Social Media Comments

Dante Donati and Lena Song*

December 2025

Preliminary draft - please do not circulate

Abstract

Comment sections are a central feature of social media platforms, enabling users to interact with content and with one another. This paper studies how the presence and stance of comments influence subsequent user behavior on Facebook. We develop a novel experimental pipeline that manipulates and randomizes comment visibility and stance using Meta's built-in A/B testing infrastructure to estimate their causal impact on engagement. In collaboration with a leading racial justice organization, we conduct a large-scale field experiment involving over one million U.S. users randomly assigned to one of four treatment arms: (i) no visible comments (control), (ii) opposing, (iii) supportive, and (iv) mixed comments displaying both stances. We find that the presence of a comment section increases engagement. Opposing comments, in particular, significantly amplify reactions, comments, and link clicks relative to the control, whereas supportive comments have no effect. Most subsequent comments and reactions are positive in tone, consistent with *reactance* as a potential mechanism. Effects are concentrated among male users and those in conservative areas, suggesting that *curiosity* may also play a role. Our results underscore the role of comment-moderation policies, with important implications for businesses, users, platforms, and policymakers.

Keywords: Comments, Field Experiments, Platforms, Social Media, User Generated Content

JEL codes: C93, D12, D90, J15, L82, L86, M37

*Donati: Columbia Business School. dd3137@gsb.columbia.edu. Song: University of Illinois Urbana-Champaign. lenasong@illinois.edu. We thank Charles Amuzie from Color of Change for support. We thank seminar and conference participants at Berlin School of Economics, Carnegie Mellon University, University of Illinois Urbana-Champaign, Marketing Science, Columbia Business School, and CESifo Venice Summer Institute: Workshop on Digital Platforms for helpful feedback. We are grateful to Anna Bezhanishvili, Seungwoo Kim, Jungyun Kim, Thomas Lilly and Navtej Singh for excellent research assistance. The research was approved by the Institutional Review Boards at Columbia University (AAAU8166). This experiment was registered in the American Economic Association Registry for randomized control trials under trial number AEARCTR-0013812. We acknowledge financial support from the Russel Sage Foundation (Grant G-2309-44994), the Digital Future Initiative and the Bernstein Center at Columbia Business School, and the Provost Office at Columbia University.

1 Introduction

In its idealized form, social media has often been described as a digital public square—a space where people can come together to share information and engage in open dialogue.[1] One feature that facilitates discussions is the comment section. On platforms like Facebook and YouTube, users can engage not only with content but also with one another through comments and replies. However, comment sections often reflect the views of a vocal minority whose discussions tend to be highly polarized (Kim and Noh, 2025). Because comments are publicly visible, the views expressed by a few can influence how others interpret information, form opinions, and behave, consistent with theories of social learning, persuasion, and social norms (Kamenica and Gentzkow, 2011; DellaVigna and Gentzkow, 2010; Bursztyn, González and Yanagizawa-Drott, 2020). This raises concerns about whether social media fulfills its public-square ideal, especially in light of ongoing challenges related to misinformation, polarization, and hate speech.[2] Understanding the causal effects of comments is therefore central to evaluating online discourse and informing emerging content-moderation policies.

This paper examines how users interact in social media comment sections and how those interactions shape subsequent behavior. In collaboration with Color of Change, the largest online racial justice organization in the United States, we ran organic field experiments in which more than one million Facebook users were randomly assigned to see posts about racial justice. Because racial justice is a domain of intense public debate and highly polarized opinions (Pew Research Center, 2024), it provides a particularly useful setting for studying cross-cutting interactions online and for isolating how the stance of the comment section affects user engagement.

First, we document how users engage with this content and analyze the discussions that emerge in the comment sections. We then leverage these organically generated comments to estimate the causal impact of the comment section itself. Although comments are often used as measures of engagement and studied using observational data (Huang, Choi and Wan, 2024; He, Hong and Raghu, 2025; Moehring, 2024), isolating their effects from those of the original posts is empirically challenging: posts that receive many early comments may have inherent characteristics that make them more engaging ex ante, and platform algorithms tend to recommend posts with higher early activity, further increasing their visibility and subsequent engagement.[3] To address this challenge, we build on existing platform features to design a pipeline that manipulates comment visibility and stance as a novel research instrument. Using this pipeline, we provide causal evidence that the presence and stance of pre-existing comments subsequently influence the behavior of other users.

To collect organic engagement, we presented five post designs as sponsored content to roughly

[1] See, for example, `www.nytimes.com/2022/04/26/technology/twitter-elon-musk-free-speech.html`, `www.washingtonpost.com/politics/2024/02/28/supreme-court-revives-debate-over-social-media-public-square/`, `www.apnews.com/article/supreme-court-social-media-florida-texas-dc523bc9a6ef7b0f7b0aa933d0a43cca`.

[2] For reviews on these challenges and the broader political and social impacts of social media, see Zhuravskaya, Petrova and Enikolopov (2020), Aridor et al. (2024), and Sunstein (2018).

[3] This reflects a common identification issue in the user-generated content literature (Eliashberg and Shugan, 1997).

135,000 Facebook users across 4,095 ZIP codes. These ZIP codes were grouped into 30 clusters stratified by their 2020 Republican vote share to vary the prevailing ideology in the audience: low Republican vote share ($< 30\%$), mixed (45–55%), and high ($> 70\%$). The posts covered voter suppression, criminal-justice reform, education equity, environmental justice, and technology fairness, and were co-created by professional designers with guidance from the nonprofit to align with branding guidelines and encourage authentic user responses. Over two weeks, users generated more than 12,000 reactions, roughly 1,500 comments directed at the organization, and 1,750 link clicks.

We distinguish between two types of user interaction: reactions (likes and emoji responses) and comments. In addition, we classify all comments using a large language model along four dimensions: political stance (supportive vs. opposing), sentiment (positive vs. negative), offensiveness, and informativeness. We show that engagement patterns differ sharply across areas with different ideological compositions. Users in conservative areas commented at higher rates (1.3% of those reached) than users in progressive areas (0.8%), yet they reacted less frequently (7.9% vs. 10.1%). Comments directed at the organization were overwhelmingly conservative in right-leaning areas (76% vs. 55%), more negative in sentiment (80% vs. 56%), and more likely to be offensive (47% vs. 35%). These patterns persisted after accounting for exposure frequency and ad creatives, indicating that audience ideology shapes not only engagement volume but also the tone and substance of discourse. Consistent with the literature on gender differences in vocal engagement (Klinowski, 2023), women were more likely to react than men and far less likely to comment, with the ideological gap in commenting driven almost entirely by men.

These findings highlight that engagement is not the same as endorsement: posts can generate engagement driven largely by negative or opposing comments, especially in ideologically opposed communities. This supports concerns that comment sections may amplify polarized or extreme voices rather than reflect the broader audience, effectively creating a micro-echo chamber even when the post itself is cross-cutting (Kim and Noh, 2025).

Using these organically generated comments, we provide causal evidence on the effect of the comment section on subsequent user engagement in a field experiment involving over one million Facebook users. We leverage platform features to create a novel pipeline that manipulates the visibility and stance of the comment section. To isolate the effect of comment stance from that of the post itself, we re-marketed a subset of posts from the previous phase of the study — each pre-populated with two organic comments — to roughly one million new users across 1,881 ZIP codes, combined into 18 clusters spanning the three ideology groups previously identified. Using Meta's A/B-testing infrastructure, in each cluster, we randomized participants into four conditions: (1) no comments (control), (2) supportive comments only, (3) opposing comments only, and (4) a mixed condition with one supportive and one opposing comment. We measured attention to the comment section, user interactions, and direct traffic to the organization's website.

We implemented several design choices to address potential concerns. To minimize violations of the Stable Unit Treatment Value Assumption (SUTVA), a real-time filtering pipeline hid all new comments, minimizing the influence users exposed to the same post could have on one another. To isolate the effect of the comment section from other visible interactions, we equalized the number of shares and average reactions across conditions. To address the risk of divergent delivery—the tendency of ad algorithms to learn and serve treatment arms to different user types based on early engagement (Braun and Schwartz, 2025; Eckles, Gordon and Johnson, 2018)—we followed and augmented best practices from recent work (Burtch et al., 2025). Specifically, we split budgets evenly across arms, launched all ads simultaneously, imposed a one-impression cap per user, and optimized for reach rather than engagement. In addition, we ran the campaign over multiple days to achieve audience saturation for each ad—ensuring that nearly all users within a defined area were reached—to further limit algorithmic learning and divergence. We verified balance across gender, age, and delivery metrics such as impressions and costs.

We show that comment sections significantly influence subsequent user engagement: the opinions of a small, vocal minority can shape the behavior of a larger audience. Comparing the treatment conditions to the control condition in which posts were displayed without any comments, we find that displaying any pre-populated comments increased all subsequent engagement with the post by 0.065 percentage points, a 13% rise relative to the baseline ($p < 0.01$). Further dissecting by types of engagement, we show that the presence of comments increased the likelihood of users clicking to view the comment section by about 0.05 percentage points on a 0.24% baseline, with no significant differences between supportive, opposing, or mixed conditions. However, comment stance had pronounced effects on downstream engagement: ads with opposing comments drove significantly more interactions—reactions, saves, and user comments—than those with supportive comments ($p < 0.01$). Compared with the no-comment condition, the presence of negative comments increased these interactions by roughly 45% ($p < 0.05$). Click-through rates were highest in the opposing-only condition, rising by 0.034 percentage points above the 0.228% control rate (a 15% relative increase, $p < 0.01$), leading to more visits to the organization's website and reducing cost-per-click.

These average effects mask substantial heterogeneity in the valence of subsequent interactions, as well as across genders and ideological compositions. Further analysis of the stance of subsequent engagement shows that most interactions within opposing comment sections are supportive. This suggests that exposure to disagreement prompts users to reinforce the original message—a pattern consistent with a *reactance* mechanism, in which individuals respond to opposing views by reaffirming their own position. We also find that the impact is much larger for men than for women, and stronger in conservative than in progressive areas. These patterns also point to a *curiosity* mechanism, whereby visible disagreement captures attention, particularly among audiences whose prior beliefs diverge from the post's message. In both cases, while opposing comments may not

persuade subsequent users, they increase the perceived utility of engaging—either to counter-argue and defend the main message or because the content becomes more interesting, contentious, or entertaining to explore.

Our study has high ecological validity. First, we examine the effects of comment sections on a large, diverse population of social media users across a wide range of ZIP codes. Second, users were unaware they were part of an experiment and interacted with the posts as they naturally would. Finally, in our experiment, we used organic comments, ensuring that the content shown to new users reflected genuine Facebook user responses rather than researcher-generated or AI-generated text. Together, these design choices make our findings highly generalizable to real-world online discourse.

The results have several implications for policy and practice. Regulations like the European Union's Digital Services Act and the United Kingdom's Online Safety Act place growing responsibility on platforms to monitor and manage user-generated content, including both posts and comment sections.[4] Our results reveal a strategic tension for social media platforms in their comment moderation: prioritizing negative, opposing comments may boost on-platform user engagement, but for content producers who post on these platforms, such comments may be polarizing and pose brand-safety risks. For advertisers and campaign managers, our results imply that they need to calibrate their moderation policies to their objectives. Tolerating opposing comments can boost engagement and potentially reduce ad costs, while stricter moderation helps protect brand image. For nonprofit organizations, these findings underscore the need for caution when extending outreach beyond their supporters. Without moderation, their comment sections can quickly become an echo chamber with polarized opinions. Balancing visibility and brand safety requires deliberate comment management strategies.

Our paper builds on several strands of literature. First, it is related to the literature on the economics of social media Zhuravskaya, Petrova and Enikolopov (2020); Aridor et al. (2024). The existing literature in social media has used the comment section as a data source to study a range of questions (see, for example, Yang, Ren and Adomavicius 2019; Moehring 2024). In this literature, comment sections are often used as a measure of engagement and studied using observational data. However, one open question is the causal effect of the comment section itself. To answer this, we build on existing platform features to design a pipeline that manipulates comment visibility and stance as a novel research instrument. Using this pipeline, we provide causal evidence that the stance of pre-existing comments subsequently influences the behavior of other users.

More broadly, our paper builds upon and expands the literature on the drivers and consequences of user-generated content (UGC). Existing studies have examined the determinants of customer conversation and word-of-mouth production (e.g., Chen and Berger 2013; Dubois, Bonezzi and

[4]Digital Services Act (2022, EU): `https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng`; Online Safety Act (2023, UK): `https://www.legislation.gov.uk/ukpga/2023/50/contents`

De Angelis 2016; Deng et al. 2022), the features that drive engagement with advertising content on social media (e.g., Lee, Hosanagar and Nair 2018), and the impact of customer reviews on demand and firm performance (e.g., Chevalier and Mayzlin 2006; Mayzlin, Dover and Chevalier 2014; Xu, Armony and Ghose 2021; Donati 2025). In line with recent theoretical work by Nistor and Selove (2024), our study empirically examines the role of social media comments as a form of UGC and investigates how their stance shapes subsequent user engagement with racial justice posts.

The paper is organized as follows: Section 2 presents the setting of the study, Section 3 describes the generation and analysis of organic interactions, Section 4 presents causal evidence on the impact of the comment section, and Section 5 concludes.

2 Study Setting

2.1 Context

The comment section is a common feature of almost all major social media platforms. The most widely used platforms in the United States, Facebook and YouTube, display comment sections directly beneath posts or videos, allowing users to respond publicly, reply to one another, and react through likes or other engagement tools. Other platforms, such as Reddit, rely heavily on comment sections as their primary mode of discussion. Beyond social media, online news outlets such as The New York Times, The Wall Street Journal, and CNN maintain dedicated comment sections at the end of articles, often moderated or restricted to subscribers. Across these contexts, comment sections serve as central spaces where users respond to content, interact with one another, and participate in broader online discourse.

By providing a shared space for deliberation, comment sections create opportunities for individuals to encounter a wide range of viewpoints. In particular, it facilitates interaction among individuals with differing views who may not have the opportunity to engage in face-to-face conversations. Such interactions could, in principle, foster productive exchange on divisive issues: users may articulate their opinions, engage in reasoned debate, and develop a better understanding of opposing perspectives. At the same time, comment sections can also devolve into unproductive or even hostile exchanges, and reduce users' willingness to participate in public discussions.

Our study focuses on racial justice, one of the most divisive issues in the United States. In 2021, the George Floyd protests sparked a racial reckoning across the country. On social media, discussions about race and racial justice, such as those with the hashtag #blacklivesmatter, surged in 2020 (Anderson et al., 2020). Yet racial attitudes remain divided in the U.S. According to a Pew Research Center survey in 2024, 80% of Democrats or Democratic-leaning independents say that White people benefit from advantages in society that Black people do not have, while only 22% of Republican or Republican-leaning independents expressed the same view (Pew Research Center, 2024).

Our study provides evidence on how these divisions are reflected in online interactions and how they shape the behavior of other users. We partner with Color of Change, the largest online racial justice organization in the United States. As a non-profit, Color of Change uses digital platforms to mobilize supporter and hold institutions accountable. We designed social media posts in line with Color of Change's brand guidelines to ensure the ecological validity of our study.

To reach a large audience, we deliver the posts as sponsored content on Facebook using Meta Ads Manager. Our design leverages social media advertising for several reasons. First, understanding the role of comment sections on advertisements is inherently important, as social media ads represent a significant share of the trillion-dollar advertising industry. In 2024, global social media ad spend exceeded $240 billion, including over $80 billion in the U.S.[5] While social media ads are proven effective, the role of user comments in enhancing or diminishing ad impact remains underexplored. Comments on ads allow users to share information and opinions, potentially influencing future viewers, akin to other forms of user-generated content (UGC). Despite significant interest from firms in leveraging UGC and social influence, there is little empirical evidence on the impact of social media comments on ad effectiveness. Understanding how these comments influence the effectiveness of marketing efforts is vital for brands striving to maintain a positive image and create meaningful connections with their audiences.

Second, using Facebook advertising allows us to access a large and diverse sample while minimizing experimenter demand effects. Recent work (e.g., Donati and Rao 2025; Donati et al. 2024) has explored social media ads as a research tool, as they enable the delivery of content to a broad audience and allow researchers to observe user behavior in the natural context in which the content is usually encountered.

Third, we develop a novel pipeline for comment section manipulation using existing features in the Meta Ads Manager. This enables us to manage comment sections across a large number of posts and systematically collect relevant data. Because our approach relies solely on tools already available on the platform, it yields practical insights for social media managers seeking to moderate and manage comment sections.

2.2 Post Design

To generate the comment sections used in our analysis, we first developed a series of designs for social media posts. These were created in collaboration with our NGO partner to ensure contextual relevance and consistency with brand guidelines. We then conducted a pre-test to assess clarity, engagement potential, and appropriateness. A subset of the highest-performing designs was subsequently used to create posts that were advertised to elicit organic comments, which served as the basis for our main experiment.

[5] `https://datareportal.com/reports/digital-2025-sub-section-global-advertising-trends`

2.2.1 Content Creation

To create professional content (ad creatives and copy), we hired four graphic designers, assigning each to focus on 2–3 key issues. These issues, identified in collaboration with our partner organization, Color of Change, included voter suppression, environmental justice, criminal justice and police reform, education reform, and technology fairness.[6] Each designer developed multiple concepts and taglines for their assigned ads. These were reviewed by the partner organization, which approved, revised, or rejected the proposals. Approved concepts were then finalized by our designers, adhering to the branding guidelines of the organization.

Appendix Figure A1 displays the final graphics (a total of ten, two for each issue) and their corresponding headlines exactly as they would appear to Facebook users. Each visual is designed to tell a compelling story about its assigned issue, emphasizing the urgency and importance of taking action. The content aims to engage users by sparking curiosity and encouraging further exploration of each topic. The use of bold imagery and compelling design elements is intended to capture attention while users browse their Facebook feed.

2.2.2 Pre-testing

We conducted several pre-tests to systematically select the posts used in the study and to refine the campaign parameters. In Pre-test A, we used Facebook's A/B testing tool across all 10 banners to identify, within each issue, which posts were most likely to generate a high number of clicks.[7] Appendix Table A1 summarizes the click-through rates (CTRs) – the ratio of link clicks over reach – for this test. These vary between 0.12% to 0.25%. For each issue, the banners with higher CTR are displayed on the right in Appendix Figure A1.

We conducted two additional tests, Pre-tests B and C, where we capped the frequency at one impression per person. Test B was conducted with a large potential audience (approximately 200,000 users per banner), while Test C targeted a smaller audience (approximately 6,000 users per banner). These adjustments were made to simulate a campaign that closely resembles the one planned for our main experiment.

Appendix Table A2 presents the aggregate results for Pre-tests B and C.[8] Notably, the CTRs for Pre-tests B and C are significantly lower than those reported for Pre-test A. This discrepancy arises because Pre-test A did not impose a frequency cap, allowing users to see each banner an average of two times and thereby increasing the likelihood of clicks. By contrast, Pre-tests B and C adopt a configuration similar to that used in the main experiment, in which we saturate an

[6] See Appendix A for a detailed description of these issues.

[7] In Pre-test A, we specified the audience (ZIP codes with a high share of progressive populations), budget ($50 per banner), duration (1 week), and optimization goal (reach), without imposing a frequency cap.

[8] The banners on education were excluded from these tests due to their low performance in Pre-test A and budgetary considerations.

audience group by imposing a frequency cap of one impression per user. While this approach may result in lower CTRs, it is essential to mitigate potential divergent delivery bias in Facebook A/B tests caused by the ad platform's algorithm (Burtch et al., 2025), as further discussed in Section 4.1.

2.3 Facebook Audience Selection

We reach Facebook users via sponsored content. A key advantage of social media advertising is that it enables both broad distribution and precise control over audience targeting (Aridor et al., Forthcoming).

To examine how responses vary by audience characteristics, we use ZIP codes as a targeting criterion. We group similar ZIP codes into strata based on observable characteristics, allowing us to compare outcomes across different types of communities while maintaining experimental control. The ZIP code characteristics were collected from several sources:

- **Meta Audience Estimates**: We use information on audience size provided by Meta, which reports the estimated number of users advertisers could potentially reach over a given period.[9] This data was collected through the Marketing API for each ZIP code on October 15, 2024.
- **Voting Behavior**: To proxy political preferences and ideology at the ZIP code level, we rely on the 2020 voting results. These come at the precinct level and were obtained from The Upshot.[10] We assigned each precinct to its nearest ZIP code according to Euclidean distance of the centroids using GIS software, and then aggregated voting information across all precincts matched with the same ZIP code.
- **Population and Racial Composition**: We use 2020 Census data to obtain information on the total and Black populations in each ZIP code and compute the share of Black residents.[11]

We categorize ZIP codes into three ideology groups based on the Republican vote share: *Blue* (Republican vote below 30%), *Swing* (between 45% and 55%), and *Red* (above 70%). Each ideology group was further divided into two subgroups based on racial composition (low and high share of Black population relative to the group median). Hence, a total of six ideology-race categories were created. Each ZIP code is used only once during our experiments, ensuring that audiences are exposed to the content in a single, well-defined condition.

[9] https://www.facebook.com/business/help/1665333080167380?id=176276233019487

[10] https://github.com/TheUpshot/presidential-precinct-map-2020

[11] https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2020&layergroup=ZIP+Code+Tabulation+Areas

2.4 Pre-Analysis Plan

The pre-analysis plan specified the intervention, the content creation process, estimating equations, moderators, and variable construction. The analysis follows the plan.

3 Generation and Analysis of Organic Engagement

To analyze the impact of comment sections, we begin with a large-scale social media campaign designed to elicit organic engagement with posts about racial justice. This initial phase generates user comments and reactions in a naturalistic setting, and provides direct evidence on how individuals engage with divisive content across audience types and topics. These organically generated engagement forms the basis for subsequent experimental manipulation.

Generating organic engagement is an important feature of our design for several reasons. First, organically generated comments reflect genuine, naturalistic user behavior, which is crucial for ensuring external validity. Unlike researcher- or AI-generated content, organic comments capture the authentic language, tone, and perspectives that users produce and encounter on social media. Second, these comments are themselves substantively important to study. By targeting specific audiences through Facebook's ad infrastructure, we can link engagement patterns to detailed audience characteristics – such as demographics and ZIP code–level ideology – providing richer insights than are typically available. Third, using organic comments enhances the ethical integrity of the experiment, as users interact with content created by other real users rather than being unknowingly exposed to artificially constructed narratives.

3.1 Methodology

We ran a Facebook ad campaign featuring five banners—one for each issue—that had achieved higher click-through rates (CTR) in pre-tests. Each banner promoted content related to racial justice, covering topics in *education*, *environmental justice*, *policing and criminal justice reform*, *technology fairness*, and *voting rights*. The campaign was optimized to maximize engagement with the posts—showing ads to users most likely to react, comment, or share—in order to collect authentic interactions that reflected spontaneous responses to important yet divisive content.

To examine variation in engagement across audiences and ad characteristics, we created 30 distinct audience strata. Each stratum consisted of sets of ZIP codes randomly sampled and grouped by ideological similarity and racial composition, with an estimated Facebook audience size of about 800,000 users on average.[12] The 30 strata corresponded to six combinations of political ideology (conservative, moderate, and liberal) and racial composition (above or below the median share of Black residents within each ideological category). For each of the six combinations (e.g., conserva-

[12]We excluded ZIP codes used in pre-tests.

tive areas with a below-median Black population share), we constructed five independent strata, yielding a total of 30 strata or audience groups. Within each stratum, we used Facebook's native A/B testing tool to randomly allocate users to be potentially exposed to one of the five issue banners. In total, the campaign included 150 posts (30 strata $\times$ 5 topics), reaching approximately 131,000 individuals and generating 12,000 reactions, 1,750 unique link clicks, 1,500 direct comments,[13] and 650 shares.

The randomization from A/B testing enables comparisons of engagement patterns across topics while holding audience composition constant. The resulting data allow us to characterize the intensity and nature of organic engagement and to identify patterns of interaction by political ideology and demographic composition. However, we do not interpret potential differences as causal effects of content. Although audiences are randomly assigned to potential exposure, actual exposure is determined by Facebook's ad-delivery algorithm, which endogenously allocates impressions based on predicted engagement probabilities Braun and Schwartz (2025). As a result, the observed engagement patterns reflect the joint influence of both content characteristics and algorithmic delivery.

Our analysis primarily focuses on engagement outcomes that are visible to other users, including comments and reactions. First, we compare overall engagement rates (expressed as a percentage of total reach) across ideological groups and genders, computing confidence intervals using standard errors clustered at the advertisement level (150 posts). Second, we assess the *valence* of these interactions. For reactions, we classify *likes*, *hearts*, and *cares* as supportive of the posts.[14] For comments, we use GPT–4 to categorize their valence. Finally, we examine additional textual dimensions that characterize comment tone and stance, including sentiment, offensiveness, and toxicity.

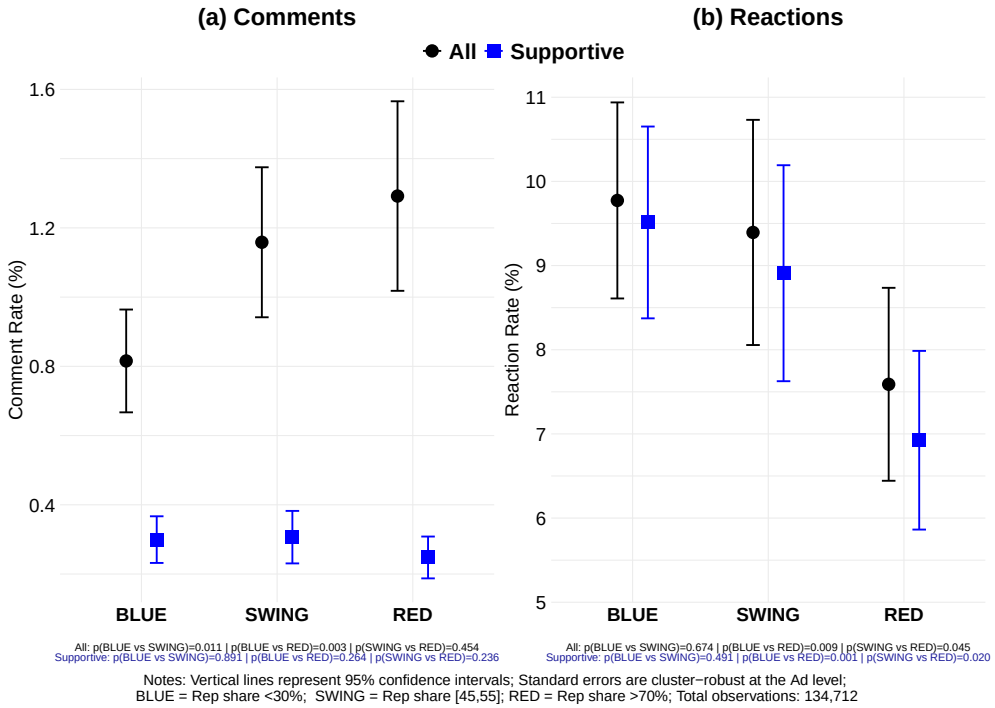
3.2 Engagement Rates

We first document patterns of engagement—both comments and reactions—across different areas. As shown in Figure 1, there are pronounced differences in how users engage with racial justice content across areas with different ideological compositions. When considering all interactions irrespective of their valence, comment rates increase substantially from liberal to conservative areas, rising from about 0.8 percent of reach in Blue ZIP codes to 1.3 percent in Red ones ($p < 0.01$). Reaction rates, by contrast, follow the opposite pattern, declining from roughly 10 percent in Blue areas to 7.5 percent in Red areas ($p < 0.01$). These differences indicate that users in more conservative areas are less likely to engage through quick, low-effort reactions but more likely to participate vocally by commenting on posts. In more progressive areas, engagement occurs

[13] Direct comments are those directed at the Facebook page/post itself, as opposed to replies to other users' comments.

[14] Other reaction types, such as *laugh*, *wow*, *sad*, and *angry*, are context-dependent and therefore harder to interpret; we focus on those that most reliably convey positive engagement.

Figure 1: Comment and Reaction Rates by Ideological Composition



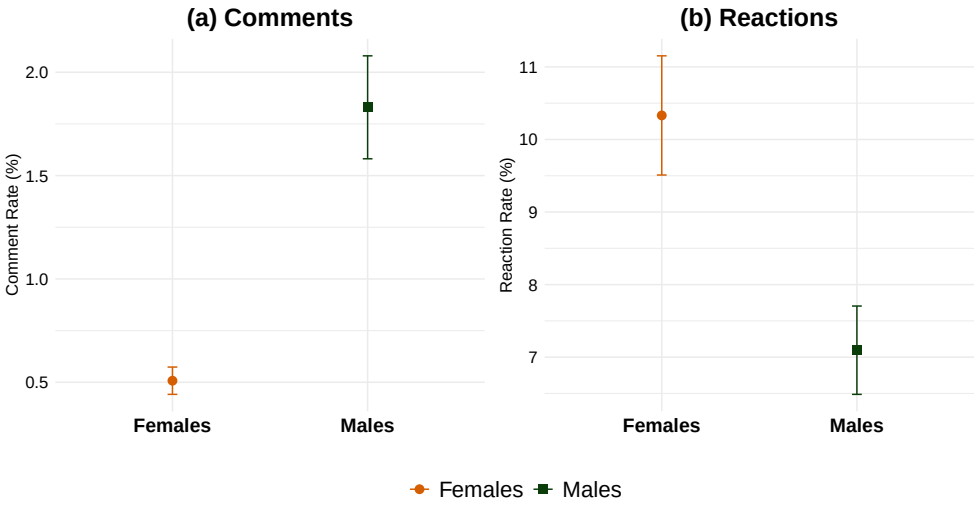
Notes: Vertical lines represent 95% confidence intervals; Standard errors are cluster−robust at the Ad level;
BLUE = Rep share <30%; SWING = Rep share [45,55]; RED = Rep share >70%; Total observations: 134,712

primarily through reactions, suggesting a more passive and silent mode of interaction. Because reactions are far more common than comments, summing the two measures across the subfigures indicates that users in progressive areas are, as expected, more likely to engage with racial justice posts overall.

When focusing on *supportive engagement*—defined as reactions or comments expressing agreement or approval—the ideological gradient is notably flatter for comments. Supportive comments remain consistently low across areas, ranging around 0.3 percent of reach, with no statistically significant differences between Blue, Swing, and Red ZIP codes ($p > 0.20$). In contrast, supportive reactions decline significantly from roughly 9.5 percent in Blue areas to about 7 percent in Red areas ($p < 0.01$), mirroring the overall reaction pattern. This suggests that while the overall volume of vocal participation (comments) rises in conservative areas, supportive responses remains relatively small and stable. Taken together, these results imply that ideological context shapes not only the intensity but also the *type* of engagement: audiences in liberal areas interact more through silent reactions, whereas those in conservative areas engage more vocally, using comments more frequently to express or debate opposing views.

We then compare engagement patterns by gender. Figure 2 reports comment and reaction rates for female and male users, with all rates expressed as a share of total reach. Men are substantially more likely to comment on posts than women: the average comment rate among men is more

Figure 2: Comment and Reaction Rates by Gender



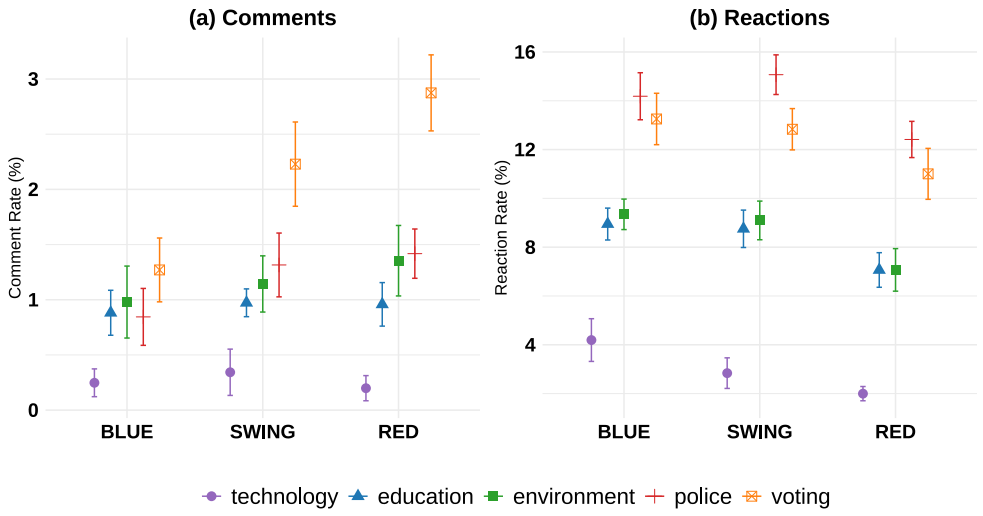
Notes: Vertical lines represent 95% confidence intervals.
Standard errors are cluster−robust at the Ad level.
Females: 75,451, Males: 59,329

than three times higher, at about 1.8 percent of reach compared to 0.5 percent among women. By contrast, reaction rates are about one and a half times higher among women, averaging 10.4 percent of reach compared to 7.1 percent among men. These differences are statistically significant and highlight a clear gender divide in the mode of engagement. Women are more likely to engage silently through quick, low-effort reactions, whereas men engage more vocally by commenting on posts. Moreover, since the previous analysis shows that most comments express disagreement with the posts, while most reactions are supportive, these gender differences suggest that men are more likely to express criticism through comments, whereas women are more likely to signal support through reactions. This pattern is consistent with well-documented gender differences in opinion expression, where men are disproportionately represented among those who criticize scientific manuscripts and presentations (Klinowski, 2023; Handlan and Sheng, 2023), and women tend to produce more favorable reviews than men (Bayerl et al., 2024). Our results show that even in online settings – where users are anonymous or interacting with strangers, and social or reputational costs are low – women are still significantly less likely to voice dissenting opinions.

We present the results separately for each issue in Figure 3. Issues such as voter suppression and police reform & criminal justice generate substantially more vocal engagement in Red ZIP codes, producing large differences relative to Blue ZIP codes. By contrast, topics such as education reform and technology fairness elicit lower levels of vocal engagement overall and exhibit little variation across areas. This pattern suggests that users in more conservative areas are not uniformly more vocal; rather, specific issues within the broader domain of racial justice appear to trigger heightened vocal engagement and often dissent. By contrast, reactions follow the general pattern in which users

in more progressive areas are more likely to react, and this pattern holds consistently across all issues.

Figure 3: Comment and Reaction Rates by Ideological Composition and Issue

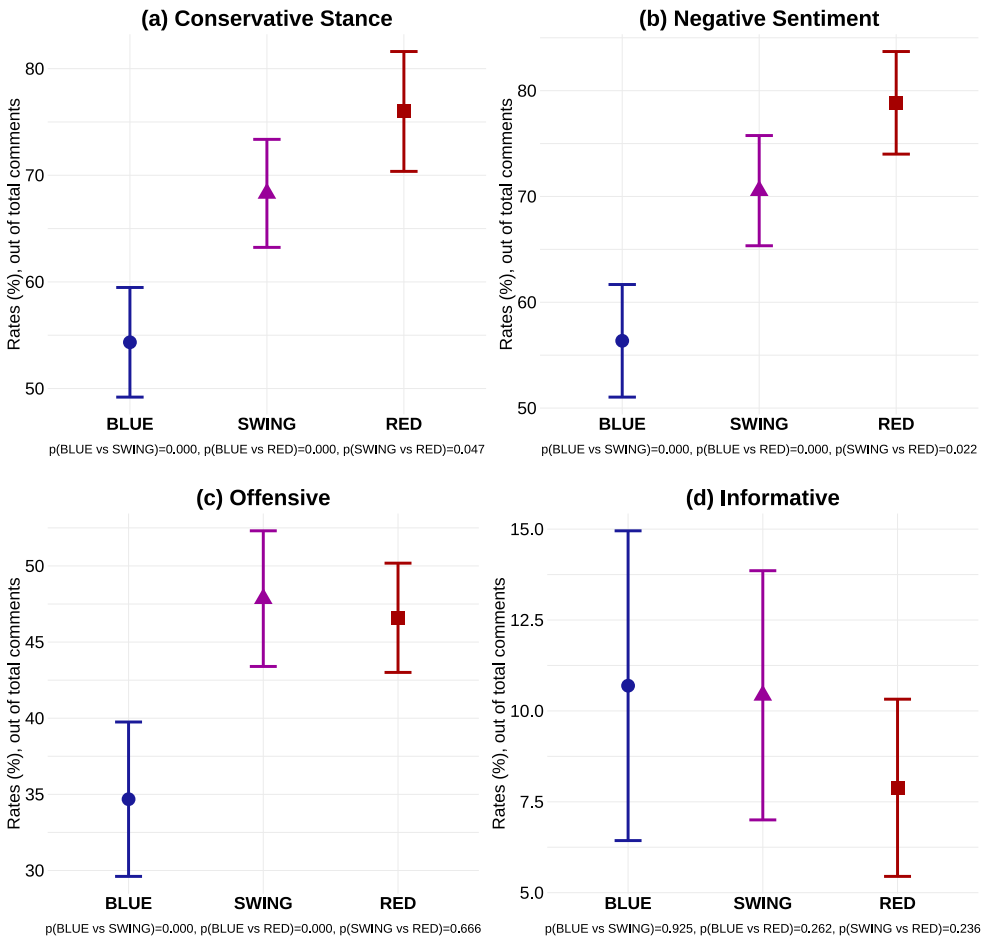


Notes: Vertical lines represent 95% confidence intervals.
Standard errors are cluster-robust at the Ad level.
Observations — technology: 32,350, education: 26,135, environment: 24,375, police: 26,458, voting: 25,394

3.3 Composition of the Comment Section

Figure 4 shows how the composition of comments varies across areas with varying ideological compositions, with all rates expressed as a share of total comments. Comments originating from conservative areas are substantially more likely to express a conservative stance and a negative sentiment. The share of conservative-leaning comments rises from about 55 percent in Blue ZIP codes to roughly 75 percent in Red areas (Panel (a), $p < 0.01$), while the share of comments with negative sentiment increases from around 56 percent to nearly 80 percent (Panel (b), $p < 0.01$). Panel (c) shows that the prevalence of offensive language increases from roughly 35 percent in Blue areas to almost 50 percent in Swing and Red areas, with a statistically significant difference relative to Blue areas ($p < 0.01$). Finally, Panel (d) indicates that the share of informative comments—those providing factual content or elaboration—remains relatively low, between 7.5 and 11 percent on average, with no statistically significant differences across contexts ($p > 0.10$). Taken together, these results suggest that conversations in conservative areas are more likely to adopt a critical or oppositional tone and to align with conservative viewpoints, but they are not necessarily less informative. In contrast, discussions in liberal areas feature a smaller share of ideologically charged and negative comments, suggesting a comparatively more neutral discussion environment.

Figure 4: Comments Characteristics by Ideology



Notes: Vertical lines represent 95% confidence intervals; Standard errors are cluster-robust at the Ad level;
BLUE = Rep share <30%; SWING = Rep share [45,55]; RED = Rep share >70%; Total observations: 1,393

4 The Causal Impact of the Comment Section

In this section, we provide novel causal evidence on the effect of the comment section. Specifically, we study how the presence and stance of a comment section influences subsequent users' engagement with the content.

4.1 Design

We manipulate the comments that participants see below our ads in a new ad campaign, using Meta A/B testing tool and an automated pipeline that hides new comments from users.

We investigate how exposure to the comment section and the different narratives expressed in the comment section of a post (collected in the ad campaign described above) affects individuals' subsequent engagement with that post (comments/views), as well as their intentions (clicks).

We are interested in the average treatment effect, as well as in the heterogeneous effects across individuals in different audience groups.

4.1.1 Intervention Design

To select and stratify audience types, we follow a similar approach as the one described in Section 3. We exclude ZIP codes used in the generation part, and create 18 audience strata, three for each ideology/race combination.

Within each audience group, we use A/B testing to randomly split its population into four conditions: no visible comments (No Comments), visible comments that include both opposing and supportive comments (Mixed), opposing comments only (Opposing), and supportive comments only (Supportive). Figure 5 provides an example of what the users can see, depending on the random assignment for their audience group.

Figure 5: Experimental Arms

(a) No Comments (b) Supportive (c) Mixed (d) Opposing

4.1.2 Issue Selection

From the five issues used in the comment generation phase, we selected environmental justice to focus on in order to maximize statistical power. This topic was chosen because it ranks near the middle in terms of overall engagement in the engagement generation campaign, ensuring sufficient variation in user responses without being dominated by extreme levels of attention or controversy. In addition, the topic remained timely and relevant at the time of the experiment.

4.1.3 Post Selection

We showed a subset of posts from the comment generation phase with existing interactions to new audience groups.

To select the posts for the treatment conditions, we identified triplets of posts using the following procedure. As part of the analysis in Section 3, comments are classified using a 5-point scale for political ideology: Strongly progressive or left-leaning, Slightly or moderately progressive, Centrist/unclear or no explicit stance, Slightly or moderately conservative, or Strongly conservative or right-leaning. In general, comments that support the original post or express pro-racial justice views are classified as progressive, while those that oppose the post or its message are classified as more conservative. To select posts for the Mixed, Opposing, and Supportive Comments conditions, we focused on posts that have at least two supportive and two opposing direct comments (i.e., comments that are directly responding to the post, rather than comments that reply to another comment). For each triplet of posts with similar number of reactions, we assigned one post to each of the Mixed Comments, Opposing Comments, or the Supportive Comments conditions.

For the No Comment condition, we selected posts that are not used in any other conditions and have a number of reactions similar to the average number of reactions in each triplet group.

4.1.4 Comment Selection

To select comments for the Opposing or Supportive Comments conditions, we chose two comments that align with the corresponding stance. For the Mixed Comments condition, we selected one opposing and one supportive comment. These selected comments remained visible to the audiences of the main experiment, while all other direct comments were hidden.

To better isolate the effect of the number of comments versus the narrative of the comments on outcomes, we manipulated the number of comments in the posts displayed to audience groups. In the No Comment condition, we deleted as many comments as possible so that the comment counter is zero or close to zero when it is first delivered to the audience in the experiment. [15] In the triplet of posts used for the treatment conditions, we equalized the number of initial comments across conditions by adding or deleting comments, while allowing the total number of comments to vary across triplets (ranging from 20 to 40).[16] This procedure ensured that the comment counter in the No Comment condition was substantially lower than counter displayed in the treatment conditions.

[15] Some comments, such as those violating Facebook policy, are not visible to the research team and therefore cannot be deleted. As a result, for some comment sections, it may not be possible to reduce the comment counter to zero.

[16] We also equalized the number of shares across posts by sharing them ourselves.

4.1.5 SUTVA Violations

A potential threat to the internal validity of social media experiments is the violation of the Stable Unit Treatment Value Assumption (Aridor et al., Forthcoming). This concern is particularly relevant to our research question, given the inherently social nature of comment sections. New comments posted by users could influence the perceptions or behaviors of other users, thereby confounding the treatment effects. For example, a new comment supportive of racial justice posted in the Opposing condition would alter the intended composition of the comment section.

To address this concern, we implemented a real-time, automated comment-hiding pipeline that immediately hides any new user comments from other users once they are posted.[17] This design ensured that the No Comment condition displayed no comments and that users in the treatment conditions saw only the comments corresponding to the assigned stance. It also helped minimizing interactions among users exposed to the same post, thereby mitigating possible SUTVA violations resulting from social interactions within the comment section.

4.1.6 Divergent delivery

Field experimentation in online display advertising presents several challenges to causal inference (Johnson, 2023). In particular, even within A/B tests, advertising platforms' algorithms may optimize campaign delivery over time for predicted user–ad relevance. As a result, different users can be targeted across experimental conditions based on engagement early in the campaign, generating algorithmic selection bias over time (Eckles, Gordon and Johnson, 2018; Ali et al., 2019; Braun and Schwartz, 2025). This threat to internal validity - the divergent delivery bias – poses a key concern when the goal is to identify the causal effect of specific ad features, rather than the joint effect of algorithmic delivery and ad features.

To minimize the risk of divergent delivery and ensure that our estimates reflect the causal effect of the comment section on behavior, rather than the effect of the comment section and platform delivery, we followed and augmented best practices from recent work (Burtch et al., 2025). First, we split budgets evenly across arms, launched all ads simultaneously, and capped exposure at one impression per user. Second, we optimize the campaign for reach rather than engagement. Third, we set a budget large enough to saturate the predefined audience and ran the campaign over multiple weeks to ensure that nearly all users within a defined area were reached. These design features minimize differences in ad delivery across conditions (Braun and Schwartz, 2025). Consequently, any observed differences across conditions can be more confidently interpreted as the causal effect of the comment section on user behavior, rather than as artifacts of algorithmic delivery dynamics.

Since we equalized the number of shares and average reactions across conditions and held the post content constant, the only element that varies across conditions is the comment section.

[17] The commenter will not know that their comment has been hidden; the comment remains visible to the commenter as well as to the researcher.

Together with the precautionary measures described above, this design allows us to isolate the effect of the comment section.

4.2 Data

Our data come from two sources. First, we obtain engagement metrics from the Meta Ads Manager dashboard,[18] which reports daily outcomes for each advertisement disaggregated by gender and age group. Second, we collect the full text and timestamps of all user-generated comments directly from the corresponding Facebook posts.

4.2.1 Outcomes

We analyze a set of engagement outcomes that capture both on-platform activity and off-platform behavior. All measures are expressed in *unique* terms, meaning that each user is counted at most once per outcome. Specifically, we focus on the following metrics:

- **Reach**, defined as the number of distinct users who were shown the ad at least once;
- **All engagement with the post**, capturing any form of user interaction with the ad;
- **Post expansions**, measuring whether users expanded the ad panel to view the comment section. This is the only outcome not directly available from the Meta Ads Manager interface. We construct this metric from other measures provided by Meta.[19]
- **Interactions**, defined as the sum of unique reactions (likes and other emoji responses), comments, and shares (reposting);
- **Link clicks**, indicating users' intent to learn more about the campaign by clicking on the external link;
- **Landing-page views**, recorded via the Facebook Pixel, capturing off-platform engagement with the organization's website.

Unless otherwise specified, we report the engagement rates computed as the ratio of each outcome to total reach.

[18] `https://www.facebook.com/business/tools/ads-manager`

[19] We compute *Unique Post Expansions* as

$$\text{Unique Post Expansions} = \text{Unique Clicks All} - \text{Unique Page Engagement}.$$

The *Unique Clicks All* metric aggregates all unique user clicks on an ad, including link clicks, profile clicks, reactions, comments, shares, saves, and other interactions. *Unique Page Engagement* is defined as the sum of all identifiable interactions with the post or the advertiser's page (link clicks, reactions, comments, shares, saves, and profile clicks). The residual therefore isolates users who expanded the ad container, serving as a proxy for interest in the comment section.

In addition to these behavioral outcomes, we collect the full text and timestamps of all user-generated comments. This allows us to examine not only the volume but also the *stance* of user discourse. We use a large language model (GPT-4) to classify each comment into categories of *supportive* or *non-supportive* toward the organization's message, based on its semantic content and sentiment. Analogously, we categorize user reactions according to their valence: "likes," "loves," and "cares" are coded as supportive, while other reaction types (such as "angry," "sad," or "wow") are coded as neutral or not supportive. These additional measures allow us to quantify how pre-existing comment narratives shape the tone and direction of subsequent engagement, thereby linking the stance of visible comments to the ideological composition and sentiment of later user responses.

4.2.2 Sample Characteristics

Table 1 reports descriptive statistics for the sample used in the main experiment. The final dataset comprises 1,054,015 unique users across all experimental conditions. Treatment assignment is well balanced across arms, with roughly one quarter of individuals allocated to each of the four conditions (Control, Supportive, Mixed, and Opposing comments).

Engagement outcomes exhibit substantial heterogeneity in magnitude and dispersion, reflecting the skewed nature of user activity on social media platforms. On average, 0.54 percent of reached users engaged with the post in any form, 0.27 percent expanded the ad panel to view the comment section, and 0.24 percent clicked on the external link. Interaction rates—comprising reactions, comments, and shares—averaged 0.02 percent of reach, while 0.18 percent of users visited the organization's landing page.

The demographic composition of the reached audience mirrors the U.S. Facebook user base. Approximately 52 percent of reached users were male and 48 percent female. The age distribution is centered on individuals aged 25–44, who account for over half of the total, while younger (18–24) and older (65+) users constitute smaller shares. Overall, these statistics confirm that our ads reached a demographically diverse sample representative of the platform's active user population.

Table 1: Summary statistics

Variable name	Mean (%)	St. Dev.
Treatment assignment		
Arm: Control (no comments)	25.019	43.312
Arm: Supportive	24.881	43.232
Arm: Mixed	24.971	43.284
Arm: Opposing	25.129	43.376
Main Outcomes		
All engagement	0.535	7.292
Post expansions	0.271	5.203
Interactions	0.022	1.487
Link clicks	0.240	4.889
Page views	0.183	4.276
Demographics		
Gender: females	47.672	49.946
Gender: males	52.328	49.946
Age: 18-24	8.692	28.172
Age: 25-34	30.165	45.897
Age: 35-44	27.158	44.478
Age: 45-54	16.262	36.902
Age: 55-64	10.241	30.319
Age: 65+	7.482	26.309
Observations:	**1,054,015**	

Notes: Values are expressed as percentages relative to reach.

4.2.3 Balance Checks

Table 2 reports covariate balance across the four experimental conditions. The sample comprises approximately 1.05 million individuals, evenly distributed across treatment arms, with about 263,000 users per group. We examine three observable individual-level characteristics: gender, middle-age status (ages 35–64), and senior status (ages 65 and above). Mean values and standard deviations are shown by group, and pairwise differences with the control arm are tested using two-sample t-tests with standard errors clustered at the advertisement level (see Section 4.3 for details).

Across all covariates, differences between treatment and control groups are small in magnitude and statistically insignificant. The p-values from the corresponding tests uniformly exceed conventional significance thresholds, indicating that random assignment produced well-balanced groups across key demographic dimensions. This balance supports the internal validity of the experimental design and suggest that any subsequent differences in engagement outcomes can be attributed to the randomized variation in comment visibility and stance, rather than to pre-existing differences in audience composition.

Table 3 reports balance checks for ad-level cost and performance metrics across the four experi-

Table 2: Balance Checks: Individual-level Covariates

	Group Mean / (SD)				*t*–test *p*–value		
Variable	(1) Control	(2) Supportive	(3) Mixed	(4) Opposing	(1)–(2)	(1)–(3)	(1)–(4)
Male	0.522 (0.500)	0.522 (0.500)	0.524 (0.499)	0.525 (0.499)	0.983	0.899	0.878
Middle Aged (35-64)	0.538 (0.499)	0.536 (0.499)	0.537 (0.499)	0.535 (0.499)	0.868	0.944	0.813
Senior (65+)	0.074 (0.262)	0.076 (0.266)	0.075 (0.263)	0.074 (0.262)	0.736	0.938	0.973
Observations	263,706	262,246	263,197	264,866			

Notes: Each observation is a user. Standard errors in the t-tests are clustered at the advertisement level (72 ads).

mental conditions. Each observation corresponds to one advertisement, for a total of 72 ads evenly distributed across treatment arms. We compare total spend, cost per mille (CPM), frequency, reach, and spend per user to verify that Meta's delivery algorithm exposed ads in each treatment arm to comparable audience sizes and costs.

Mean values are virtually identical across groups, and none of the pairwise differences relative to the control group are statistically significant. Total spend per ad averages approximately $150, with CPMs around $9.6 and mean reach near 14,600 users. The estimated p-values for all tests are well above conventional significance thresholds, confirming that the experimental conditions were implemented under comparable delivery and budget parameters. These results indicate that Meta's optimization algorithm did not differentially allocate impressions or spending across treatment arms, reinforcing the internal validity of our causal design.

The frequency and reach metrics further support the correct implementation of the experimental design. Our objective was to saturate audiences such that each individual would be reached at most once. The observed average frequency of approximately 1.11, combined with an upper-bound estimated audience size of about 13,750 users per ad, is consistent with near-complete audience saturation and balanced delivery across conditions.

4.3 Empirical Strategy

We examine how the presence and stance of the comment section affect subsequent user engagement with a social media post. The experiment randomizes the post's comment section across 18 predefined audience strata, defined as sets of ZIP codes grouped by ideological and racial composition. Within each stratum, individuals are randomly assigned to one of four ads, corresponding to the treatment conditions: *No Comments* (control), *Opposing*, *Supportive*, and *Mixed*.

At the aggregate level, we observe the outcomes of 72 ads—one for each combination of audience

Table 3: Balance Checks: Ad-level Delivery Metrics

	Group Mean / (SD)				t–test p–value		
Variable	(1) Control	(2) Supportive	(3) Mixed	(4) Opposing	(1)–(2)	(1)–(3)	(1)–(4)
Total Spend	150.763 (1.028)	150.276 (1.202)	150.395 (0.722)	150.542 (1.078)	0.196	0.218	0.531
CPM	9.594 (2.130)	9.639 (2.075)	9.626 (2.065)	9.548 (2.046)	0.949	0.964	0.947
Frequency	1.123 (0.026)	1.118 (0.020)	1.116 (0.023)	1.118 (0.023)	0.502	0.356	0.520
Reach	14650.333 (3234.248)	14569.222 (3150.366)	14622.056 (3136.580)	14714.778 (3103.458)	0.939	0.979	0.952
Spend per User	0.011 (0.003)	0.011 (0.002)	0.011 (0.002)	0.011 (0.002)	0.994	0.952	0.894
Observations	18	18	18	18			

Notes: Each observation is an ad. Standard errors in the t-tests are heteroskedasticity-robust.

stratum $z \in \{1, \ldots, 18\}$ and treatment condition $k \in \{\text{No Comments}, \text{Opposing}, \text{Supportive}, \text{Mixed}\}$. Each ad corresponds to a social media post delivered to a specific audience stratum under a specific treatment condition. For each ad, we record the total number of unique individuals reached and the total number who engaged in a given action (e.g., clicking the link or reacting to the post). These outcomes are further disaggregated by day, gender, and age group. In the main analysis, we aggregate the data across days so that each observation reflects the full duration of the campaign.

To analyze treatment effects at the individual level, we construct a synthetic dataset in which each observation represents one individual exposure to an ad. Let p_z^k denote the observed proportion of individuals in stratum z assigned to treatment condition k who took the action. We model the individual-level outcome Y_{iz} as a realization of a Bernoulli random variable:

$$Y_{iz} \sim \text{Ber}(p_z^k),$$

where individual i is assigned to treatment condition k within stratum z. For the purpose of generating this synthetic microdata, we assume that individual observations are independently and identically distributed within each (k, z) cell, with mean p_z^k. This assumption is justified by the random assignment of individuals to treatment conditions within each stratum. The synthetic sampling approach enables estimation of treatment effects using individual-level regressions, despite the availability of only aggregate data (Eckles, Karrer and Ugander, 2017; Gordon et al., 2019).

We estimate the following linear probability model via ordinary least squares:

$$Y_{iz} = \alpha + \beta_k T_i^k + X_{iz}'\gamma + \delta_z + \varepsilon_{iz}, \tag{1}$$

where Y_{iz} is the simulated outcome of individual i in stratum z; T_i^k is a binary indicator equal to one if individual i is assigned to treatment condition k; X_{iz} is a vector of individual-level covariates including age and gender; δ_z denotes stratum fixed effects; and ε_{iz} is an individual-level error term. The set of treatment conditions is defined as $k \in \{\text{Opposing}, \text{Supportive}, \text{Mixed}\}$, with the *No Comments* condition omitted and serving as the reference category. The coefficients β_k thus capture the causal effect of each comment stance relative to the control.

For inference, we cluster standard errors at the stratum–treatment level, corresponding to the 72 unique ads in the experiment. While the assumption that observations are independently and identically distributed within each (z, k) cell is required to simulate individual-level outcomes, it is not required for valid inference. Clustering is critical in our context because all individuals in a given stratum–treatment cell belong to the same ZIP code audience group and are exposed to the exact same ad content—including the same comments—which may induce correlation in their responses. By clustering at the (z, k) level, we allow for arbitrary dependence in outcomes within each treatment cell and ensure conservative inference even in the presence of correlated behavior among individuals who viewed the same ad.

We also examine heterogeneous treatment effects across key subgroups of interest. Specifically, we explore differential responses by gender and age, as well as across audience strata defined by prevailing political ideology. To estimate subgroup-specific effects, we re-estimate regression (1) separately within each subgroup. This approach allows us to assess whether individuals respond differently to the narrative framings depending on their demographic characteristics or the ideological orientation of the area in which they reside. All models are estimated using the same linear probability specification as in the main analysis, and standard errors are clustered at the stratum–treatment level.

The only exception concerns the analysis of the valence of subsequent interactions. Because the stance of comments and the types of reactions are not available from the Meta Ads Manager but are instead retrieved directly from the posts, we cannot assign these outcomes to a specific stratum (z) or include user-level controls. Consequently, the model specification and level of clustering for these outcomes differ slightly from those used in the main analysis. Further details are reported in the corresponding result tables.

As a robustness check, we verify that our results are not sensitive to the choice of the linear probability model by re-estimating treatment effects using a logistic regression specification. In addition, we present model-free evidence by directly comparing mean outcomes Y across individuals assigned to the four experimental conditions, without relying on the simulated individual-level observations. In this case, statistical significance of the differences across groups is assessed using a χ^2 test for equality of proportions with continuity correction (Yates, 1934). This nonparametric approach provides a transparent benchmark for evaluating the magnitude and direction of treatment effects without imposing functional form assumptions.

4.4 Results

4.4.1 Main Results

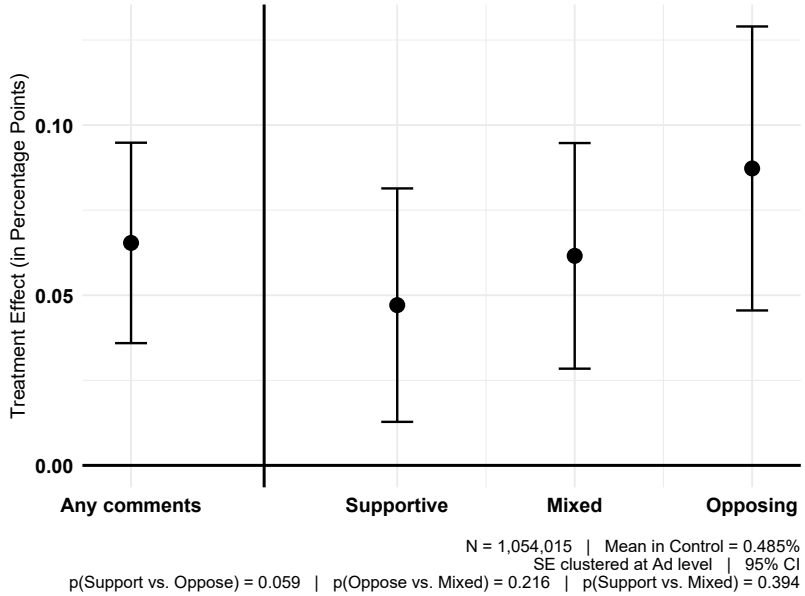
We present estimates of the causal impact of the comment section on user engagement, obtained from the linear probability model described above, with standard errors clustered at the advertisement level. To aid interpretation, effects are reported in percentage points (pp) and as percent changes relative to the baseline mean in the *No Comments* condition. Figure 6 presents results for overall engagement, while Figure 7 summarizes the effects on post expansions, interactions, link clicks, and landing-page views. Full regression tables are reported in the Appendix Tables A3 and A4.

All Engagement. Figure 6 shows that displaying any comments increases total engagement by 0.065 pp ($p < 0.01$), corresponding to a 13.4 percent increase relative to the control mean of 0.485 percent. By stance, Opposing comments generate the largest increase (0.087 pp, $p < 0.01$; 17.9 percent), followed by Mixed (0.062 pp, $p < 0.01$; 12.8 percent) and Supportive (0.047 pp, $p < 0.01$; 9.7 percent). The difference between Opposing and Supportive comments is statistically significant at the 10 percent level ($p = 0.059$) and sizable, as the effect of Opposing comments is nearly twice as large. The results indicate that not only does the presence of comments matter for subsequent engagement, but also their stance. In particular, critical or contentious remarks draw substantially more overall engagement than supportive ones, suggesting that negative commentary tends to amplify user activity around divisive content. We next examine which specific actions drive this pattern.

Post Expansions. Figure 7(a) shows that any comments increase the probability that users expand the ad panel to view the comment section by 0.046 pp ($p < 0.01$), corresponding to a 19.4 percent increase over the baseline mean of 0.237 percent. All three stance conditions produce similar effects—Opposing (0.044 pp, $p < 0.01$), Mixed (0.047 pp, $p < 0.01$), and Supportive (0.048 pp, $p < 0.01$)—and none of the pairwise differences are significant ($p > 0.80$). This pattern is consistent with expectations: users can only observe the stance of the comments after expanding the post, so the decision to view the comment section should not differ systematically across comment types.

Interactions. Figure 7(b) indicates that the pooled "any comments" effect on reactions, comments, and shares is small and statistically insignificant (0.003 pp). However, stance-specific estimates reveal that Opposing comments substantially increase interactions by 0.009 pp ($p < 0.05$), a 45 percent rise relative to the control mean of 0.020 percent. Mixed comments have a modest, insignificant effect (0.002 pp), while Supportive comments slightly reduce interactions (-0.003

Figure 6: The Impact of the Comment Section on All Subsequent Engagement



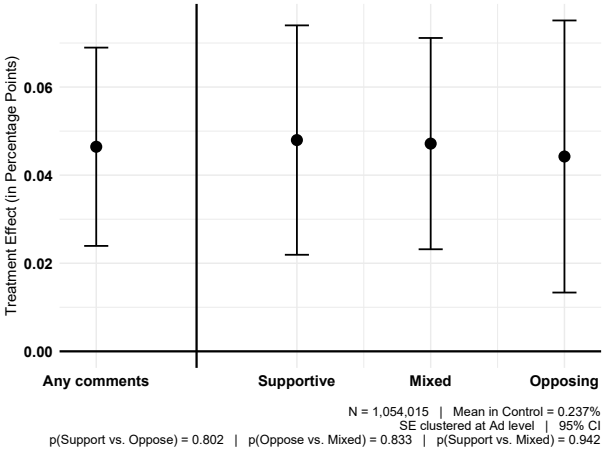
pp, not significant). Differences between Opposing and Supportive conditions are highly significant ($p = 0.002 < 0.01$), whereas differences involving the Mixed condition are smaller and only marginally significant ($p \approx 0.09$). This pattern suggests that antagonistic or contentious comments elicit greater visible participation from subsequent users.

Link Clicks. Figure 7(c) shows that the presence of any comments raises the probability of clicking on the external link by 0.016 pp ($p < 0.05$), representing a 7.0 percent increase over the baseline mean of 0.228 percent. Opposing comments again produce the largest effect (0.034 pp, $p < 0.01$; 14.9 percent), while Mixed (0.012 pp) and Supportive (0.003 pp) are smaller and not significant. The difference between Opposing and Supportive comments is statistically significant ($p = 0.008 < 0.01$), whereas comparisons involving the Mixed condition are not ($p > 0.10$). Hence, negative comments are particularly effective in driving click-through engagement.

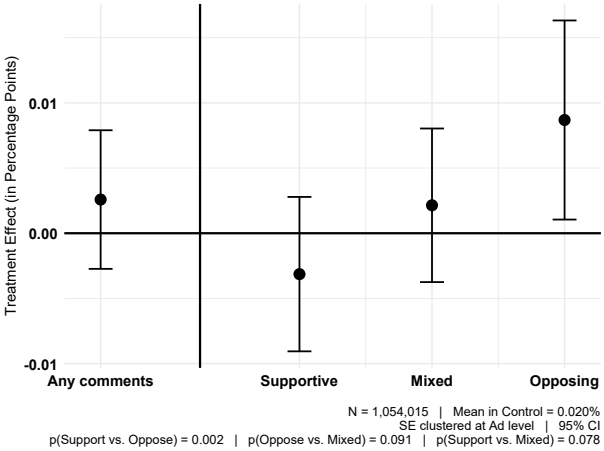
Landing-Page Views. Figure 7(d) suggests that the presence of any comments increases off-platform engagement by 0.016 pp ($p < 0.10$), corresponding to a 9.4 percent gain relative to the control mean of 0.171 percent. Opposing comments yield the largest increase (0.019 pp, $p < 0.10$; 11.1 percent), followed by Mixed (0.017 pp) and Supportive (0.011 pp), both not significant. However, differences across stances are small and statistically indistinguishable ($p > 0.50$), indicating that while comments modestly enhance downstream engagement, their stance might have a limited influence on conversions beyond the on-platform engagement.

Figure 7: The Impact of the Comment Section on Selected Outcomes (as % of total reach)

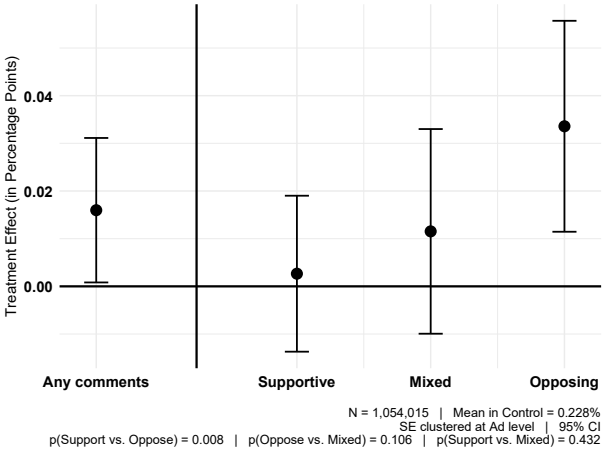
(a) Post Expansions



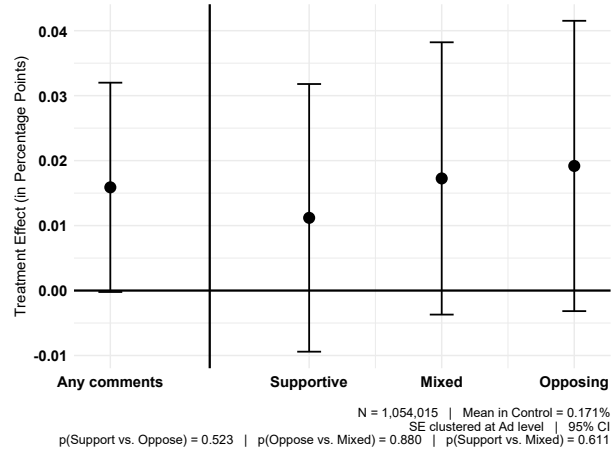
(b) Interactions (reactions, comments, shares)



(c) Link Clicks



(d) Landing-page Views



Discussion. Across all outcomes, the presence of a comment section modestly raises user engagement, both on- and off-platform. However, the overall increase is largely driven by higher attention to the post itself—as reflected in post expansions—which does not vary with the stance of existing comments. By contrast, comment stance shapes the nature of subsequent engagement. Opposing comments consistently generate higher rates of interactions and link clicks, while Supportive comments do not significantly outperform the control, and differences between Mixed and Supportive tones are generally negligible. Taken together, the results indicate that negative or contentious discourse amplifies participation and interest more effectively than positive or supportive comments, highlighting a potential engagement–polarization trade-off in online comment sections.

4.4.2 Valence of Subsequent Interactions

We next examine how the stance of the comment section affects the tone of subsequent engagement. Specifically, we compare the effects on *all interactions*—the total number of user reactions, comments, and shares—with those on *supportive interactions*, which include positive reactions (*likes*, *loves*, *cares*) and comments expressing agreement with the organization's message. Because certain reaction types (such as *laugh* or *wow*) are ambiguous in tone, we interpret results for non-supportive or neutral interactions with caution and report them in Appendix Table A6 for completeness. The estimates are obtained from a parsimonious specification using data retrieved directly from the posts, without user-level covariates, and standard errors are corrected for heteroskedasticity.

Figure 8: The Impact of the Comment Section on All and Supportive Interactions

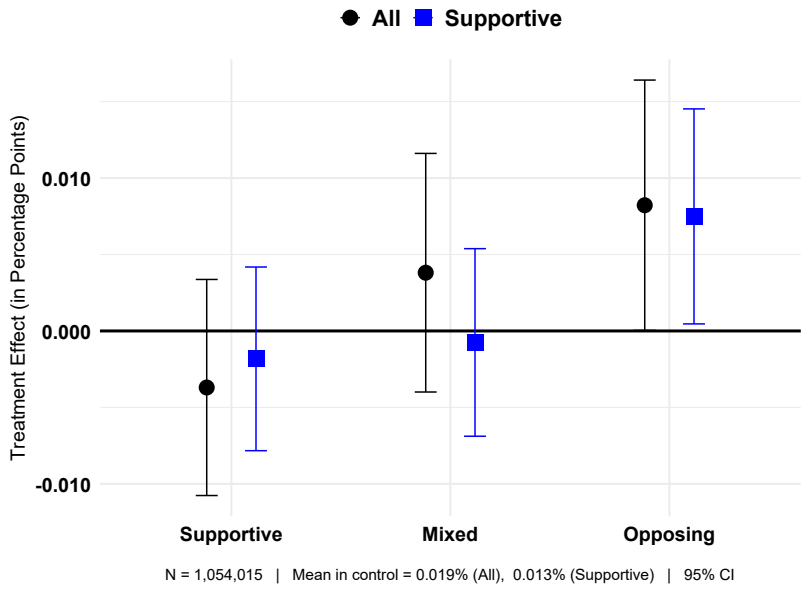


Figure 8 compares the effects of different comment stances on overall and supportive interactions.[20] Opposing comments significantly increase both measures, raising the rate of all interactions by 0.0082 percentage points (pp, $p < 0.05$) and supportive interactions by 0.0075 pp ($p < 0.05$), which corresponds to a 56 percent increase relative to the control mean. In contrast, Mixed and Supportive comment sections have no measurable effect on total or supportive engagement. Pairwise tests confirm that Opposing comments generate significantly more supportive follow-up activity than either Mixed ($p = 0.020$) or Supportive ($p = 0.008$) conditions, suggesting that critical remarks can mobilize users to express agreement rather than detachment.

[20]Results for all interactions are very similar but not identical to those reported in Section 4.4.1. In this case, reactions and comments were collected directly from the posts rather than from the Meta Ads Manager dashboard. Minor discrepancies may arise if users deleted their comments or removed reactions after the initial data extraction, leading to slight differences between datasets.

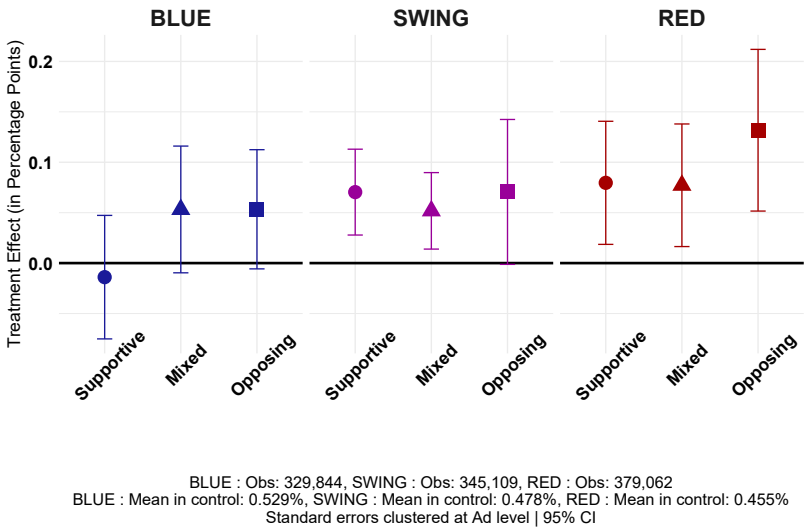
For completeness, Appendix Table A6 reports results for other or non-supportive interactions, which include reactions such as *angry*, *sad*, and ambiguous responses. The estimated effects are small and statistically imprecise. Only the Mixed condition shows a marginally significant increase (+0.0046 pp, $p < 0.10$), while Opposing and Supportive conditions have no discernible impact.

Overall, these results reinforce the finding that exposure to opposing or contentious remarks heightens engagement primarily through supportive expressions. Rather than fostering hostility, critical comments appear to activate a form of reactive solidarity, leading users to respond positively to the organization's message. This behavioral pattern is consistent with theories of psychological reactance and reinforcement, which posit that exposure to opposing views can strengthen existing attitudes and stimulate expressive engagement among like-minded individuals (Klapper, 1960; Brehm, 1966).

4.4.3 Heterogeneous Effects Across Political Ideology

We next examine whether the impact of comment stance varies across areas with different prevailing political ideologies, distinguishing between *Blue* (mostly liberal), *Swing* (mixed), and *Red* (mostly conservative) ZIP codes. Figure 9 displays the heterogeneous treatment effects on overall engagement, and Figure 10 summarizes the corresponding effects for post expansions, interactions, link clicks, and landing-page views. The detailed point estimates are reported in Appendix Tables A7, A8, and A9. Across all contexts, the presence of a comment section increases engagement on average, but the magnitude and statistical precision of these effects rise sharply with the conservativeness of the area.

Figure 9: Heterogeneous Treatment Effects on All Engagement across Political Ideologies



In *Blue areas*, effects are small and generally imprecise. Opposing and Mixed comments mod-

estly increase overall engagement by 0.05 pp ($p < 0.10$), corresponding to roughly a 10 percent rise relative to the control mean of 0.53 percent, whereas Supportive comments have no measurable effect. No statistically significant differences emerge across stances for any specific outcome. These patterns suggest that in liberal environments, exposure to comment sections—regardless of the tone of existing comments—does not meaningfully alter user behavior, consistent with audiences already predisposed to engage with the campaign's message.

In *Swing areas*, engagement becomes more responsive to the presence of a comment section. All three comment stances produce statistically significant increases in total engagement, with Opposing, Mixed, and Supportive comments raising overall activity by 0.07, 0.05, and 0.07 percentage points (pp), respectively. The strongest effects are observed on link clicks, where Opposing comments increase click-through rates by 0.037 pp ($p < 0.05$), a roughly 18 percent rise relative to the baseline, and on page views, where Mixed comments increase website visits by 0.044 pp ($p < 0.01$), corresponding to an increase of about 29 percent relative to control. Although differences between stances remain small, this pattern suggests that users in politically mixed areas respond primarily to the visibility of comments rather than to their ideological orientation, possibly perceiving the existence of discussion itself as a signal of relevance.

In *Red areas*, treatment effects become both larger and more differentiated across comment stances. Opposing comments generate the strongest responses across all outcomes: total engagement rises by 0.13 percentage points (pp, $p < 0.01$), post expansions by 0.07 pp ($p < 0.05$), link clicks by 0.045 pp ($p < 0.05$), and page views by 0.040 pp ($p < 0.05$). These effects correspond to relative gains of roughly 20–25 percent compared with baseline means. Mixed and Supportive comments also increase engagement, but the magnitudes are smaller and less precisely estimated. Pairwise tests confirm that the differences between Opposing and Supportive conditions are statistically significant for interactions ($p < 0.01$) and link clicks ($p < 0.05$), indicating that critical or contentious remarks mobilize higher levels of participation and intent, particularly among users in conservative areas.

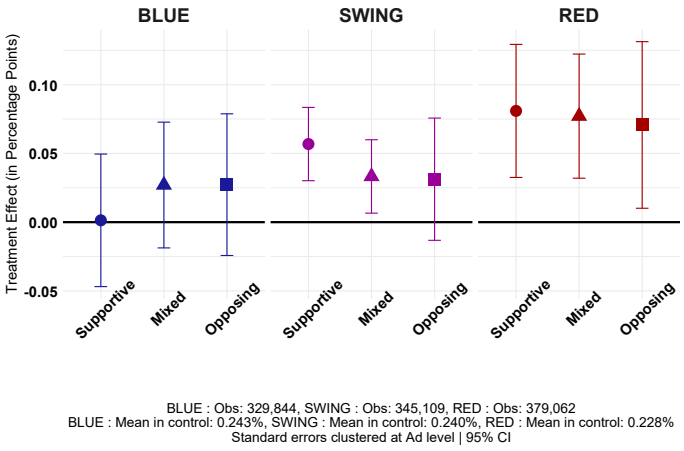
Baseline engagement levels also vary systematically with ideology. In the control condition, overall engagement averages 0.53 percent in Blue areas, 0.48 percent in Swing areas, and 0.46 percent in Red areas, reflecting greater inherent alignment between the campaign's message and audiences' ideology in more progressive areas.[21] Comment sections therefore play a more pronounced role in conservative areas, where baseline interest is lower: by displaying visible discussion—even when critical—they make the content more salient and relevant to users who might otherwise overlook it.

Taken together, these results reveal a clear ideological gradient. While engagement in liberal and politically balanced areas responds similarly across comment tones, user activity in conser-

[21]This pattern is consistent with prior literature (e.g., Song 2024 in the racial justice context), which finds that individuals are more likely to engage with social media content that aligns with their preexisting attitudes.

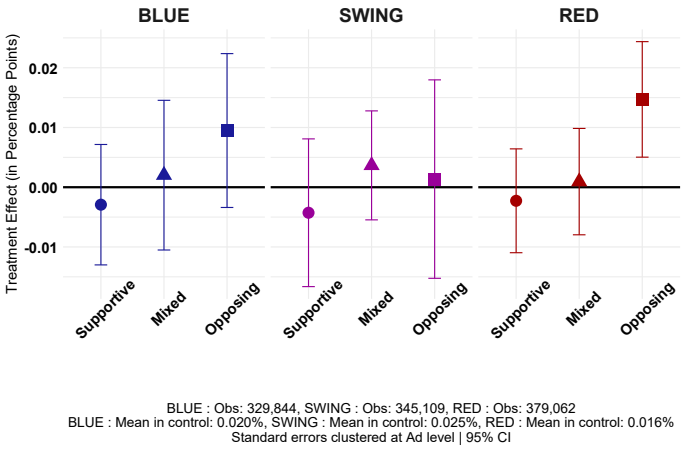
Figure 10: Heterogeneous Treatment Effects on Selected Outcomes across Political Ideologies

(a) Post Expansions



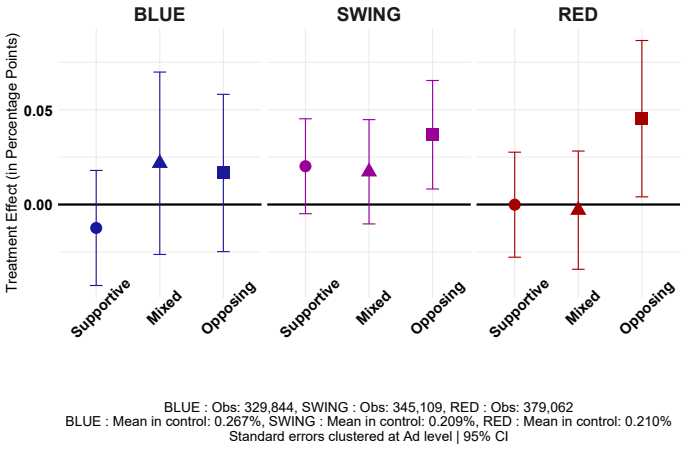
BLUE : Obs: 329,844, SWING : Obs: 345,109, RED : Obs: 379,062
BLUE : Mean in control: 0.243%, SWING : Mean in control: 0.240%, RED : Mean in control: 0.228%
Standard errors clustered at Ad level | 95% CI

(b) Interactions (Reactions, Comments, Shares)



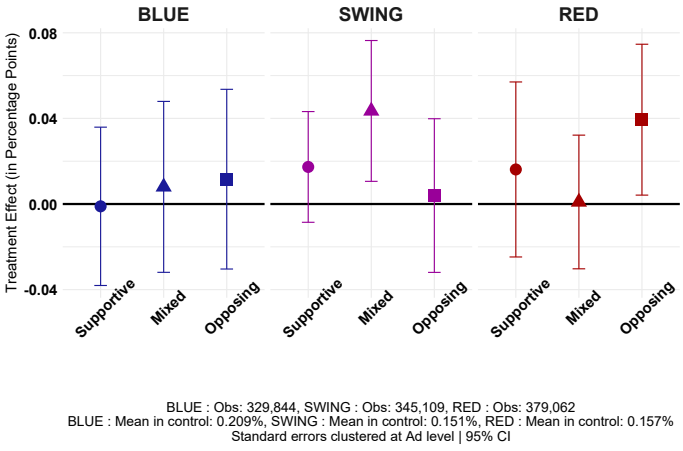
BLUE : Obs: 329,844, SWING : Obs: 345,109, RED : Obs: 379,062
BLUE : Mean in control: 0.020%, SWING : Mean in control: 0.025%, RED : Mean in control: 0.016%
Standard errors clustered at Ad level | 95% CI

(c) Link Clicks



BLUE : Obs: 329,844, SWING : Obs: 345,109, RED : Obs: 379,062
BLUE : Mean in control: 0.267%, SWING : Mean in control: 0.209%, RED : Mean in control: 0.210%
Standard errors clustered at Ad level | 95% CI

(d) Landing-page Views



BLUE : Obs: 329,844, SWING : Obs: 345,109, RED : Obs: 379,062
BLUE : Mean in control: 0.209%, SWING : Mean in control: 0.151%, RED : Mean in control: 0.157%
Standard errors clustered at Ad level | 95% CI

vative regions becomes markedly more sensitive to opposing narratives. The same social media feature—the comment section—thus amplifies attention and interaction most strongly where ideological agreement with the content of the post is low, turning critical or contentious commentary into a mechanism for mobilizing attention rather than discouraging it. This pattern is consistent with a *curiosity mechanism*, in which exposure to comments that challenge the post's viewpoint sparks user interest and prompts additional exploration, such as expanding the post or clicking through to learn more.

When viewed alongside the valence analysis in Section 4.4.2, this heterogeneity may also point to a complementary mechanism of *reactance*. Opposing comments tend to increase the *volume* of engagement—especially in conservative areas—and the additional interactions they generate

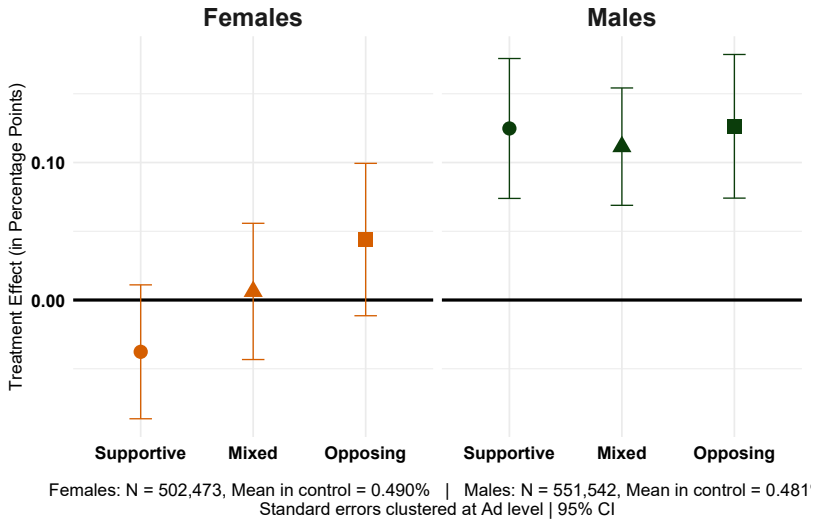
appear, on average, to be more *supportive* in tone. Such a pattern would be consistent with theories of psychological reactance and reinforcement, which posit that exposure to opposing views can trigger expressive behaviors aimed at reaffirming one's prior attitudes (Brehm, 1966; Klapper, 1960).

However, because we cannot directly match the stance of individual comments or reactions to a stratum, we cannot determine whether the additional engagement in conservative areas is indeed supportive. To further investigate these potential mechanisms, we will conduct an artefactual field experiment described in Section 4.5 to evaluate whether curiosity and reactance operate as the channels through which comment sections shape engagement.

4.4.4 Heterogeneous Effects across Genders

We next examine whether the impact of comment stance differs across gender, distinguishing between male and female users. Figure 11 displays the heterogeneous treatment effects on overall engagement, while Figure 12 summarizes the corresponding effects for post expansions, interactions, link clicks, and landing-page views. The detailed point estimates are reported in Appendix Tables A10 and A11. Across outcomes, we find little evidence that comment sections meaningfully affect engagement among women, whereas the effects are large, precise, and positive among men, particularly for opposing and mixed comment stances.

Figure 11: Heterogeneous Treatment Effects on All Engagement Across Genders

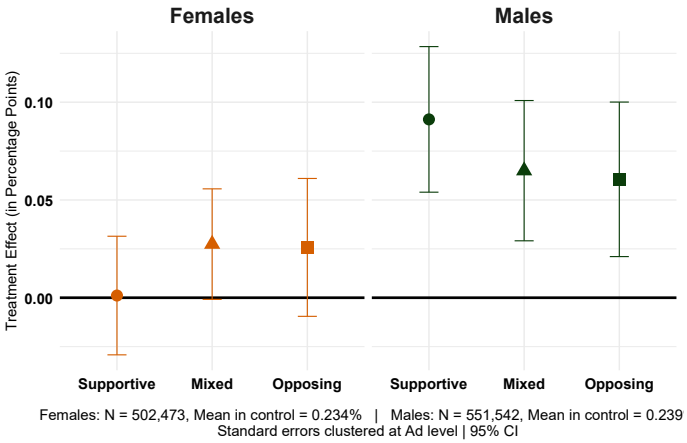


Among *female* users, estimated effects are small and often statistically insignificant. Opposing comments increase overall engagement by 0.044 percentage points (pp), corresponding to roughly a 9 percent rise relative to the control mean of 0.49 percent, but the effect is not statistically significant. Mixed comments have no measurable impact on total engagement, while Supportive comments

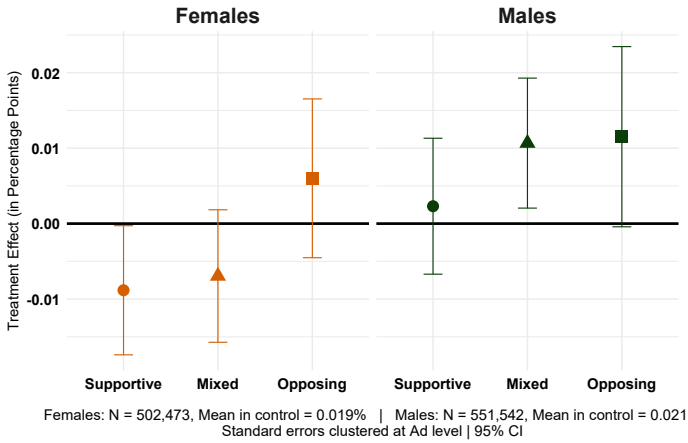
are associated with small negative coefficients across most outcomes. For example, Supportive comments reduce link clicks by 0.032 pp ($p < 0.05$) and interactions by 0.009 pp ($p < 0.05$), indicating that positive or reinforcing comment sections may slightly reduce active engagement among women. Pairwise tests confirm that differences between Opposing and Supportive conditions are statistically significant for interactions ($p = 0.002$) and link clicks ($p = 0.009$), largely driven by the decline under the Supportive condition rather than gains from Opposing comments. Overall, female engagement appears relatively unresponsive to comment visibility or stance.

Figure 12: Heterogeneous Treatment Effects on Selected Outcomes Across Genders

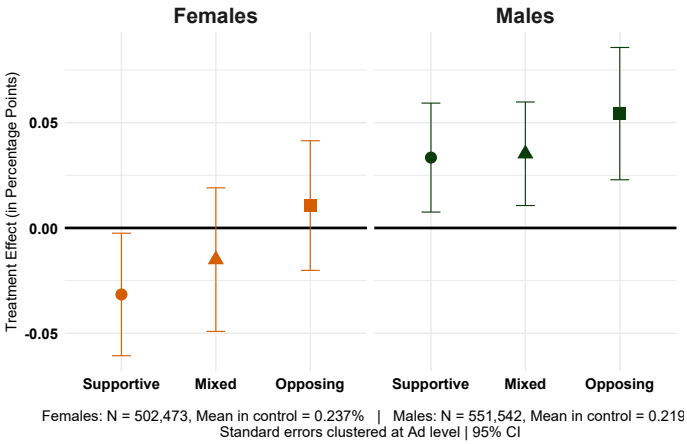
(a) Post Expansions



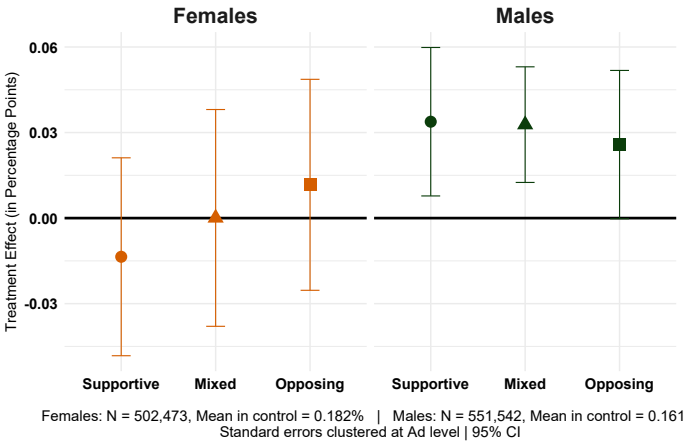
(b) Interactions (Reactions, Comments, Shares)



(c) Link Clicks



(d) Landing-page Views



In contrast, *male* users exhibit large and precisely estimated effects across all outcomes. Opposing comments increase overall engagement by 0.126 percentage points (pp, $p < 0.01$), a 26 percent gain relative to the control mean of 0.48 percent. Mixed and Supportive comments also

raise engagement by 0.112 pp ($p < 0.01$) and 0.125 pp ($p < 0.01$), respectively, corresponding to 23–26 percent increases. While the presence of comments increases the likelihood of post expansion regardless of stance, only Opposing and Mixed comments significantly raise interaction rates (+0.012 and +0.011 pp, respectively). All comment stances increase link-click and page-view activity among men, with Opposing comments raising click-through rates by 0.054 pp ($p < 0.01$), a roughly 25 percent rise relative to baseline. These patterns indicate that men are considerably more responsive both to the visibility and to the stance of comments. They engage vocally when presented with critical or diverse narratives, whereas their less visible actions—such as clicking or browsing—do not appear to depend on comment stance.

Taken together, these results reveal a clear gender gradient in responsiveness to online discussion. While women's engagement remains largely unaffected by the presence or stance of comments, men respond strongly through higher rates of both passive and active engagement. The evidence suggests that the overall amplification effects of comment sections are driven primarily by male users. This heterogeneity highlights that the dynamics of online engagement depend not only on the stance of visible comments but also on the demographic composition of the audience.

4.4.5 Robustness and Placebo Tests

We explore several alternative specifications to assess the robustness of our results. Specifically, we re-estimate all main models using alternative combinations of control variables and fixed effects, and test the sensitivity of our findings to different estimators by employing a Logit specification. In addition, we compute engagement rates directly—without simulating individual-level data—and report the corresponding estimates together with χ^2 test for equality of proportions with continuity correction. Overall, our results remain robust across all specifications.

4.5 Testing Potential Mechanisms

In ongoing work, we use an artefactual field experiment to investigate the behavioral mechanisms underlying the causal effects identified in the natural field experiment. Specifically, we will test two complementary explanations that could account for the observed increase in engagement following exposure to comment sections: *curiosity* and *reactance*.

The *curiosity mechanism* posits that exposure to visible comments—especially those expressing disagreement—heightens users' desire to explore the discussion, independent of their prior attitudes. Under this view, engagement rises because controversy attracts attention, leading users to expand posts or click through for more information.

The *reactance mechanism*, on the other hand, suggests that opposing or critical comments trigger a psychological motivation to reaffirm one's position or defend the in-group. In this case, increased engagement reflects not curiosity but an expressive response to attitudinal challenge.

To disentangle these mechanisms, we design an artefactual field experiment in which participants are randomly exposed to controlled comment sections varying in tone and stance. We then measure both their information-seeking behavior (as an indicator of curiosity) and their expressive responses (as evidence of reactance). By isolating these behavioral channels in a controlled setting, this auxiliary experiment will allow us to identify whether curiosity, reactance, or a combination of both drives how comment sections of divisive issues influence subsequent user behavior.

5 Conclusion

Our study provides causal evidence that comment sections can shape how users engage with social media content. Focusing on the topic of racial justice, we show that the nature and tone of visible comments influence not only whether users engage with a post but also how they interact with it—through reactions, comments, and link clicks. We document important variation by gender and local ideology, highlighting that vocal minorities can steer the trajectory of online discourse in ways that may not reflect the broader audience. These findings contribute to ongoing debates about the role of user-generated content in shaping public opinion on divisive issues.

Our study introduces a novel experimental pipeline that leverages platform features to manipulate comment visibility and content at scale. This approach offers a framework for studying how online discourse influences behavior in real-world settings. Future research can adapt this pipeline to investigate comment section effects in other domains – such as commercial products or public health messaging – where user perceptions and social influence play a critical role in shaping outcomes. By combining organic engagement with causal identification, this method opens new possibilities for understanding and designing comment sections across a range of topics and platforms.

References

Ali, Muhammad, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. "Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes." *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1–30.

Anderson, Monica, Michael Barthel, Andrew Perrin, and Emily A. Vogels. 2020. "#BlackLivesMatter surges on Twitter after George Floyd's death." *Pew Research Center*.

Aridor, Guy, Rafael Jiménez-Durán, Ro'ee Levy, and Lena Song. Forthcoming. "Experiments with Social Media." In *Handbook of Experimental Methods in the Social Sciences*. Edward Elgar Publishing.

Aridor, Guy, Rafael Jiménez-Durán, Ro'ee Levy, and Lena Song. 2024. "The Economics of Social Media." *Journal of Economic Literature*.

Bayerl, Andreas, Yaniv Dover, Hila Riemer, and Daniel Shapira. 2024. "Gender rating gap in online reviews." *Nature Human Behaviour*, 1–14.

Braun, Michael, and Eric M Schwartz. 2025. "Where A/B Testing Goes Wrong: How Divergent Delivery Affects What Online Experiments Cannot (and Can) Tell You About How Customers Respond to Advertising." *Journal of Marketing*, 89(2): 71–95.

Brehm, Jack W. 1966. *A Theory of Psychological Reactance*. New York:Academic Press.

Bursztyn, Leonardo, Alessandra L. González, and David Yanagizawa-Drott. 2020. "Misperceived Social Norms: Female Labor Force Participation in Saudi Arabia." *American Economic Review*, 110(10): 2997–3029.

Burtch, Gordon, Robert Moakler, Brett R Gordon, Poppy Zhang, and Shawndra Hill. 2025. "Characterizing and Minimizing Divergent Delivery in Meta Advertising Experiments." *arXiv preprint arXiv:2508.21251*.

Chen, Zoey, and Jonah Berger. 2013. "When, why, and how controversy causes conversation." *Journal of Consumer Research*, 40(3): 580–593.

Chevalier, Judith A, and Dina Mayzlin. 2006. "The effect of word of mouth on sales: Online book reviews." *Journal of marketing research*, 43(3): 345–354.

DellaVigna, Stefano, and Matthew Gentzkow. 2010. "Persuasion: Empirical Evidence." *Annual Review of Economics*, 2(1): 643–669.

Deng, Yipu, Jinyang Zheng, Warut Khern-Am-Nuai, and Karthik Kannan. 2022. "More than the quantity: The value of editorial reviews for a user-generated content platform." *Management Science*, 68(9): 6865–6888.

Donati, Dante. 2025. "The End of Tourist Traps: The Impact of Review Platforms on Quality Upgrading." *Marketing Science*.

Donati, Dante, and Nandan Rao. 2025. "Adaptive Survey Sampling via Ad Platforms." *Available at SSRN 5495148*.

Donati, Dante, Nandan Rao, Victor Hugo Orozco Olvera, and Ana Maria Munoz Boudet. 2024. "Can facebook ads prevent malaria? two field experiments in india." The World Bank.

Dubois, David, Andrea Bonezzi, and Matteo De Angelis. 2016. "Sharing with friends versus strangers: How interpersonal closeness influences word-of-mouth valence." *Journal of Marketing Research*, 53(5): 712–727.

Eckles, Dean, Brett R Gordon, and Garrett A Johnson. 2018. "Field studies of psychologically targeted ads face threats to internal validity." *Proceedings of the National Academy of Sciences*, 115(23): E5254–E5255.

Eckles, Dean, Brian Karrer, and Johan Ugander. 2017. "Design and analysis of experiments in networks: Reducing bias from interference." *Journal of Causal Inference*, 5(1): 20150021.

Eliashberg, Jehoshua, and Steven M Shugan. 1997. "Film critics: Influencers or predictors?" *Journal of marketing*, 61(2): 68–78.

Gordon, Brett R, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. 2019. "A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook." *Marketing Science*, 38(2): 193–225.

Handlan, Amy, and Haoyu Sheng. 2023. *Gender and tone in recorded economics presentations: Audio analysis with machine learning.* SSRN.

He, Qinglai, Yili Hong, and TS Raghu. 2025. "Platform governance with algorithm-based content moderation: An empirical study on Reddit." *Information Systems Research*, 36(2): 1078–1095.

Huang, Justin T, Jangwon Choi, and Yuqin Wan. 2024. "Politically biased moderation drives echo chamber formation: An analysis of user-driven content removals on Reddit." *Available at SSRN*.

Johnson, Garrett A. 2023. "Inferno: A guide to field experiments in online display advertising." *Journal of economics & management strategy*, 32(3): 469–490.

Kamenica, Emir, and Matthew Gentzkow. 2011. "Bayesian Persuasion." *Quarterly Journal of Economics*, 126(4): 1713–1768.

Kim, Sangbeom, and Seonhye Noh. 2025. "Disproportionate Voices: Participation Inequality and Hostile Engagement in News Comments." *arXiv preprint arXiv:2508.16040*.

Klapper, Joseph T. 1960. *The Effects of Mass Communication.* Glencoe, IL:Free Press.

Klinowski, David. 2023. "Voicing disagreement in science: Missing women." *Review of Economics and Statistics*, 1–40.

Lee, Dokyun, Kartik Hosanagar, and Harikesh S Nair. 2018. "Advertising content and consumer engagement on social media: Evidence from Facebook." *Management Science*, 64(11): 5105–5131.

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier. 2014. "Promotional reviews: An empirical investigation of online review manipulation." *American Economic Review*, 104(8): 2421–2455.

Moehring, Alex. 2024. "Personalized Rankings and User Engagement: An Empirical Evaluation of the Reddit News Feed."

Nistor, Cristina, and Matthew Selove. 2024. "Influencers: The power of comments." *Marketing Science*, 43(6): 1153–1167.

Pew Research Center. 2024. "Racial attitudes and the 2024 election." *Web report*, Accessed December 4, 2025.

Song, Lena. 2024. "Closing the distance: The effects of social media content on support for racial justice."

Sunstein, Cass R. 2018. "Republic: Divided democracy in the age of social media."

Xu, Yuqian, Mor Armony, and Anindya Ghose. 2021. "The interplay between online reviews and physician demand: An empirical investigation." *Management Science*, 67(12): 7344–7361.

Yang, Mochen, Yuqing Ren, and Gediminas Adomavicius. 2019. "Understanding user-generated content and customer engagement on Facebook business pages." *Information Systems Research*, 30(3): 839–855.

Yates, Frank. 1934. "Contingency tables involving small numbers and the χ 2 test." *Supplement to the Journal of the Royal Statistical Society*, 1(2): 217–235.

Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov. 2020. "Political effects of the internet and social media." *Annual Review of Economics*, 12: 415–438.

Online Appendix: Not for Publication

The Impact of Comments on Social Media Campaigns

Dante Donati and Lena Song

A Background Appendix

A.1 List of Issues and Description

- Voter Suppression: Black communities face deliberate barriers like restricted polling access, strict ID laws, and voter roll purges, aimed at limiting their voting power. Misinformation campaigns also target Black voters to reduce turnout, undermining fair representation. Breaking down these barriers is crucial to ensure Black voices are heard in democratic processes.

- Environmental Justice: Black communities often live near pollution sources like factories and highways, leading to higher rates of health issues such as asthma. These neighborhoods are frequently overlooked in clean-up efforts and lack green spaces. Environmental justice aims to provide Black communities with clean air, safe water, and healthy environments.

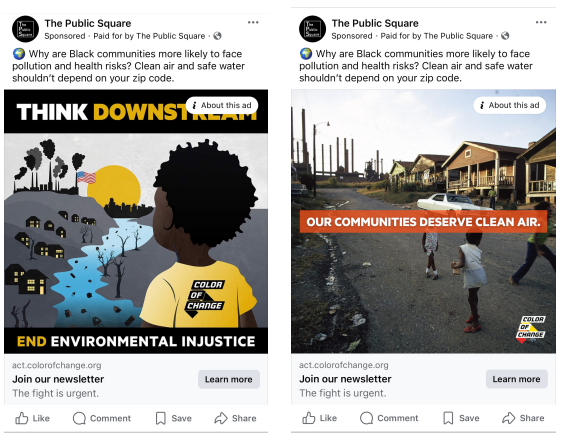
- Criminal Justice or Police Reform: Black communities experience disproportionate police violence, profiling, and harsher sentencing. This systemic bias erodes trust in law enforcement and perpetuates disadvantages. Police reform is essential for fair treatment, accountability, and ensuring Black communities feel protected, not targeted, by the justice system.

- Education Reform: Black students often attend underfunded schools with fewer resources, larger classes, and limited access to advanced courses. These disparities create achievement gaps and limit future opportunities. Education reform seeks equitable funding and support to provide Black students with the quality education they deserve.

- Technology fairness: Black communities face systemic biases in technology, from algorithmic discrimination in hiring and lending to facial recognition tools that disproportionately misidentify Black individuals. These inequities perpetuate existing racial disparities and limit opportunities. Ensuring technology fairness involves designing inclusive systems, addressing bias in algorithms, and creating tools that serve all communities equitably

Figure A1: Ad Banners and Headlines

(a) Education



(b) Environment

(c) Voting

(d) Technology

(e) Police

A.2 Pre-tests

Table A1: Results from Pre-test A

Issue	Ad Name	Link Clicks	Reach	CTR (%)
technology	pixels	15	6115	0.245
technology	man	13	6683	0.195
voting	lady	14	5777	0.242
voting	flag	11	5792	0.190
police	lady	13	6146	0.212
police	hands	13	6315	0.206
environment	kid	13	6235	0.209
environment	street	11	6290	0.175
education	future	9	6216	0.145
education	history	6	4884	0.123

Table A2: Results from Pre-tests B and C

Issue	Ad Name	Link Clicks	Reach	CTR (%)
voting	lady	32	26794	0.119
environment	kid	33	28351	0.116
police	hands	32	27771	0.115
technology	pixels	29	28208	0.103

B Additional Results for Causal Impact of the Comment Section

Table A3: The Impact of Any Comment on Subsequent Engagement (as % of total reach)

	Dependent variable:				
	All Engagement	Post Expansions	Interactions	Link Clicks	Page Views
	(1)	(2)	(3)	(4)	(5)
Any comments	0.065***	0.046***	0.003	0.016**	0.016*
	(0.015)	(0.012)	(0.003)	(0.008)	(0.008)
Constant	0.538***	0.156**	0.004	0.379***	0.347**
	(0.120)	(0.064)	(0.010)	(0.070)	(0.135)
Zipcode Fixed Effects	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓
Zipcode FEs × Controls	✓	✓	✓	✓	✓
Mean Y in Control (C)	0.485	0.237	0.020	0.228	0.171
Observations	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015
R^2	0.0004	0.0005	0.0003	0.0002	0.0002

Notes: Standard errors are clustered at the advertisement level (72 ads). The unit of observation is defined at the level of the reach. Use gender and age as control, along with two-way interactions. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A4: The Impact of the Comment Section on Subsequent Engagement (as % of total reach)

	Dependent variable:				
	All Engagement	Post Expansions	Interactions	Link Clicks	Page Views
	(1)	(2)	(3)	(4)	(5)
Opposing	0.087***	0.044***	0.009**	0.034***	0.019*
	(0.021)	(0.016)	(0.004)	(0.011)	(0.011)
Mixed	0.062***	0.047***	0.002	0.012	0.017
	(0.017)	(0.012)	(0.003)	(0.011)	(0.011)
Supportive	0.047***	0.048***	−0.003	0.003	0.011
	(0.018)	(0.013)	(0.003)	(0.008)	(0.011)
Constant	0.539***	0.156**	0.004	0.380***	0.347***
	(0.115)	(0.064)	(0.010)	(0.066)	(0.134)
Zipcode Fixed Effects	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓
Zipcode FEs × Controls	✓	✓	✓	✓	✓
Mean Y in Control (C)	0.485	0.237	0.020	0.228	0.171
p(Support vs. Oppose)	0.059	0.802	0.002	0.008	0.523
p(Oppose vs. Mixed)	0.216	0.833	0.091	0.106	0.880
p(Support vs. Mixed)	0.394	0.942	0.078	0.432	0.611
Observations	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015
R^2	0.0004	0.0005	0.0003	0.0002	0.0002

Notes: Standard errors are clustered at the advertisement level (72 ads). The unit of observation is defined at the level of the reach. Use gender and age as control, along with two-way interactions. Interactions include comments, reactions and shares.
$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A5: The Impact of the Comment Section on Subsequent Engagement (as % of total reach)

	Dependent variable:				
	All Engagement	Post Expansions	Interactions	Link Clicks	Page Views
	(1)	(2)	(3)	(4)	(5)
Opposing	0.086***	0.044***	0.008**	0.033***	0.019*
	(0.021)	(0.016)	(0.004)	(0.011)	(0.011)
Mixed	0.061***	0.047***	0.002	0.011	0.017
	(0.017)	(0.012)	(0.003)	(0.011)	(0.011)
Supportive	0.047***	0.048***	−0.003	0.002	0.011
	(0.017)	(0.013)	(0.003)	(0.008)	(0.011)
Constant	0.516***	0.194***	0.028***	0.295***	0.274***
	(0.048)	(0.038)	(0.008)	(0.028)	(0.025)
Zipcode Fixed Effects	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓
Mean Y in Control (C)	0.485	0.237	0.020	0.228	0.171
p(Support vs. Oppose)	0.063	0.776	0.003	0.007	0.515
p(Oppose vs. Mixed)	0.225	0.822	0.096	0.105	0.876
p(Support vs. Mixed)	0.399	0.920	0.084	0.422	0.608
Observations	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015
R^2	0.0003	0.0003	0.0001	0.0001	0.0001

Notes: Standard errors are clustered at the advertisement level (72 ads). The unit of observation is defined at the level of the reach. Use gender and age as control. Interactions include comments, reactions and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A6: The Impact of the Comment Section on the Valence of Subsequent Interactions (in %)

	Dependent variable: Interactions by valence		
	All	Supportive	Others/Non-supportive
	(1)	(2)	(3)
Opposing	0.0082**	0.0075**	0.0007
	(0.0042)	(0.0036)	(0.0021)
Mixed	0.0038	−0.0008	0.0046*
	(0.0040)	(0.0031)	(0.0025)
Supportive	−0.0037	−0.0018	−0.0019
	(0.0036)	(0.0031)	(0.0019)
Constant	0.0210***	0.0143***	0.0067***
	(0.0034)	(0.0027)	(0.0019)
Zipcode Set FEs	✓	✓	✓
Mean Y in Control (C)	0.0190	0.0133	0.0057
p(Support vs. Oppose)	0.003	0.008	0.186
p(Oppose vs. Mixed)	0.310	0.020	0.128
p(Support vs. Mixed)	0.048	0.722	0.005
Observations	1,054,015	1,054,015	1,054,015
R^2	0.00002	0.00001	0.00001

Notes: Heteroskedasticity-robust standard errors (HC1).
The unit of observation is defined at the level of the reach.
Interactions include comments, reactions, and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A7: The Impact of the Comment Section in Blue Areas

	Dependent variable (in %):				
	All Engagement	View Comments	Interactions	Link Clicks	Page Views
	(1)	(2)	(3)	(4)	(5)
Opposing	0.053*	0.027	0.010	0.017	0.012
	(0.030)	(0.026)	(0.007)	(0.021)	(0.021)
Mixed	0.053*	0.027	0.002	0.022	0.008
	(0.032)	(0.023)	(0.006)	(0.024)	(0.020)
Supportive	−0.014	0.001	−0.003	−0.012	−0.001
	(0.031)	(0.025)	(0.005)	(0.016)	(0.019)
Constant	0.561***	0.152**	−0.007	0.416***	0.361***
	(0.117)	(0.067)	(0.012)	(0.072)	(0.138)
Zipcode Fixed Effects	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓
Mean Y in Control (C)	0.529	0.243	0.020	0.267	0.209
p(Support vs. Oppose)	0.052	0.274	0.029	0.195	0.590
p(Oppose vs. Mixed)	0.996	0.990	0.270	0.862	0.885
p(Support vs. Mixed)	0.065	0.200	0.363	0.184	0.689
Observations	329,844	329,844	329,844	329,844	329,844
R^2	0.0004	0.0004	0.0003	0.0002	0.0002

Notes: Standard errors are clustered at the advertisement level (72 ads). The unit of observation is defined at the level of the reach. Use gender and age as control. Interactions include comments, reactions and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A8: The Impact of the Comment Section in Swing Areas

	Dependent variable (in %):				
	All Engagement	View Comments	Interactions	Link Clicks	Page Views
	(1)	(2)	(3)	(4)	(5)
Opposing	0.071*	0.031	0.001	0.037**	0.004
	(0.037)	(0.023)	(0.009)	(0.015)	(0.018)
Mixed	0.052***	0.033**	0.004	0.017	0.044***
	(0.019)	(0.014)	(0.005)	(0.014)	(0.017)
Supportive	0.070***	0.057***	−0.004	0.020	0.017
	(0.022)	(0.014)	(0.006)	(0.013)	(0.013)
Constant	0.317***	0.187***	0.021	0.109**	0.136
	(0.068)	(0.052)	(0.017)	(0.053)	(0.137)
Zipcode Fixed Effects	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓
Mean Y in Control (C)	0.478	0.240	0.025	0.209	0.151
p(Support vs. Oppose)	0.996	0.282	0.556	0.282	0.341
p(Oppose vs. Mixed)	0.604	0.932	0.789	0.237	0.023
p(Support vs. Mixed)	0.382	0.124	0.219	0.844	0.029
Observations	345,109	345,109	345,109	345,109	345,109
R^2	0.0005	0.001	0.0003	0.0002	0.0002

Notes: Standard errors are clustered at the advertisement level (72 ads). The unit of observation is defined at the level of the reach. Use gender and age as control. Interactions include comments, reactions and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A9: The Impact of the Comment Section in Red Areas

	Dependent variable (in %):				
	All Engagement	View Comments	Interactions	Link Clicks	Page Views
	(1)	(2)	(3)	(4)	(5)
Opposing	0.132***	0.071**	0.015***	0.045**	0.039**
	(0.041)	(0.031)	(0.005)	(0.021)	(0.018)
Mixed	0.077**	0.077***	0.001	−0.003	0.001
	(0.031)	(0.023)	(0.004)	(0.016)	(0.016)
Supportive	0.080**	0.081***	−0.002	−0.0001	0.016
	(0.031)	(0.025)	(0.004)	(0.014)	(0.021)
Constant	0.482***	0.288**	−0.015**	0.212***	0.175***
	(0.145)	(0.115)	(0.007)	(0.041)	(0.055)
Zipcode Fixed Effects	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓
Mean Y in Control (C)	0.455	0.228	0.016	0.210	0.157
p(Support vs. Oppose)	0.145	0.712	0.000	0.027	0.336
p(Oppose vs. Mixed)	0.127	0.807	0.001	0.027	0.056
p(Support vs. Mixed)	0.919	0.837	0.370	0.849	0.505
Observations	379,062	379,062	379,062	379,062	379,062
R^2	0.0005	0.001	0.0002	0.0002	0.0002

Notes: Standard errors are clustered at the advertisement level (72 ads). The unit of observation is defined at the level of the reach. Use gender and age as control. Interactions include comments, reactions and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A10: The Impact of the Comment Section on Female Engagement

	Dependent variable:				
	All Engagement	View Comments	Interactions	Link Clicks	Page Views
	(1)	(2)	(3)	(4)	(5)
Opposing	0.044	0.026	0.006	0.011	0.012
	(0.028)	(0.018)	(0.005)	(0.016)	(0.019)
Mixed	0.006	0.027*	−0.007	−0.015	0.0001
	(0.025)	(0.014)	(0.004)	(0.017)	(0.019)
Supportive	−0.038	0.001	−0.009**	−0.032**	−0.014
	(0.025)	(0.016)	(0.004)	(0.015)	(0.018)
Constant	0.560***	0.112	0.003	0.447***	0.375**
	(0.116)	(0.106)	(0.004)	(0.110)	(0.154)
Zipcode Fixed Effects					
Controls					
Mean Y in Control (C)	0.490	0.234	0.019	0.237	0.182
p(Support vs. Oppose)	0.004	0.133	0.002	0.009	0.156
p(Oppose vs. Mixed)	0.182	0.912	0.008	0.166	0.553
p(Support vs. Mixed)	0.080	0.032	0.617	0.354	0.461
Observations	502,473	502,473	502,473	502,473	502,473
R^2	0.001	0.001	0.0005	0.0003	0.0004

Notes: Standard errors are clustered at the advertisement level (72 ads). The unit of observation is defined at the level of the reach. Use age as control. Interactions include comments, reactions and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A11: The Impact of the Comment Section on Male Engagement

	Dependent variable:				
	All Engagement	View Comments	Interactions	Link Clicks	Page Views
	(1)	(2)	(3)	(4)	(5)
Opposing	0.126***	0.061***	0.012*	0.054***	0.026*
	(0.027)	(0.020)	(0.006)	(0.016)	(0.013)
Mixed	0.112***	0.065***	0.011**	0.035***	0.033***
	(0.022)	(0.018)	(0.004)	(0.013)	(0.010)
Supportive	0.125***	0.091***	0.002	0.033**	0.034**
	(0.026)	(0.019)	(0.005)	(0.013)	(0.013)
Constant	0.692***	0.301***	−0.006*	0.397**	0.405***
	(0.124)	(0.059)	(0.004)	(0.154)	(0.068)
Zipcode Fixed Effects Controls					
Mean Y in Control (C)	0.481	0.239	0.021	0.219	0.161
p(Support vs. Oppose)	0.954	0.132	0.173	0.173	0.613
p(Oppose vs. Mixed)	0.521	0.822	0.898	0.198	0.602
p(Support vs. Mixed)	0.556	0.158	0.119	0.875	0.939
Observations	551,542	551,542	551,542	551,542	551,542
R^2	0.0005	0.001	0.0003	0.0003	0.0003

Notes: Standard errors are clustered at the advertisement level (72 ads). The unit of observation is defined at the level of the reach. Use age as control. Interactions include comments, reactions and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.