a data-driven approach. We adopt a similar solution here by leveraging the recently developed DRIV model to give us consistent individual-level CATE values (Syrgkanis et al. 2019). This estimator builds on the double machine learning (DML) approach proposed in Chernozhukov et al. (2018). We refer readers to Ellickson et al. (2023) for a recent application of the DML approach in the marketing setting. Table 7 presents a summary of how the DRIV estimator compares against the ITT, 2SLS, and OrthoIV estimators discussed earlier.

6.1. DRIV Model and Estimation

The DRIV approach provides a unified framework that allows us to estimate an instrumental variable model with flexible functional forms, treatment effect heterogeneity, and heterogeneous treatment intensity. As

Table 7. Comparison of Estimator Properties

	Properties of the estimates			
	Counterfactual validity	Marginal treatment effects	Heterogeneous treatment intensity	Heterogeneous treatment effects
ITT	✗	✗	✗	✗
2SLS	✓	✓	✗	✗
OrthoIV	✓	✓	✗	✓
DRIV	✓	✓	✓	✓

with the 2SLS model, our goal is to use the instrumental variable Z_p to model how each individual producer's activity a_p is a function of their features X_p and their received consumer engagement e_p:

$$a_p = \theta(X_p)e_p + g(X_p) + \varepsilon$$
$$\mathbb{E}[\varepsilon | Z_p, X_p] = 0. \tag{5}$$

The marginal effect of each additional engagement is modeled as a flexible function θ that depends on the producer's features X_p. Two producers with different X values will return different treatment effects $\theta(X_p)$, so this modeling approach will generate heterogeneous treatment effects unlike the ITT approach described in Section 4.1 or the 2SLS ATE approach described in Section 4.2. Furthermore, the treatment intensity e_p is allowed to depend on both the treatment versus control bucket (Z_p) and their individual features X_p.

We now provide a high-level summary of the estimation steps. First, split the data into training (70%) and test (30%) data. Following the standard practice in double machine learning approaches, we will fit the preliminary nuisance functions on one subset of the training data and then estimate the second-stage models on a different subset (and vice versa). This practice is referred to as cross-fitting, and it ensures that the errors from potential over-fitting in the first stage do not propagate into the second-stage models.

Step 1: First estimate a set of nuisance functions from the different partitions of the training data separately. These submodels include the following:

- A flexible model to predict user activity based on user features: $\hat{a}(X_p) = \mathbb{E}[a_p | X_p]$.
- A flexible model to predict user engagements or treatment intensity based on user features: $\hat{e}(X_p) = \mathbb{E}[e_p | X_p]$.
- A flexible model to predict user engagements or treatment intensity based on user features and experiment condition: $\hat{h}(X_p) = \mathbb{E}[e_p | X_p, Z_p]$.
- A flexible model to predict how user engagements and the experiment condition are jointly affected by user features: $\hat{f}(X_p) = \mathbb{E}[e_p \cdot Z_p | X_p]$.

Step 2: Calculate the following residualized versions of the models using the preliminary models from Step 1 (estimated on a different partition of the training data):

$$\tilde{a}_p = a_p - \hat{a}(X_p)$$
$$\tilde{e}_p = e_p - \hat{e}(X_p)$$
$$\tilde{Z}_p = Z_p - \hat{Z}(X_p)$$
$$\hat{\beta}(X_p) = \mathbb{E}[\tilde{e}_p \cdot \tilde{Z}_p | X_p] = \mathbb{E}[(e_p - \hat{e}_p) \cdot (Z_p - \hat{Z}_p) | X_p]$$
$$= \hat{f}(X_p) - \hat{a}(X_p)\hat{e}(X_p).$$

Step 3: Using a preliminary estimate of $\hat{\theta}_{pre}$ and our estimate of $\hat{\beta}$, estimate a doubly robust treatment effect, $\hat{\theta}_{DR}(X_p)$, by minimizing the loss function:

$$\hat{\theta}_{DR}(X_p) = \arg\inf_{\theta} \frac{2}{N} \sum_{p} \left(\hat{\theta}_{pre}(X_p) + \frac{(\tilde{a}_p - \hat{\theta}_{pre}(X_p)\tilde{e}_p)\tilde{Z}_p}{\hat{\beta}(X_p)} - \theta(X_p) \right)^2, \tag{6}$$

where N is the total number of observations in the training sample. When p is one partition of the training sample, then $\hat{\theta}_{pre}$ and the nuisance function estimates come from a different partition, and vice versa. The preliminary estimate of θ_{pre} comes from minimizing the following square loss function on a separate part of the training sample:

$$\hat{\theta}_{pre} = \arg\inf_{\theta} \frac{2}{n} \Sigma_p [a_p - \hat{a}(X_p) - \theta(X_p)(\hat{h}(X_p, Z_p) - \hat{e}(X_p))]^2.$$

Note that θ_{DR} is robust to the potential misestimation of $\hat{\theta}_{pre}$ or $\hat{\beta}$. Thus, as long as one of these estimates is right, the final estimate is consistent. This approach is similar in spirit to the doubly robust estimators in standard causal inference settings (see Dudík et al. (2011) as an example), with $\hat{\beta}$ playing a role similar to a propensity score. The loss function in Equation (6) can be minimized using any parametric or semiparametric estimator.

The full details of the model and consistency proofs for this procedure can be found in Syrgkanis et al. (2019), and the estimation routine is publicly available as part of the EconML library using Python. A major benefit of this modeling framework is that the different submodels in Step 1 can be estimated using flexible machine learning methods. For our implementation, we first log-transform all the continuous features such

as "tweets composed" and "favorites received" using a $\ln(X+1)$ transformation. We then use a combination of LightGBM gradient-boosted decision trees (Ke et al. 2017) and Lasso for the nuisance submodels from Step 1 and a Lasso model for the CATE function $\hat{\theta}_{DR}$. These models are chosen based on their predictive performance in the holdout test sample. The good performance of the Lasso in the second step is consistent with earlier research that shows those Lasso-based CATE estimators often outperform other methods (Simester et al. 2020, Yoganarasimhan et al. 2023). We provide a comparison between different model variants in Online Appendix G.

Finally, we note that, in addition to the DRIV approach, researchers can also use other recently developed double machine learning based estimators to estimate individual-level treatment effects that account for both heterogeneous treatment intensity and heterogeneous treatment effects (Athey et al. 2019, Farrell et al. 2021). We do not take a stance on the pros and cons of these different approaches because our focus is on substantive and managerial insights. Nevertheless, the Syrgkanis et al. (2019) approach offers a few advantages in our specific setting, such as the availability of the estimator in the EconML package that allows for easier adoption in real industry applications, the flexibility to experiment with different machine learning methods for the first stage nuisance models, and the ability to use a simple parametric model for the last step to aid with interpretability and substantive insights.

6.2. DRIV Results

The key outputs from our estimation procedure are the predicted CATE estimates $\hat{\theta}_{DR}(X_p)$. Table 8 provides summary statistics for these estimated CATE values. We find that most of the estimated CATE values are small in magnitude. The outcome variable a_p is measured in minutes, so the average value of 0.14 implies that receiving one incremental engagement as a producer causes them to increase their usage by only about 8 seconds (i.e., 0.14 minutes ≈ 8.4 seconds) on average. The standard deviation of the CATE estimates is very high relative to the mean, which demonstrates the importance of allowing for heterogeneous treatment effects across users. This pattern is similar to the findings in earlier papers that estimate CATE values based on ITT experiments (Syrgkanis et al. 2019).

We now interpret our model results and examine its predicted outcomes through two approaches: coefficient estimates and feature comparison.

6.2.1. Coefficient Estimates.

Because we use a Lasso model in Step 3 to estimate the CATE ($\hat{\theta}_{DR}(X_p)$), we are able to generate coefficient estimates for each of the pretreatment producer features in our data. These coefficient estimates are displayed in Table 9, and they summarize how a one-unit increase in each particular variable affects the user's minutes active, holding all other variables constant. This is helpful for understanding which factors contribute to a particular user's CATE value being large or small.[12] Nevertheless, these coefficient results do not yield actionable guidance for the platform, because the goal is to figure out which kinds of users have the highest CATE values, not necessarily *why* they have the highest values. This issue is particularly noticeable because many of the producer features are highly correlated with each other. As such, it may not be that useful for the platform to know what the effect of each variable is when holding other factors constant; instead, it would be better for them to generate a profile of the kinds of customers who have high versus low CATE values. Therefore, in the next section, we analyze how producer-level features differ based on their CATE values.

6.2.2. User-Level Heterogeneity: Feature Comparison.

We now examine the heterogeneity in CATE estimates $\hat{\theta}_{DR}(X_p)$. Our goal is to understand if and how the users who are the most responsive to engagement are distinct from the rest of the population. Note that this question is important both substantively and managerially. From a substantive perspective, this provides us with insights into the profile of responsive users. Further, this information allows managers to develop targeting strategies and recommendation policies that can target the right set of users.

We divide users into two groups: those in the top 1% of CATE values and those in the bottom 99% of CATE values. The former group represents the set of users who benefit the most from incremental engagement, whereas the latter group serves as the baseline level of response for the user base at large. The average CATE value is 2.36 for users in the top 1% of CATE values, but it is only 0.12 for users in the bottom 99%. This large discrepancy reinforces the importance of estimating heterogeneous treatment effects here, especially if we are considering targeted interventions that would go to some users but not others.

To understand how these groups differ from each other in observable ways, we can summarize their values for the pretreatment producer features X_p. A full

Table 8. Summary Statistics of the CATE Estimates $\theta(X_p)$ from the DRIV Model for All Users in the Data

Outcome	Mean	Standard deviation	25th percentile	Median	75th percentile
Minutes active	0.14	0.52	−0.16	0.11	0.37

Note. These CATE estimates represent how much users would be expected to increase their "minutes active" after receiving one additional engagement.

Table 9. Coefficient Results from the DRIV Model

Variable	Coefficient
Tweets composed	0.228
Favorites received	−0.056
Retweets received	0.035
Replies received	−0.021
Follows received	−0.095
Follows made	−0.174
Favorites given	0.036
Retweets sent	−0.008
Replies sent	−0.139
Quote retweets sent	0.228
Minutes active	0.216
Active days	−0.352
Monetizable active days	−0.235
Active followers	0.070
Total followers	−0.028
Following	0.016
Indicator push notification enabled	−0.008
Account age (days)	−0.024
User state: Heavy Tweeter	−0.154
User state: Heavy non-Tweeter	0.144
User state: Medium Tweeter	0.094
User state: Medium non-Tweeter	0.403
User state: Light	0.271
User state: Very light	0.196
User state: Near zero	0.654
Country: Japan	0.046
Country: United States	0.042
Country: Brazil	0.096
Country: Philippines	−0.594
Country: United Kingdom	−0.207
Country: Indonesia	−0.341
Country: Turkey	−0.136
Country: South Africa	0.194
Country: Mexico	1.940
Country: Argentina	−0.089
Intercept	0.200

Notes. These coefficients are derived from the Lasso final stage submodel and they describe how each producer feature enters the CATE effect function $\hat{\theta}_{DR}(X_p)$. All continuous variables are log-transformed using a $\ln(X+1)$ transformation. For categorical variables, "User state: New" and "Country: Rest of the World" are the omitted baseline levels.

comparison between the two groups across all the features X_p is provided in Table 10. To more easily visualize a few of the key differences, we also show detailed histograms for some of the key pretreatment producer features for each group in Figure 3.

We find differences between these groups in users' activity levels, account age, and popularity. First, we find that users with high CATE values tend to have more followers, and they are also following more accounts. They also have older accounts (more account days), but they do not use the platform as regularly (fewer active days). Users with high CATE values are also more likely to be in the lower-intensity user state groups (everything other than heavy tweeter and heavy non-tweeter). On most of the other observable features, there are no major differences between users in the top 1% of CATE values versus the other 99% of users. This analysis suggests that longstanding users who are no longer using the platform regularly are likely to respond most positively to receiving incremental engagement; as such, these users should be targeted if the goal is to increase users' time spent on the platform. Finally, note that while we focused on the 99%–1% cohorts in this analysis, the broader substantive results are similar if we use larger thresholds/cutoffs (e.g., 95%–5%).[13]

7. Focusing on Other Outcomes

The results in Section 6.2 show that on average, receiving additional engagement on one's social media content leads to an increase in time spent on the platform. This outcome is of first-order importance to social media platforms because time spent on the platform is often the key monetizable outcome for ad-supported applications and websites. However, there may be other metrics that are of interest to the platform as well. We now examine how receiving engagement affects four other measures of user activity: tweets composed, favorites given, monetizable active days, and total engagements given.

7.1. How Does Receiving Engagement Improve Users' Production vs. Consumption?

In addition to understanding how receiving additional engagement causes users to increase their time spent on the platform, the platform may also be interested in discovering how that additional time is being spent. In particular, we can now examine whether additional engagement causes producers to produce more content, consume more of other people's content, or both.

First, we examine these outcomes using ITT and 2SLS models in which we replace the outcome variable, which was previously the overall minutes spent on the platform. We estimate two sets of models: one that uses a purely production-focused outcome (tweets composed) and one that uses a purely consumption-focused outcome (favoriting other users' tweets). The results from these preliminary approaches are presented in Tables 11 and 12, respectively. In both sets of results, we find that there is a positive effect on both tweets composed and favorites given.

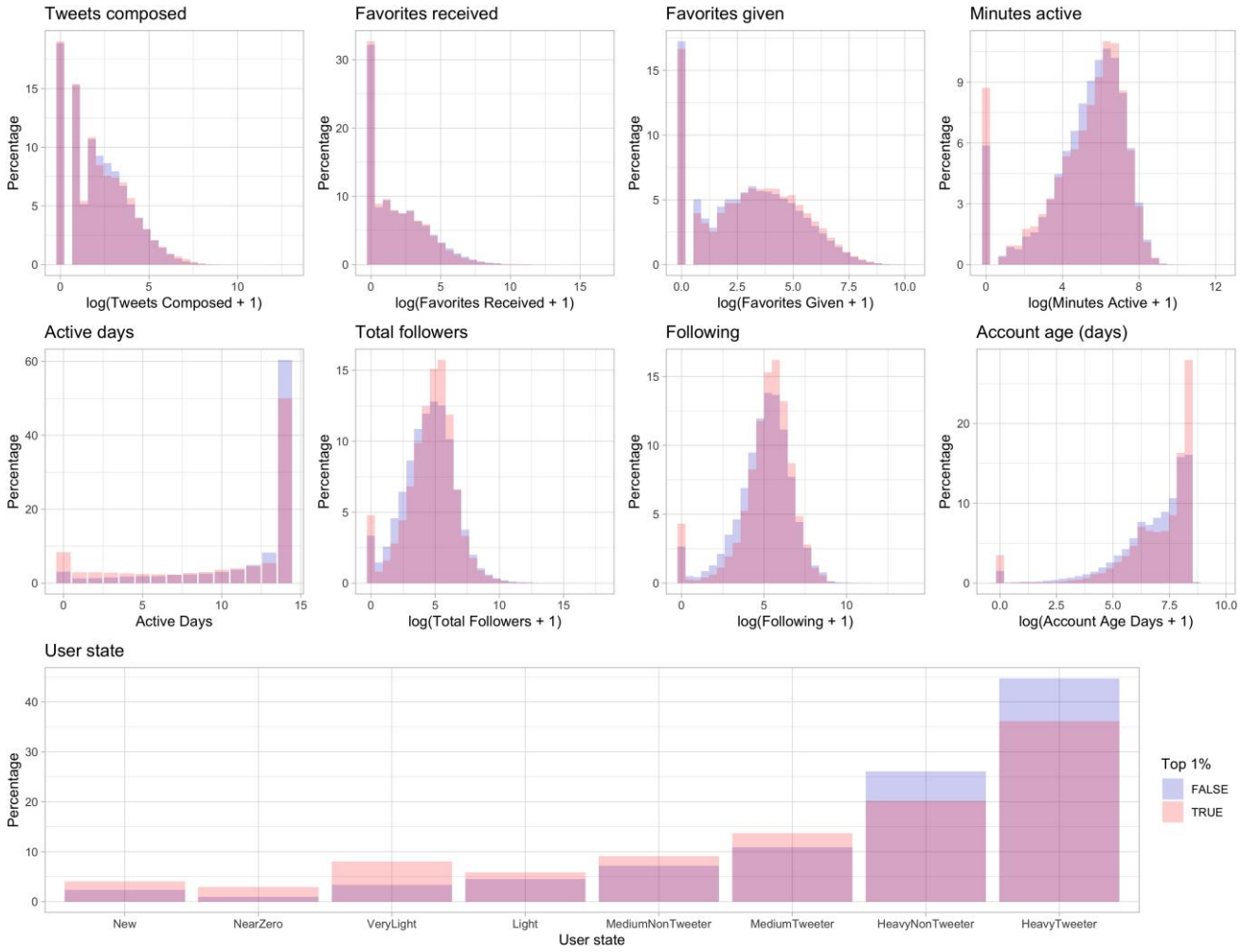
To investigate this issue further while dealing with heterogeneous treatment intensity and heterogeneous treatment effects, we now re-estimate our DRIV model specification (see Section 6.1) with tweets composed and favorites given as the outcome variables. The estimated CATE values from these DRIV models are shown in the top two rows of Table 13. On average, we find that receiving one additional engagement causes users to increase their tweets composed by 0.042 and to increase their favorites given by 0.054. This corresponds to average

Table 10. Summary Statistics of Pretreatment Producer Features (X_p) Among Users in the Top 1% of CATE Estimates vs. Others

	Users in the top 1% of CATE values (minutes active)				
Variable	Mean	Standard deviation	25th percentile	Median	75th percentile
Tweets composed	62.78	1,593.57	1	6	28
Favorites received	161.86	3,404.82	0	3	23
Retweets received	38.81	1,649.75	0	0	1
Replies received	22.45	252.63	0	1	5
Follows received	8.64	218.01	0	0	2
Follows made	3.86	17.09	0	0	2
Favorites given	170.28	517.75	3	26	123
Retweets sent	31.37	175.25	0	2	12
Replies sent	22.27	115.22	0	1	7
Quote retweets sent	8.47	63.34	0	0	3
Minutes active	611.02	901.67	55	290	808
Active days	10.23	4.99	7	13	14
Monetizable active days	10.15	5.03	6	13	14
Active followers	525.50	9,227.77	26	83	211
Total followers	1,132.27	32,712.44	36	127	341
Following	452.81	1,579.88	90	222	471
Indicator push notification enabled	0.75	0.44	0	1	1
User state: Heavy Tweeter	0.36	0.48	0	0	1
User state: Heavy non-Tweeter	0.20	0.40	0	0	0
User state: Medium Tweeter	0.14	0.34	0	0	0
User state: Medium non-Tweeter	0.09	0.29	0	0	0
User state: Light	0.06	0.23	0	0	0
User state: Very light	0.08	0.27	0	0	0
User state: Near zero	0.03	0.17	0	0	0
User state: New	0.04	0.20	0	0	0
Account age (days)	2,106.25	1,559.68	537	2,005	3,666
	Users in the bottom 99% of CATE values (minutes active)				
Variable	Mean	Standard deviation	25th percentile	Median	75th percentile
Tweets composed	41.22	250.84	1	7	25
Favorites received	202.63	9,369.62	0	4	25
Retweets received	35.84	2,159.96	0	0	1
Replies received	21.50	428.11	0	1	6
Follows received	17.64	469.97	0	1	4
Follows made	10.87	57.81	0	1	5
Favorites given	155.02	496.31	2	20	101
Retweets sent	31.54	186.74	0	1	10
Replies sent	21.12	111.29	0	1	9
Quote retweets sent	3.55	21.44	0	0	1
Minutes active	622.77	922.04	74	292	804
Active days	11.75	3.88	11	14	14
Monetizable active days	11.33	4.30	10	14	14
Active followers	613.16	14,495.75	19	70	218
Total followers	1,134.13	52,887.95	24	98	330
Following	442.18	1,933.04	64	180	435
Indicator push notification enabled	0.73	0.44	0	1	1
User state: Heavy Tweeter	0.45	0.50	0	0	1
User state: Heavy non-Tweeter	0.26	0.44	0	0	1
User state: Medium Tweeter	0.11	0.31	0	0	0
User state: Medium non-Tweeter	0.07	0.26	0	0	0
User state: Light	0.05	0.21	0	0	0
User state: Very light	0.03	0.18	0	0	0
User state: Near zero	0.01	0.10	0	0	0
User state: New	0.02	0.15	0	0	0
Account age (days)	1,719.98	1,428.47	430	1,325	2,987

Note. The underlying CATE estimates represent how much the users would be expected to increase their "minutes active" after receiving one additional engagement.

Figure 3. (Color online) Histograms for Key Pretreatment Producer Features (X_p) Among Users in the Top 1% of CATE Estimates vs. Others



Note. The underlying CATE estimates represent how much the users would be expected to increase their "minutes active" after receiving one additional engagement.

increases of 0.10% for the production outcome (tweets composed) and 0.03% for the consumption outcome (favorites given), relative to those variables' pretreatment baseline values (shown in Table 1).

Overall, the results from our different models indicate that receiving incremental engagement encourages users to consume and engage with other users' content, and produce more content, but it has an even larger relative effect (roughly three times bigger) on their production of content. This indicates that interventions based on increasing users' engagement can lead to broad improvements across multiple dimensions that are of interest to the platform and that the benefits are significant for users' overall enjoyment of the platform rather than being isolated to their enjoyment from producing content. However, if the platform specifically wanted to maximize production *or* consumption rather than minutes active, then it would benefit from targeting users who are most responsive on that particular dimension.

In Online Appendix H, we show that the CATE values for production and consumption are negatively correlated with each other and that the platform would end up targeting very different sets of users if the goal was to maximize users' production versus consumption. In particular, we find that targeting relatively inactive users is ideal for maximizing consumption (favorites given) but if the goal is to maximize product (tweets composed), then the platform should target a broader set of users. We also show in Online Appendix H that targeting users to maximize production also leads to a small increase in their consumption—This demonstrates that these two forms of usage are not necessarily substitutes from the perspective of the user.

Table 11. Effect of Instrument Z (Control or Boost) on the Production-Focused Outcome (Tweets Composed) and the Consumption-Focused Outcome (Favorites Given) After the Experiment

	Dependent variable	
	Tweets composed (1)	*Favorites given* (2)
Boost condition	0.6629*	2.433***
	(0.293)	(0.459)
Constant	38.260***	145.319***
	(0.266)	(0.261)
R^2	0.000	0.000
No. of observations	4,871,594	4,871,594

Note. Robust standard errors are shown in parentheses.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

7.2. How Does Receiving Engagement Improve Users' Monetizable Active Days?

In Section 6.2, we focused on users' total minutes active as the main variable of interest. This is consistent with the standard monetization strategy used by social media platforms, which is to increase users' time spent on the platform because that allows them to be shown more advertisements. However, platforms are also interested in the overall size of their user base and getting these users to use the service on a regular basis. These active-day metrics are important because platforms typically report them to investors regularly. Furthermore, having a large user base helps attract new advertisers and can increase the diversity and broader appeal of the content on the platform (Godes and Mayzlin 2004). For these reasons, we now focus on the number of monetizable active days; see Table 1 for the summary statistics of this variable in the pretreatment period. This variable refers to the number of days a user used the platform and provided nonzero ad revenue, so it provides a measure of how regularly each user is using the platform, rather than measuring how intensely they are using it. Similar daily log-in measures have previously been used as a measure of platform usage (see Gallus et al. (2020) as an example).

Table 12. 2SLS Estimates on the Production-Focused Outcome (Tweets Composed) and the Consumption-Focused Outcome (Favorites Given) After the Experiment

	Dependent variable	
	Tweets composed (1)	*Favorites given* (2)
Engagement received	0.0015*	0.0053***
	(0.0006)	(0.001)
Constant	37.869***	143.89***
	(0.429)	(0.474)
No. of observations	4,871,594	4,871,594

Note. Robust standard errors are shown in parentheses.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

We re-estimate the DRIV model as described in Section 6.1 with monetizable active days as the outcome variable a_p and present the results in the last row of Table 13.[14] We find that the average incremental effect is positive but small in magnitude: on average, each incremental engagement leads to an increase of 0.0044 monetizable active days. One reason for this smaller effect size could be the fact that the values of this outcome variable are both constrained and skewed. For instance, in the 14-day pretreatment period, over half of the users in our sample were active for all 14 days (Table 1). From an intervention perspective, this is challenging because it means that the majority of the users cannot be improved on this metric. Further, compared with the CATE estimates for minutes active (as summarized in Table 8), the CATE estimates for monetizable active days are not distributed as widely; that is, the standard deviation is not that large compared with the mean. In Online Appendix I.2, we present a detailed analysis of how users in the top 1% of CATE on this metric differ from users in the bottom 99%.

Broadly speaking, monetizable active days represent the extensive margin of usage on the platform, whereas minutes active represent the intensive margin. If the platform's primary objective is to get more users to use the platform regularly, then the platform should focus on increasing monetizable active days. On the other hand, if the platform's primary objective is to increase the time that each user spends within each day or within each login session, then the platform should focus on increasing the minutes active. In Online Appendix I.3, we present a more detailed comparison of the differences in the profiles of users who are the most responsive on these two different usage metrics.

7.3. Effect on Engagements Given

Thus far, we have shown that receiving engagements makes producers more likely to spend time on the platform and return to the platform more regularly. In addition, we also found that producers created more content and favorited more tweets from other users. From the platform's perspective, favoriting other users' tweets is helpful because it may encourage the recipients to increase their own usage of the platform, to create more content, and to engage more with others. However, favoriting is just one of the three user behaviors that represent engagement. To consider this issue more holistically, we now examine how receiving additional engagement causes each user to alter their total engagement given to others.

We start with some preliminary analysis. First, in Table 14, we show the ITT estimates of the effect of the instrument on the outcome variable. There is a positive and significant effect; being in the boost condition leads to an approximately 1.86% improvement in the number

Table 13. Summary Statistics of the CATE Estimates $\theta(X_p)$ for All Users in the Data for the DRIV Models Trained on Different Outcomes

Outcome	Mean	Standard deviation	25th percentile	Median	75th percentile
Tweets composed	0.042	0.181	−0.067	0.009	0.105
Favorites given	0.054	0.153	−0.037	0.054	0.139
Monetizable active days	0.0044	0.0017	0.0034	0.0045	0.0055

Note. These CATE estimates represent how much users would be expected to increase their "tweets composed," their "favorites given," and "monetizable active days" after receiving one additional engagement.

of engagements given, compared with the baseline of 193 for the control group. This is almost twice the effect as on minutes active (see the discussion in Section 4.1). Next, we estimate a 2SLS model with total engagements given as the outcome variable and present the results in Table 15. Once again, we find that receiving engagement has a small but positive effect on downstream engagements given. Nevertheless, these ITT and 2SLS estimates suffer from the same challenges discussed earlier in Section 4.

Therefore, next, we re-estimate our DRIV model with "total engagement given" as the outcome variable and present a summary of the CATE values from this analysis in Table 16. We find that the average CATE is 0.10, that is, receiving one engagement leads producers to give an incremental 1/10 engagement to other users. Furthermore, we see that there is significant heterogeneity in CATE values across users (similar to the CATE estimates for active minutes).

To further understand which users are the most responsive on this outcome variable, we conduct an analysis similar to that in Section 6.2.2. As before, we divide users into two groups: Those in the top 1% of CATE values, and those in the bottom 99% of CATE values. The average CATE value is 1.01 for users in the top 1% of CATE values, but it is only 0.09 for users in the bottom 99%. We show detailed histograms for some of the key pretreatment producer features for each group in Figure 4, and a full comparison across all the user features X_p is provided in Online Appendix J. We find that the users with the highest CATE values (i.e., the users who provide the most engagements to other people after they receive one incremental engagement) are systematically different from the rest of the population: They tend to be less active users (fewer minutes active and active days) with lower levels of content production (tweets composed) and engagement with others (favorites given and favorites received). These substantive findings can be used by platforms and managers to target users if they seek to promote downstream engagement on others' content.

8. Quantifying the Returns to Engagement Across Target Groups

We now use the individual-level CATE estimates to understand the returns from different types of targeting approaches. We focus on quantifying the total *gain in minutes active* for the platform if it were to target the *most responsive producers* (top 1% of $\theta(X_p)$ values) with *one incremental engagement*. This targeting scenario assumes that the platform is using the user-level CATE values when deciding whom to target rather than using other observable user-level variables, because this provides the most direct measure of which users will respond the most after receiving additional engagements.

We use minutes active as the primary outcome of interest since it is most closely tied to platform revenue. Moreover, in Section 8.2, we show how a producer's response to other outcomes (e.g., engagement) can be transformed into a measure of incremental downstream change in minutes active. We estimate the counterfactual gains from one incremental engagement rather than the overall gains from a specific intervention (e.g., boosting the most responsive users, or prominently showcasing them on the front page, etc.) because the overall effect of such interventions is a function of both the total engagements received under the intervention as well as the CATE estimate. Although our CATE estimates have counterfactual validity, the engagement received is endogenous to the experiment (see Online Appendix E for details) and is therefore unlikely to be counterfactually valid. Thus, we focus on counterfactual exercises that rely only on CATE estimates.

Finally, focusing on the top 1% of users (or a relatively small percentage) has a few natural advantages. First, it ensures that the platform will be discovering producers who will respond most positively if they were to receive additional engagement, so targeting this group can yield substantial benefits for the

Table 14. Effect of Instrument Z (Control or Boost) on the "Engagements Given" as the Outcome Variable

Dependent variable: *Engagements given*	Coefficients and standard errors
Boost condition	3.5879***
	(0.587)
Constant	193.077***
	(0.334)
R^2	0.000
No. of observations	4,871,594

Note. Robust standard errors are shown in parentheses.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Table 15. 2SLS Estimates of the Effect of Incremental Engagement (e_p) on Post-treatment Engagements Given (a_p), Where the Instrument Is the Bucket (Z)

Dependent variable: *Engagements given*	Coefficients and standard errors
Engagement received	0.0079***
	(0.0013)
Constant	190.96***
	(0.6071)
No. of observations	4,871,594

Note. Robust standard errors are shown in parentheses.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

platform. This is likely to be the case because the treatment effects $\theta(X_p)$ are highly heterogeneous (as we saw in Table 8, some producers have very high values, but most do not). Second, targeting a small percentage of producers allows the platform to maintain a consumer experience that is nearly fully in line with their preferences. On the other hand, if many producers get targeted/boosted, then the consumers' content suggestions can deviate significantly from what they actually want to see. This may negatively affect their experience on the platform and may lead to worse outcomes for the platform over the long run. Therefore, in the rest of this section, we will consider counterfactuals where we will examine the impact of giving one incremental engagement to the top 1% of producers with the highest estimated treatment effects, and measuring the overall increase in time spent on the platform.

8.1. Direct Gains Compared with Benchmark Target Groups

Recall that our DRIV models yield individual-level CATE values that indicate how each user would change their behavior if they were to receive one additional engagement from other users. To understand how much the platform can benefit from this information, we now consider what would happen if the platform were to target users who are in the top 1% of CATE values for minutes active versus target users who are in the top 1% of CATE values for engagement given.

For comparison, we also evaluate the effects of targeting two other baseline groups chosen based on user activity levels: users who are in the top 1% of minutes active, and users who are in the bottom 1% of minutes active. Such heuristics/baselines are commonly used by the managers of the platform to identify and target users. For instance, one existing approach at the platform at the time of the experiment was to focus on low-activity users and employ interventions that gave them additional engagement. Comparing our targeting strategy with the two heuristic approaches allows us to (a) provide preliminary benchmarks on the returns to using the individual-level CATE estimates from our approach compared with simpler heuristics and (b) generate insights on how interventions based on one outcome (e.g., minutes active) perform compared with interventions based on other outcomes of interest (e.g., engagements given).

For each of these four groups, we estimate what happens to the total minutes active when we provide the targeted group of users with one additional engagement.[15] The results of this exercise are shown in Table 17. A few important patterns emerge from this analysis. First, we find that targeting users with high CATE values is more effective than targeting users based on heuristics that use activity-level thresholds, that is, the former approaches yield bigger improvements in total minutes active. For instance, if the platform wants to increase minutes active, then targeting users with the highest CATE values for minutes active yields a roughly 4–19 times improvement over the other alternatives we consider. Second, we find that targeting users with the highest CATE values for minutes active (column 1) yields about a three times improvement over targeting users with the highest CATE values for engagement given (column 2).

Interestingly, the set of users with the highest CATE values for minutes active is quite different from the set of users with the highest CATE values for engagement given. To provide some substantive insights into how the two target groups vary, we provide a comparison of the distribution of some of the key pretreatment variables across both the groups in Figure 5. We find that targeting based on engagement given CATE would focus on less active (fewer tweets composed, minutes active, etc.) and less connected users (fewer followers and following), compared with targeting based on minutes active CATE.

8.2. Gains with Spillover Effects

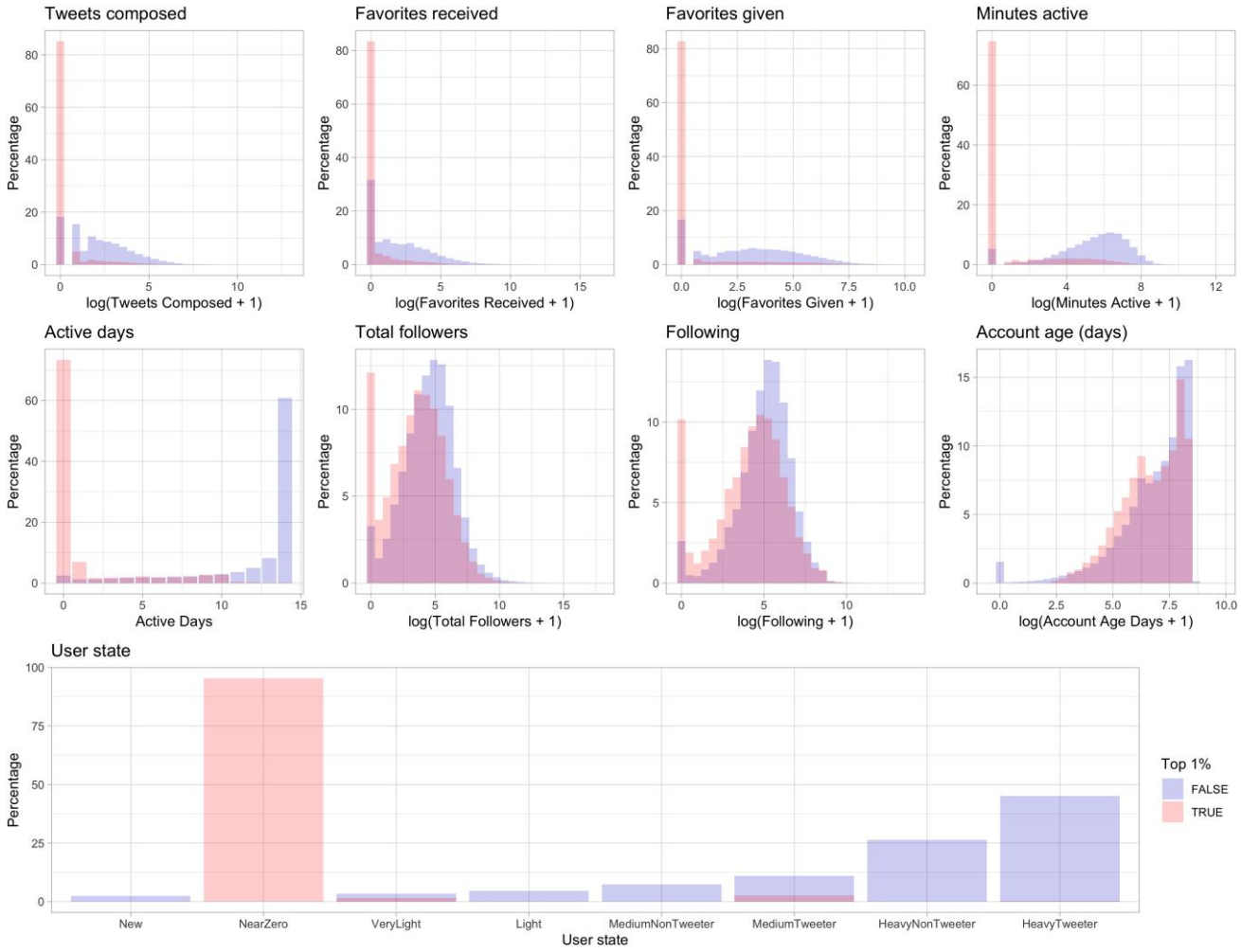
The analysis in the previous section only quantifies the effects for the focal (targeted) users, but it does not consider the potential positive spillover effects on other users connected to the targeted users. Recall the analysis in Section 7.3, which suggests that when a user receives an additional engagement, it encourages them

Table 16. Summary Statistics of the CATE Estimates $\theta(X_p)$ from the DRIV Model for All Users in the Data

Outcome	Mean	Standard deviation	25th percentile	Median	75th percentile
Total engagement given	0.1010	0.1827	−0.0034	0.0990	0.2005

Note. These CATE estimates represent how much users would be expected to increase their "total engagement given" after receiving one additional engagement.

Figure 4. (Color online) Histograms for Key Pretreatment Producer Features (X_p) Among Users in the Top 1% of CATE Estimates vs. Others



Note. The underlying CATE estimates represent how much the users would be expected to increase their "engagement given" after receiving one additional engagement.

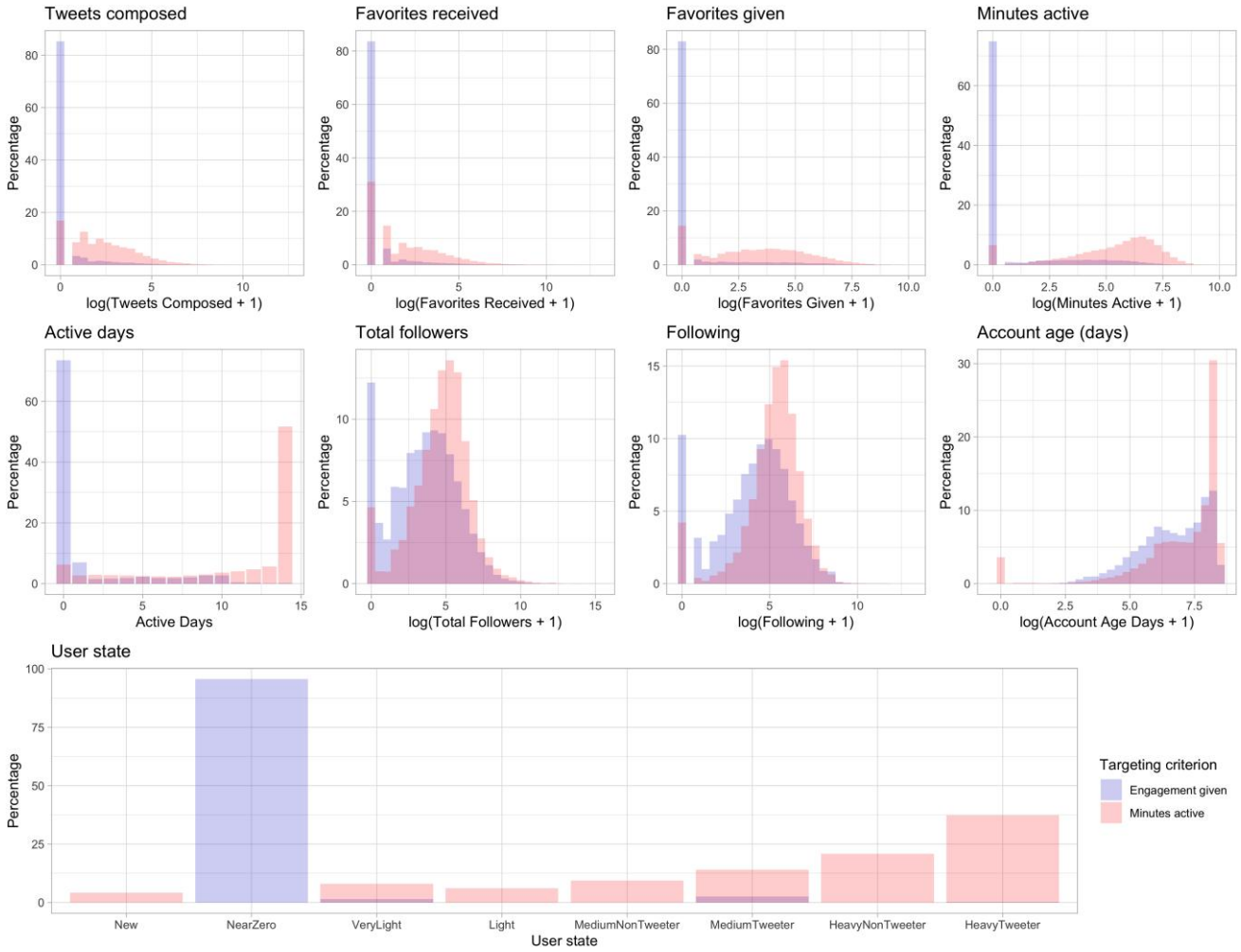
to not only increase their own usage of the platform but also to provide additional engagements to other users, which in turn would cause those users to increase their usage of the platform. These spillover effects can lead to long-term benefits for the platform because they could create a "virtuous cycle" in which users are increasing their own usage, engaging more with each other's content, and therefore increasing other people's usage. Similar positive spillover phenomena have been examined in the contexts of other social media settings; for example, video seeding strategies on YouTube (Yoganarasimhan 2012), customer relationship management

Table 17. Summary of Results from Four Potential Intervention Scenarios

	Group that gets targeted			
Outcome	(1) Top 1% minutes active CATE	(2) Top 1% engagement given CATE	(3) Top 1% most active	(4) Bottom 1% most active
Incremental minutes active	114,894	40,243	29,605	6,102
Number of targeted users	48,716	48,716	48,673	48,717

Notes. Each scenario examines the effect of providing one additional engagement to four different target groups. We consider targeting users either based on their individual CATE values or their activity (i.e., their minutes active). The number of observations varies slightly between the four interventions because of ties regarding who meets the criteria for the top 1%.

Figure 5. (Color online) Histograms for Key Pretreatment Producer Features (X_p) Among Users in the Top 1% of CATE Estimates for "Engagement Given" vs. Users in the Top 1% of CATE Estimates for "Minutes Active"



Note. These correspond to the groups of people who should be targeted if the platform is trying to improve each of these two criteria.

efforts (Ascarza et al. 2017), and social nudges on a video sharing platform (Zeng et al. 2023).

Measuring these spillover benefits is challenging in our context because we have a CATE estimate for how much each targeted user engages with other people's content (i.e., we know the CATE of engagements given), but we do not observe which specific people's content they engage with (i.e., we do not observe the user-to-user engagement edges in the network). This data limitation also means that we do not know how much a recipient of this spillover engagement would increase their time spent on the platform by (i.e., we do not know their CATE values) because their user characteristics are not observed. Therefore, we approximate the effect by assuming that any additional engagements are dispersed at random to the other users on the network. We also only consider a one-hop spillover; that is, we consider the targeted group and the users who receive engagement from people in that targeted group. This can be interpreted as a lower bound of the spillover effects, because in reality there may be further downstream positive effects (e.g., user A engages with user B, who engages with user C, who engages with user D).[16]

We can now provide an estimate of the spillover effect by adding together two outcomes for each intervention: the direct effect that the intervention has on the targeted group and the indirect effect of how the targeted group causes other people to increase their usage. For example, if we want to consider how providing one additional engagement to a specific user p will affect the total minutes active on the platform, we can use the following equation:

$$\begin{aligned}&\text{Incremental minutes active}_p\\&\quad=\text{Direct effect of engagement}\\&\qquad+\text{Indirect effect of engagement}\\&\quad=\theta_p^{\text{minutes active}}+(\theta_p^{\text{engagement}}\cdot\mathbb{E}[\theta^{\text{minutes active}}]),\end{aligned}\tag{7}$$

where $\theta_p^{\text{minutes active}}$ is user p's individual CATE value from the DRIV model for minutes active, $\theta_p^{\text{engagement}}$ is user p's individual CATE value from the DRIV model for total engagement given, and $\mathbb{E}[\theta^{\text{minutes active}}]$ is the overall average CATE value from the DRIV model for minutes active (roughly 0.14; Table 8).[17] Thus, the first term $\theta_p^{\text{minutes active}}$ captures the direct impact on the focal user p, which is simply the incremental time spent on the platform by user p. The second term $\theta_p^{\text{engagement}} \cdot \mathbb{E}[\theta^{\text{minutes active}}]$ is the one-hop spillover effect on other users in the platform: This is the average incremental time spent by the users who received incremental engagements from user p (as a result of user p receiving an additional engagement).

In order to calculate the total effects of targeting a specific group of users, we can apply Equation (7) to different target groups and then aggregate the results across users within that particular target group. In Table 18, we show the results of this exercise when we consider three potential target groups: (1) users who are in the top 1% of CATE values for minutes active, (2) users who are in the top 1% of CATE values for engagement given, and (3) users who are in the top 1% of incremental minutes active (based on the definition provided in Equation (7)). Note that this last target group is the set of users for whom the total treatment effect (the sum of direct and spillover effects) is the highest, that is, users in the top 1% of $\theta_p^{\text{minutes active}} + (\theta_p^{\text{engagement}} \cdot \mathbb{E}[\theta^{\text{minutes active}}])$.

The results in Table 18 provide two main takeaways. First, we can compare the results in columns 1 and 2 of Table 18 versus Table 17 to isolate the effects of including spillovers in our analysis. We find that the effects of spillovers are relatively small and that including these spillovers has not changed the main conclusions from our earlier analysis. Second, we can see that targeting users with the largest values for incremental minutes active (column 3) yields very small improvements over targeting users who are in the top 1% of CATE values for minutes active (column 1). This indicates that targeting users based on the CATE values for minutes active is a good targeting strategy for the platform, even if there are spill-over effects.

The close similarity in results between columns 1 and 3 is because the first term in Equation (7) (the direct effect) largely overwhelms the second term (the spillover effect). Notice that, although both $\theta_p^{\text{engagement}}$ and $\theta^{\text{minutes active}}$ are relatively similar in magnitude, the second term of Equation (7) consists of a multiplier ($\mathbb{E}[\theta^{\text{minutes active}}] = 0.14$), which leads to the spillover effects being much smaller than the direct effects.

As a result, in our full data sample of nearly 4.9 million users, the CATE for minutes active (only direct effect) and the total CATE for incremental minutes active (the sum of direct and spillover effect, as shown in Equation (7)) have a very high correlation of 99.88%. Thus, targeting the people in the top 1% of values for each of these two criteria also yields groups that are very similar to each other; in our context, there is a 96% overlap between the users who would be targeted under these two scenarios. In sum, these findings suggest that, although positive spillover effects exist, they are not significant in magnitude and the platform can simply focus on the first-order effects when choosing targeting criteria.

Finally, note that our counterfactual analysis should be interpreted carefully with the appropriate caveats. First, we only focus on the producers directly targeted and/or those within one hop of the targeted producer. However, if we view consumer attention and engagement as a zero-sum game, then any incremental engagements for the targeted group may come at the expense of other producers on the platform (the 99% who are not targeted for intervention). When we shift engagements and attention to a focal targeted group, these other producers may be relatively disadvantaged in terms of opportunities to receive consumer engagement. As a result, they may receive fewer engagements and this may negatively affect their subsequent behavior. Second, in the quest to provide additional engagement to producers, consumers may be shown content that they do not enjoy as much, and this could cause them to reduce their consumption or their usage of the platform. Given these caveats, our counterfactual analyses should be used as a first-order approximation when comparing different targeting approaches rather than used as a definitive prediction of the overall impact of any specific strategy. More broadly, if the platform wanted to implement a two-sided

Table 18. Summary of Results from Three Potential Intervention Scenarios

	Group that gets targeted		
Outcome	(1) Top 1% minutes active CATE	(2) Top 1% engagement given CATE	(3) Top 1% incremental minutes active
Incremental minutes active (with spillovers)	116,558	47,197	116,585
Number of targeted users	48,716	48,716	48,716

Notes. Each scenario examines the effect of providing one additional engagement to three different target groups, accounting for the effects both on that target group as well as the spillovers from the additional engagement that they provide to others. The first two approaches target users based on their individual CATE values from two DRIV models with different outcome variables. The third approach targets users based on their values for "incremental minutes active" with spillovers, as defined in Equation (7).

recommendation system that takes into account both consumer and producer utility (as described in Section 5), then our producer-side CATE estimates could be treated as inputs into a two-sided recommendation system that takes into account both consumer and producer utility.

9. Discussion and Conclusion

On social media platforms, users can receive engagement from other people in response to their posts. We outline a framework that allows us to measure how individual users respond to this kind of engagement: how it affects their minutes spent on the platform, their monetizable active days, the amount of content they produce, and the amount of other people's content they engage with. Our approach yields heterogeneous marginal treatment effects that summarize how each user would respond if they were to receive one additional engagement, while also dealing with the endogeneity and measurement problems inherent in measuring user behavior on social media platforms.

We apply our approach through a field experiment on Twitter in which some users were purposefully boosted compared with other users. This intervention exogenously increases the number of impressions their content receives, which in turn increases the amount of engagement they receive. We analyze this experimental data with a doubly robust instrumental variable (DRIV) machine learning model. Our results indicate that receiving engagement has a positive effect on users' time spent on the platform, their monetizable active days, and engagements given to other users; however, there is a considerable amount of user heterogeneity in each of these estimated treatment effects. We also find that receiving additional engagement causes users to engage more with other people's content as well as produce more tweets themselves.

Apart from these broad takeaways, the main focus of our research is to estimate how individual users would respond to receiving additional engagement. These estimates provide insight into the types of users who are most responsive to engagement. Our approach can be used by managers to build user profiles and gain an understanding of which user features/characteristics are tied to user responsiveness and behavior. Our findings also have important implications for possible interventions that could be used by social media platforms. We find that platforms can likely improve the overall success of the platform by identifying users who respond heavily to receiving additional engagement and then making them more prominent to other users. Showing these users' content more often will increase the total usage of the platform, which in turn should improve advertising revenues and other financial metrics for the platform. Further, we find that the platform can simply focus on the first-order effects of its interventions because spillover effects, although positive, are quite small in magnitude.

Finally, our study provides many avenues for future research. First, in our setting, we do not delve into *why* certain users respond more to engagement. It is possible that some users simply enjoy the social feedback and respond with more content, whereas others respond positively in order to leverage the feedback for financial gains. However, this is something future research can examine more closely, and doing so can provide additional insights into how the platform can incentivize and manage these different types of content producers. Second, in our setting, there are no direct financial incentives for content production/engagement on the platform. However, as discussed in Section 1, in some social media platforms, users can also directly make money from their content. In such settings, it would be useful to examine the extent to which users' content production and platform usage is driven by peer feedback versus financial incentives.

Acknowledgments

The authors thank Sophie Hilgard, Victor Lei, and Yang Tang for input; Dean Eckles, Ali Goli, Solomon Messing, Donald Ngwe, Omid Rafieian, Yanwen Wang, Linli Xu, and Zikun Ye for extensive feedback and comments that have significantly improved the paper; and participants at the following seminars and conferences for helpful comments and suggestions: University of Michigan, Harvard University, the University of Toronto, the University of British Columbia, Lehigh University, Chapman University, Stanford University, William & Mary, Hong Kong University of Science and Technology, the Virtual Digital Economy Seminar Series, the UW-UBC marketing conference, the INFORMS Marketing Science Conference, the CODE@MIT Conference on Digital Experimentation, the Workshop on Platform Analytics, and the Interactive Marketing Research Conference. An earlier version of this paper was circulated under the title "Producer and Consumer Engagement on Social Media Platforms."

Endnotes

[1] An example of a platform that uses direct payments is YouTube, with its YouTube Partner Program (YouTube 2023). Although this approach is easy to implement in cases where the content is sufficiently long and discrete, it is harder in the case of platforms like Facebook, Instagram, and Twitter where multiple pieces of small text/pictures from a variety of producers are consumed concurrently within a few seconds.

[2] There is also another stream of research that examines how users change the type of content they create after receiving visible recognition and attention. Burtch et al. (2022) find that Reddit users who receive external recognition subsequently create content that is similar to their award-winning posts, but Huang et al. (2025) find the opposite pattern when examining content creation in an online image-sharing platform.

[3] Historically, social media platforms such as Twitter used a reverse chronological order to show content. However, over the years they

have evolved to adopt recommendation algorithms that rank the content based on the consumer's interest (Huszár et al. 2022).

[4] About 18 months after our experiment concluded, Twitter published a blog post describing the outline of their recommendation algorithm and ranking system (Twitter 2023). The general principles are similar to what we describe here: The algorithm is intended to "optimize for positive engagement (e.g., Likes, Retweets, and Replies). This ranking mechanism takes into account thousands of features and outputs 10 labels to give each Tweet a score, where each label represents the probability of an engagement. We rank the Tweets from these scores." There are some minor differences between their description of the ranking algorithm and what was in place during our experiment (e.g., in our context the score was based on an unweighted combination of likes, retweets, and replies rather than a weighted combination). However, these differences and/or the specifics of the ranking algorithm are not relevant to our methodological approach or substantive findings; we only require the experimental conditions (boost/control) to be exogenous.

[5] The boost factor was chosen based on two important, but opposing, considerations. On the one hand, very low boost factors will not lead to any significant increase in the engagement received for the boosted group. That is, the boost factor needs to be sufficiently high to ensure that the experiment is a sufficiently strong instrument and to change the rankings of items on the consumers' feed. On the other hand, very high boost factors will push the boosted users to the top of all the consumers' feeds and can degrade the consumer experience. After discussions with the product teams and recommendation systems teams, we chose 16 as a good factor that balanced both these considerations. Because most relevance scores s_{ic} are relatively low and the distribution is skewed, multiplying those values by 16 would allow low-ranked content to rise up substantially in the consumer feed (e.g., from rank 300 to rank 100) but not to the very top.

[6] In late 2022 (over a year after our study ended), this detail was changed by Twitter so that view counts became visible for tweets (Clark 2022).

[7] From an analysis perspective, our approach is reasonable because completely inactive users do not get any treatment (or engagements) at all and have no impact on the outcomes observed. Nevertheless, this implies that the distribution of users in the experiment is different from the overall population of users on Twitter; that is, the users in the experiment skew more active. As such, the summary statistics shown in Section 3.2 should be interpreted as specific to users in the experiment and not as platform-level metrics.

[8] The main distinction between active days and monetizable active days is that the latter metric only includes days in which the user accessed Twitter long enough to be shown at least one advertisement.

[9] This is the exact definition used by the firm to define engagements, and it is considered the cumulative measure of engagement that is of importance to the firm. The firm did not see any managerial value in separately measuring the value of replies, retweets, and favorites. Further, this measure was used for other purposes within the firm, e.g., to calculate relevance scores of tweets and rank them. As such, it was important to keep the definition of engagement constant across the firm. Of course, if the firm cares about one engagement metric over the other (instead of the cumulative engagement measure), then it can always measure the incremental impact on one metric and control for the others, or potentially use multiple instruments to derive the relative effect of each.

[10] We can also conduct a related analysis where we estimate these ITT effects separately for different groups of users who receive different amounts of additional engagement during the experiment (e.g., users who receive lots of additional engagement versus those who receive very little additional engagement). However, interpreting the results of this exercise is challenging because there are multiple potential explanations. See Online Appendix B for a description of this analysis.

[11] Boosted users may also get more direct messages during the experiment, and as a result, spend more time on the platform. We do not have data on the number of direct messages to test for this alternative channel; however, given that direct messages on Twitter were quite rare compared to other forms of engagement (e.g., favorites, replies), we do not expect this to be a serious issue.

[12] An alternative approach would be to project the estimated CATE values on the producer features. In Online Appendix F, we conduct this analysis using both a linear regression and an elastic net, and we find that the takeaways are similar to the results in Table 9.

[13] In Section 8, we provide a more detailed discussion on the reasons for focusing interventions on a small fraction of users.

[14] Preliminary ITT and 2SLS models are shown and discussed in Online Appendix I.1.

[15] This approach is a direct evaluation method based on the estimated CATE values. Therefore, any error or noise in the estimated CATE values will propagate into the estimated quantities.

[16] Although it is theoretically possible to include additional hops of spillover in our analysis, doing so would likely yield less credible results because the marginal effects of providing engagement would shrink with each additional hop and the results could be outweighed by any noise in the estimates. Further, we see that even one-hop spillover effects are not particularly large. Hence, the spillover effects on higher hops are likely to be negligible.

[17] This approach is flexible and can be used to consider other outcomes as well. For instance, to estimate the effect on incremental monetizable active days, we can simply replace $\theta_p^{\text{minutes active}}$ and $\mathbb{E}[\theta^{\text{minutes active}}]$ with $\theta_p^{\text{monetizable active days}}$ and $\mathbb{E}[\theta^{\text{monetizable active days}}]$, respectively.

References

Ahn D-Y, Duan JA, Mela CF (2016) Managing user-generated content: A dynamic rational expectations equilibrium approach. *Marketing Sci.* 35(2):284–303.

Angrist JD, Imbens GW (1995) Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J. Amer. Statist. Assoc.* 90(430):431–442.

Angrist JD, Imbens GW, Rubin DB (1996) Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* 91(434):444.

Ascarza E, Ebbes P, Netzer O, Danielson M (2017) Beyond the target customer: Social effects of customer relationship management campaigns. *J. Marketing Res.* 54(3):347–363.

Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. USA* 113(27):7353–7360.

Athey S, Tibshirani J, Wager S (2019) Generalized random forests. *Ann. Statist.* 47(2):1148–1178.

Bradlow E (1998) Encouragement designs: An approach to self-selected samples in an experimental design. *Marketing Lett.* 9(4):383–391.

Burtch G, He Q, Hong Y, Lee D (2022) How do peer awards motivate creative content? Experimental evidence from reddit. *Management Sci.* 68(5):3488–3506.

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018) Double/debiased machine learning for treatment and structural parameters. *Econom. J.* 21(1): C1–C68.

Clark M (2022) Twitter is now showing everyone how many views your tweets get. Accessed August 15, 2025, https://www.theverge.com/2022/12/21/23522064/twitter-view-count-roll-out-personal-info.

Dawid AP (2003) Causal inference using influence diagrams: The problem of partial compliance. Green PJ, Hjort NL, Richardson S, eds. *Highly Structured Stochastic Systems*, Oxford Statistical Science Series (Oxford Academic, Oxford, UK), 45–65.

Dudík M, Langford J, Li L (2011) Doubly robust policy evaluation and learning. Getoor L, Scheffer T, eds. *Proc. 28th Internat. Conf. Machine Learn.* (Omnipress, Madison, WI), 1097–1104.

Eckles D, Kizilcec RF, Bakshy E (2016) Estimating peer effects in networks with peer encouragement designs. *Proc. Natl. Acad. Sci, USA* 113(27):7316–7322.

Ellickson PB, Kar W, Reeder JC III (2023) Estimating marketing component effects: Double machine learning from targeted digital promotions. *Marketing Sci.* 42(4):704–728.

Falk K (2019) *Practical Recommender Systems* (Manning, Shelter Island, NY).

Farrell MH, Liang T, Misra S (2021) Deep neural networks for estimation and inference. *Econometrica* 89(1):181–213.

Gallus J, Jung OS, Lakhani KR (2020) Recognition incentives for internal crowdsourcing: A field experiment at NASA. Working Paper No. 20-059, Harvard Business School, Boston.

Godes D, Mayzlin D (2004) Using online conversations to study word-of-mouth communication. *Marketing Sci.* 23(4):545–560.

Guo M, Ni J, Shen Q, Xu Y (2023) Quantity vs variety in online content production: Evidence from a knowledge sharing platforms. Johns Hopkins Carey Business School working paper 22-12, Baltimore.

Ha-Thuc V, Dutta A, Mao R, Wood M, Liu Y (2020) A counterfactual framework for seller-side a/b testing on marketplaces. Huang J, Chang Y, Cheng X, eds. *Proc. 43rd Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval* (Association for Computing Machinery, New York), 2288–2296.

Huang JT, Narayanan S (2020) Effects of attention and recognition on engagement, content creation and sharing: Experimental evidence from an image sharing social network. Working paper, Stanford University, Stanford, CA.

Huang JT, Kaul R, Narayanan S (2025) Novelty in content creation: Experimental results using deep learning–based image recognition on a large social network. *Marketing Sci.*, ePub ahead of print November 19, https://doi.org/10.1287/mksc.2023.0547.

Huszár F, Ktena SI, O'Brien C, Belli L, Schlaikjer A, Hardt M (2022) Algorithmic amplification of politics on twitter. *Proc. Natl. Acad. Sci. USA* 119(1):e2025334119.

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, et al. (2017) Lightgbm: A highly efficient gradient boosting decision tree. von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R, eds. *Adv. Neural Inform. Processing Systems*, vol. 30 (Curran Associates Inc., Red Hook, NY), 3149–3157.

Kemp S (2023) Digital 2023: Global overview report. https://datareportal.com/reports/digital-2023-global-overview-report.

Liu J, Toubia O, Hill S (2021) Content-based model of web search behavior: An application to tv show search. *Management Sci.* 67(10):6378–6398.

Lu S, Xie Y, Chen X (2023) Immediate and enduring effects of digital badges on online content consumption and generation. *Internat. J. Res. Marketing* 40(1):146–163.

Messing S (2013) Social news and civic engagement. PhD dissertation, Stanford University, Stanford, CA.

Rafieian O, Yoganarasimhan H (2023) AI and personalization. Sudhir K, Toubia O, eds. *Artificial Intelligence in Marketing* (Emerald, Oxford, UK), 77–102.

Restivo M, van de Rijt A (2014) No praise without effort: Experimental evidence on how rewards affect Wikipedia's contributor community. *Inform. Comm. Soc.* 17(4):451–462.

Shriver SK, Nair HS, Hofstetter R (2013) Social ties and user-generated content: Evidence from an online social network. *Management Sci.* 59(6):1425–1443.

Simester D, Timoshenko A, Zoumpoulis SI (2020) Targeting prospective customers: Robustness of machine-learning methods to typical data challenges. *Management Sci.* 66(6):2495–2522.

Syrgkanis V, Lei V, Oprescu M, Hei M, Battocchi K, Lewis G (2019) Machine learning estimation of heterogeneous treatment effects with instruments. Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, eds. *Adv. Neural Inform. Processing Systems*, vol. 32 (Curran Associates Inc., Red Hook, NY), 15193–15202.

Toubia O, Stephen AT (2013) Intrinsic vs. image-related utility in social media: Why do people contribute content to twitter? *Marketing Sci.* 32(3):368–392.

Twitter (2023) Twitter's recommendation algorithm. https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm.

Yoganarasimhan H (2012) Impact of social network structure on content propagation: A study using YouTube data. *Quant. Marketing Econom.* 10:111–150.

Yoganarasimhan H (2020) Search personalization using machine learning. *Management Sci.* 66(3):1045–1070.

Yoganarasimhan H, Barzegary E, Pani A (2023) Design and evaluation of optimal free trials. *Management Sci.* 69(6):3220–3240.

YouTube (2023) YouTube Partner Program overview. Accessed on March 9, 2023, https://support.google.com/youtube/answer/72857.

Zeng Z, Dai H, Zhang DJ, Zhang H, Zhang R, Xu Z, Shen Z-JM (2023) The impact of social nudges on user-generated content for social network platforms. *Management Sci.* 69(9):5189–5208.

CORRECTION

In this article, "How do Content Producers Respond to Engagement on Social Media Platforms?" by Simha Mummalaneni, Hema Yoganarasimhan, and Varad Pathak (first published in *Articles in Advance*, February 26, 2026, *Marketing Science*, DOI: 10.1287/mksc.2023.0178), The caption and notes section of Table 9 on page 14 has been updated.