Social Media Comments

Dante Donati and Lena Song*

March 2026

Preliminary draft. The latest version is available here.

Abstract

Comment sections are a central feature of social media platforms, enabling users to engage with content and one another. This paper studies how the presence and stance of comments on social media posts affect subsequent users' on-platform engagement and off-platform attitudes and behavior. For identification, we develop a novel experimental pipeline that manipulates and randomizes comment visibility and stance using Meta's built-in A/B testing infrastructure. In collaboration with a leading racial justice organization, we conduct a large-scale field experiment on Facebook reaching a million U.S. users randomly assigned to one of four conditions: (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 with the post. Importantly, comment stance matters: opposing comments significantly amplify reactions, comments, and link clicks relative to the control, whereas supportive comments have little effect. In a complementary artefactual field experiment, we find that comments increase the time spent viewing the post, and exposure to opposing comments leads to more negative perceptions of the organization, less progressive attitudes, and reduced donations. Our results reveal a fundamental tradeoff between on-platform engagement and off-platform attitudes and behavior, underscoring the role of comment ranking and moderation policies for businesses, users, platforms, and regulators.

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 Luca Braghieri, Sarah Eichmeyer, Rafael Jiménez-Durán, Maria Petrova, Andrea Prat, Carlo Schwarz, Andrey Simonov, and Yanwen Wang, and seminar and conference participants at Haas School of Business, Bass FORMS Conference, MIT Conference on Digital Experimentation, 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, Daniel Merlau, and Navtej Singh for excellent research assistance. The research was approved by the Institutional Review Boards at Columbia University (AAAU8166). This study was registered in the American Economic Association Registry for randomized control trials under trial numbers AEARCTR-0013812 and AEARCTR-0017850. 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. The authors declare no conflict of interest. All errors are the authors' own.

1 Introduction

Individuals are often exposed to the opinions of others. Online communication technologies have made such exposure ubiquitous, from discussion sections in blogs and online newspapers to review platforms. Social media platforms are a particularly important case because they have become central to how people communicate, share and consume information, and form opinions (Zhuravskaya, Petrova and Enikolopov, 2020; Aridor et al., 2024). At the same time, there are widespread concerns that these platforms contribute to increasing polarization, spreading misinformation, and amplifying xenophobia, hate speech, and social conflict.[1]

A growing literature studies how platforms shape outcomes through algorithmic curation and content exposure (Levy, 2021; Aridor et al., 2024; Gauthier et al., 2026; Guess et al., 2023*a*). But another distinctive feature of social media, and of modern communication technologies more broadly, is that they also enable *horizontal communication*, allowing users to influence one another through decentralized, peer-to-peer interactions that are often visible to large audiences. Yet much less is known about how this form of communication shapes beliefs and behavior.

Social media comment sections provide a natural setting to study these interactions. Comments are publicly visible and widely read: in our survey of U.S. adults, 85% of respondents report sometimes, often, or very often reading comments on social media, and more than 50% report doing so often or very often. At the same time, comments are often generated by a small and potentially unrepresentative subset of users (Kim and Noh, 2025). As a result, comment sections may distort perceptions of public opinion and affect beliefs and behavior through social learning and persuasion (Banerjee, 1992; DellaVigna and Gentzkow, 2010; Bursztyn and Jensen, 2017). Because platform algorithms partly reward engagement, the effects of comments may extend beyond the users who read them directly by affecting the broader visibility of posts. Understanding the causal effects of comments, and of the views they express, is therefore central to evaluating how social media platforms shape discourse and how they should be governed.

This paper studies how users interact in social media comment sections and how those interactions shape the subsequent behavior of other users, both on and off the platform. In collaboration with Color of Change, the largest online racial justice organization in the United States, we ran large-scale field experiments in which approximately one million Facebook users were randomly assigned to see posts about racial justice that differed in the stance of the comment section. To examine downstream attitudinal and behavioral effects, we complement these organic experiments with an artefactual field experiment on Prolific participants. Racial justice provides a useful setting because it is a domain of intense public debate and highly polarized views (Pew Research Center, 2024), making it possible to study cross-cutting exposure in a context where the stakes are socially important. We show that comments causally affect user engagement, attitudes, and

[1] `https://www.wsj.com/articles/the-facebook-files-11631713039`

behavior. In particular, comments increase attention to posts, and opposing comments further increase on-platform engagement while shifting attitudes in a less progressive direction and reducing donations to a progressive cause. Taken together, the results suggest that the presence and stance of comment sections create a tension between online engagement and broader offline social objectives, a trade-off that engagement-based ranking algorithms may further amplify (Germano, Gómez and Sobbrio, 2026).

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.[2] 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 135,000 Facebook users across 4,095 ZIP codes using 150 sponsored posts. 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%). Consistent with the literature on gender differences in vocal engagement (Klinowski, 2023; Karpowitz and Mendelberg, 2014), 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 always 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 (Bail, 2021; Kim and Noh, 2025; Oswald et al., 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 conducted on Facebook. 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

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

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 actions to expand the comment section, user interactions with the post (reactions, comments, and shares), 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 several weeks 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 increases 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 increases the likelihood of users clicking to expand 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 has pronounced effects on downstream on-platform engagement: ads with opposing comments generate significantly more interactions and link clicks than those with supportive comments ($p < 0.01$). Relative to the control, opposing stances increase comments, reactions, and shares by roughly 45% ($p < 0.05$) and raise click-through rates by 0.034 percentage points (a 15% increase, $p < 0.01$). As a result, advertising costs such as cost-per-click and cost-per-interaction decline proportionately.

These average effects mask substantial heterogeneity across genders and ideological compositions. The impact is much larger for men than for women, and stronger in conservative than in progressive areas. In conservative areas, opposing comments substantially increase interactions

and click-through rates, whereas effects are small and statistically insignificant in progressive areas. Similarly, male users exhibit large and precisely estimated responses to both comment visibility and stance, while female engagement appears largely insensitive to comments. These patterns are consistent with two complementary mechanisms. First, opposing comments may heighten attention by signaling visible disagreement, thereby drawing users into the discussion. Second, the effect of comment stance may depend on political identity: comments that are counter-attitudinal in a given environment may simultaneously increase salience and resonate with identity-congruent users. Together, these mechanisms suggest that the influence of visible discourse depends critically on both the ideological context and the composition of the audience.

Our Facebook 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.

At the same time, on-platform engagement metrics capture only part of the picture. To assess how comment sections shape beliefs, attitudes, and off-platform high-stakes behavior, we conducted a complementary artefactual field experiment with around 5,000 U.S. participants on Prolific, mirroring the design, creatives, and organic comments used in the Facebook experiment. We find that both supportive and opposing comments increase the time spent viewing the post by almost 30 seconds on average (a 45% increase relative to control). However, exposure to opposing comments shifts attitudes in a less progressive direction, reduces incentivized donations by 8%, and elicits stronger negative emotional responses among identity-incongruent users. Thus, although opposing comments amplify engagement on the platform, they may weaken support for the organization and its message off the platform. Taken together, the evidence highlights a tension between short-run engagement gains and downstream attitudinal and behavioral consequences.

The results have several implications for policy and practice. Regulations such as 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.[3] Our findings show that the visible comments of a vocal minority can shape the behavior of a much larger audience, underscoring the importance of how comment sections are structured and moderated.

At the same time, we document a strategic tension in comment moderation. Prioritizing negative or opposing comments may boost on-platform engagement and reduce advertising costs, but

[3]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`.

for content producers such comments may be polarizing and pose brand-safety risks. For advertisers and campaign managers, moderation policies must therefore be calibrated to organizational objectives. Tolerating opposing comments can increase engagement, while stricter moderation may better protect brand image and downstream support. Importantly, platform incentives to maximize engagement may not align with the incentives of firms, nonprofits, or political actors seeking to preserve brand integrity or policy support. Without deliberate moderation strategies, comment sections can amplify polarized voices and distort the apparent balance of opinion. Balancing visibility, engagement, and brand safety therefore requires explicit recognition of these trade-offs.

Our paper builds on several strands of literature. First, it contributes to the growing literature on the economics of social media (Zhuravskaya, Petrova and Enikolopov, 2020; Aridor et al., 2024). A recent wave of large-scale field experiments has studied how platform design choices shape user behavior and attitudes, focusing on algorithmic curation of the news feed (Levy, 2021; Nyhan et al., 2023; Guess et al., 2023*a*,*b*; Gauthier et al., 2026). These studies have advanced our understanding of the role of algorithms, but far less is known about how the social context surrounding a post – and in particular its comment section – shapes user responses. Existing work has studied the determinants of participation in online public discourse (Oswald, Schulz and Lorenz-Spreen, 2025), and used comments primarily as a data source to study a range of questions (e.g., Yang, Ren and Adomavicius 2019; Moehring 2024), typically treating them as measures of engagement rather than studying their causal impacts. We address this gap by developing a novel experimental pipeline that leverages Meta's A/B testing infrastructure to manipulate the comment visibility and stance, providing causal evidence that pre-existing comments and their stance influence user behavior both on and off the platform.

Second, our paper relates to the literature on persuasion and exposure to opposing viewpoints (see DellaVigna and Gentzkow 2010 for a review). In contrast to settings where the persuasive content is produced by media outlets or researchers, the persuasive stimulus in our setting is user-generated: comments from ordinary individuals expressing support or opposition. Comments may therefore operate not only through their informational content but also through the social signals they convey about the prevalence and acceptability of different views, whether by triggering herding and informational cascades (Banerjee, 1992; Bikhchandani, Hirshleifer and Welch, 1992; Muchnik, Aral and Taylor, 2013), or by shifting perceived social norms (Bursztyn, González and Yanagizawa-Drott, 2020). To the extent that individuals view comments as reflecting the organic opinions of peers rather than the strategic messaging of institutions, user-generated persuasion may be particularly effective in shaping attitudes and behavior.

Third, our findings speak to an emerging literature on the divergence between engagement and welfare on digital platforms. Beknazar-Yuzbashev, Jiménez-Durán and Stalinski (2024) show theoretically that platforms may have incentives to promote harmful yet engaging content, and Beknazar-Yuzbashev et al. (2025) provide supporting experimental evidence. Relatedly, Germano,

Gómez and Sobbrio (2026) show that algorithms placing greater weight on social interaction signals can increase engagement while also increasing misinformation and polarization. Our paper identifies a distinct engagement–welfare tradeoff operating through user-generated content, and more specifically, the stance composition of comment sections. We show that opposing comments—the condition that most increases on-platform engagement in terms of reactions, comments, and link clicks—simultaneously generate more negative attitudes toward the organization and reduce charitable donations off-platform. Moreover, because this tradeoff depends on the alignment between comment stance and the viewer's own ideology, the same comment section can engage and persuade different users in different directions.

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 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). Social media comments represent a distinct and understudied form of UGC: unlike product reviews, they are not evaluations of a good or service but rather spontaneous expressions of opinion that appear alongside the original post. 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 both on-platform engagement and off-platform attitudes and behavior.

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 Background and Setting

2.1 Comment Sections and Racial Justice

The comment section is a common feature of almost all major social media platforms. On Facebook and YouTube, the most widely used platforms in the United States, comment sections appear 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.

We study comment sections in the context of 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 manifest in online discourse and how they influence subsequent users' on-platform engagement and off-platform attitudes and behavior. 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.

2.2 Social Media Advertising as a Research Tool

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.[4] 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.

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

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.

2.3 Post Design and Pre-testing

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.

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.[5] 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.

Prior to launching the campaign at full scale, we conducted a series of pre-tests to select the creatives used in the study and to assess their performance under alternative delivery configurations; details are provided in Appendix A.

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

2.4 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.[6] 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.[7] 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.[8]

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.

2.5 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.

[6] https://www.facebook.com/business/help/166533308016738O?id=176276233019487

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

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

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

Between January 13 and February 10, 2025, we ran a Facebook ad campaign to generate organic engagement. The campaign featured 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.[9] 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., conservative 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

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

individuals and generating 12,000 reactions, 1,750 unique link clicks, 1,500 direct comments,[10] 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.[11] 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.

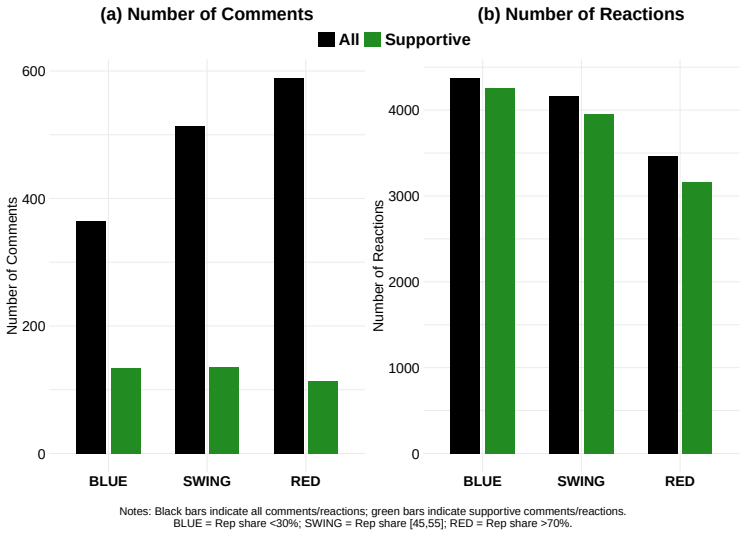
3.2 Engagement Rates

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. We find a similar pattern for the rate of commenting in Appendix Figure B1. Comment rate rises 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 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.

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

[11] Other reaction types, such as *laugh*, *wow*, *sad*, and *angry*, are context-dependent and therefore harder to interpret;

Figure 1: Comment and Reaction Counts by Location

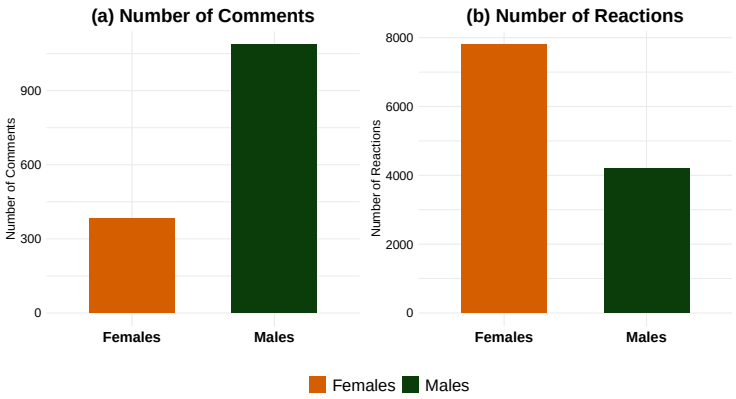


When focusing on *supportive engagement*—defined as reactions or comments expressing agreement or approval—the ideological gradient is notably flatter for comments. Supportive comments remain consistently low across areas, ranging around 0.3 percent of reach, with no statistically significant differences 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 counts by gender, and Appendix Figure B2 reports the corresponding 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 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

we focus on those that most reliably convey positive engagement.

Figure 2: Comment and Reaction Counts by Gender



Notes: Bars indicate the total number of comments/reactions.
Females: 75,451, Males: 59,329

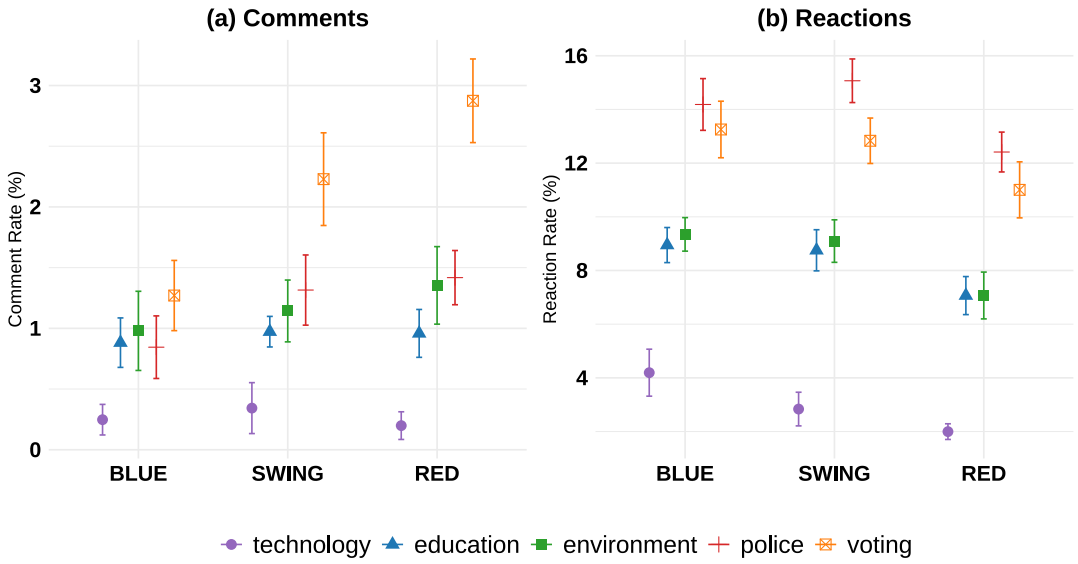
suggest that men are more likely to express criticism through comments, whereas women are more likely to signal support through reactions. This pattern is consistent with well-documented gender differences in opinion expression, where men are disproportionately represented among those who criticize scientific 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.

3.3 Composition of the Comment Section

Figure B3 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), $p < 0.01$).

Figure 3: Comment and Reaction Rates by Location 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

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.

4 The Impact of Comments on On-platform Engagement

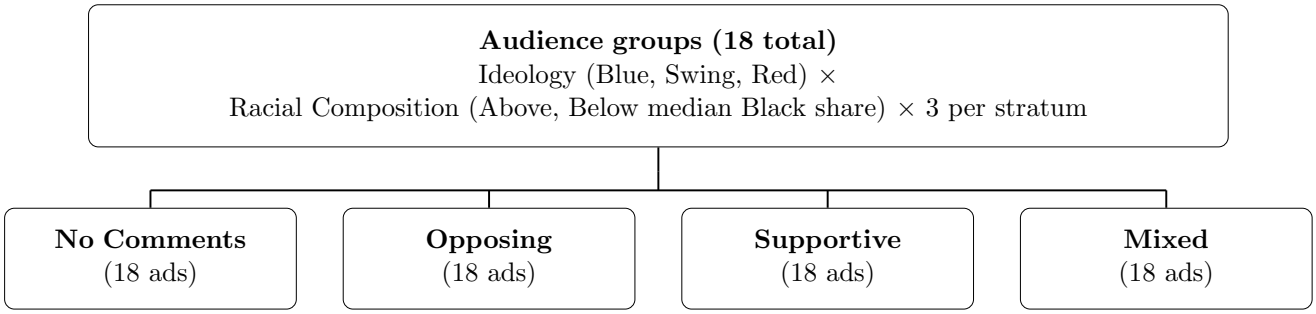
In this section, we provide novel causal evidence on the effect of the comment section. Specifically, we study how the presence and stance of a comment section influences subsequent users' engagement with the content.

4.1 Design

The experiment ran from March 26 to April 13, 2025, and reached over one million Facebook users. Figure 4 provides an overview of the experimental 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.

Figure 4: Design Overview



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 comment generation campaign, and create 18 audience groups, with three groups for each combination of political ideology (Blue, Swing, Red) and racial composition (above or below the median share of Black residents within each ideological category).

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 users see under each condition.

On Facebook, comments on ads are hidden by default. Users must actively click the comment counter or comment button, or expand the ad container, to view the comment section. Importantly, Facebook algorithmically ranks comments based on engagement and other signals.[12] To mitigate potential ranking effects, we display exactly two comments in each comment condition, thereby minimizing within-condition variation in comment visibility. In addition, at the start of the campaign, the share counter is equalized across conditions (12 shares). The comment counter is absent in the control condition and equalized across the comment arms that are part of the same A/B test, displaying 20, 30, or 40 comments. The reaction counter is similarly balanced across posts within the same A/B test, set at approximately 60, 70, or 80. Figure 5 shows an example

[12] See `https://about.fb.com/news/2019/06/making-public-comments-more-meaningful/`.

Figure 5: Experimental Arms



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

of an A/B test in which the comment counter is set to 30 in the treatment arms and the reaction counter is approximately 80.

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. [13] 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).[14] This procedure ensured that the comment counter in the No Comment condition was substantially lower than counter displayed in the treatment conditions.

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.[15] 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

[13] 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.

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

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

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. 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,[16] 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.

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

4.2.1 Outcomes

We analyze a set of engagement outcomes that capture on-platform activity. Specifically, we focus on the following metrics:

- **All engagement**, defined as any unique user activity related to the ad, including clicks on the ad container, link clicks, profile clicks, reactions, comments, shares, saves, and other observable actions.
- **Post expansions**, our proxy for attention to the comment section, defined as whether a user expanded the ad panel to view the comments. This outcome is not directly reported in the Meta Ads Manager interface and is constructed from other available metrics.[17]
- **Interactions**, defined as the sum of unique reactions (likes and other emoji responses), comments, and shares (reposting);
- **Unique link clicks**, capturing whether a user clicked on the external link, which we interpret as intent to learn more about the campaign.[18]

Unless otherwise specified, we report the engagement rates computed as the ratio of each outcome to total reach, which is defined as the number of distinct users who were shown the ad at least once.

In addition to these engagement 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.

[17] *Post Expansions = Clicks All - Page Engagement. Clicks All* captures all user clicks on an ad, including clicks on the ad container, link clicks, profile clicks, reactions, comments, shares, saves, and other interactions. *Page Engagement* includes all identifiable interactions with the ad or the advertiser's page (link clicks, profile clicks, reactions, comments, shares, saves, and other interactions). The residual therefore isolates clicks on the ad container that expand the post without generating any other observable user activity.

[18] As a robustness check, Section 4.4.5 reports results using landing-page views, which capture downstream off-platform activity. This measure has two limitations. First, landing-page views are inferred via the Facebook Pixel rather than directly recorded by Facebook, which may introduce measurement error. Second, unlike unique link clicks, landing-page views are not unique at the user level.

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).

Table 1: Summary statistics

Variable name	Mean (%)	St. Dev.
Treatment assignment (obs.)		
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	0.290	5.380
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.

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

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 ad level (72 ads).

reached a demographically diverse sample representative of the platform's active user population.

4.2.3 Balance Checks

Table 5 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 experimental 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

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. The p-values are based on t-tests using heteroskedasticity-robust standard errors.

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 14,518 users per ad on average, is consistent with complete audience saturation and balanced delivery across conditions. Appendix Table C1 reports saturation numbers under different scenarios). Under the most conservative estimate, we reached over 95% of the potential audience in the areas we targeted.

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 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 action Y. We model the individual-level outcome Y_{iz}^k as a realization of a Bernoulli random variable:

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

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.

Most 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 reach-weighted mean outcomes Y across ads assigned to the four experimental conditions, without relying on simulated individual-level observations. In this case, statistical significance is assessed using heteroskedasticity-robust t-tests of equality across groups. This approach provides a transparent benchmark for treatment effects based solely on differences in means, without imposing distributional assumptions or functional-form restrictions.

4.4 Results

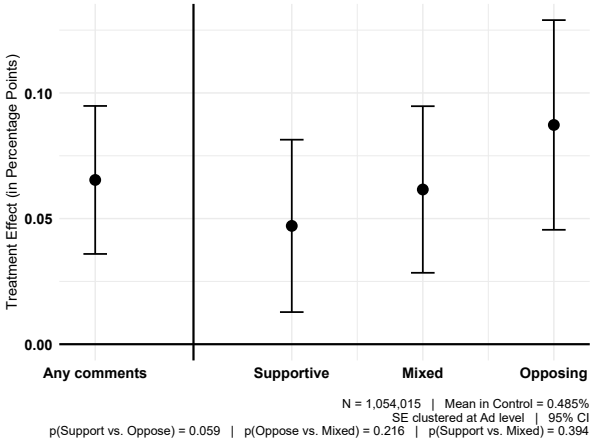
4.4.1 Main Results

We present estimates of the causal impact of the comment section on on-platform user engagement, based on the linear probability model described above, with standard errors clustered at the advertisement level. Figure 6(a) reports the effects on overall engagement, while Figures 6(b–d) present the effects on post expansions, interactions, and link clicks. Effects are reported in percentage points (pp). Full regression results are provided in Appendix Table C2.

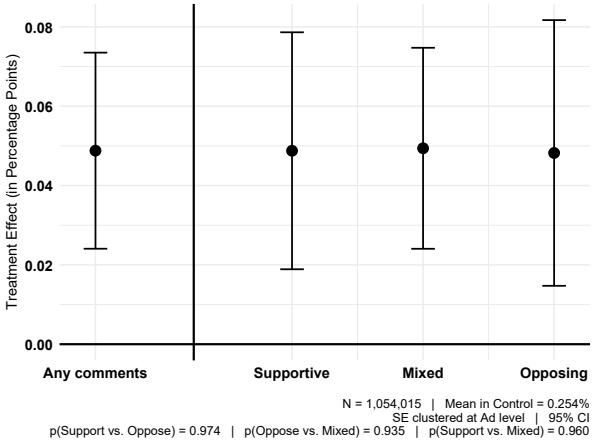
All Engagement Figure 6(a) shows that displaying any comments increases total engagement by 0.065 pp ($p < 0.01$), corresponding to a 13.4 percent increase relative to the control mean of 0.485 percent. By stance, Opposing comments generate the largest increase (0.087 pp, $p < 0.01$; 17.9 percent), followed 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

Figure 6: The Impact of the Comment Section on On-platform User Engagement
Outcomes are expressed in % of total reach

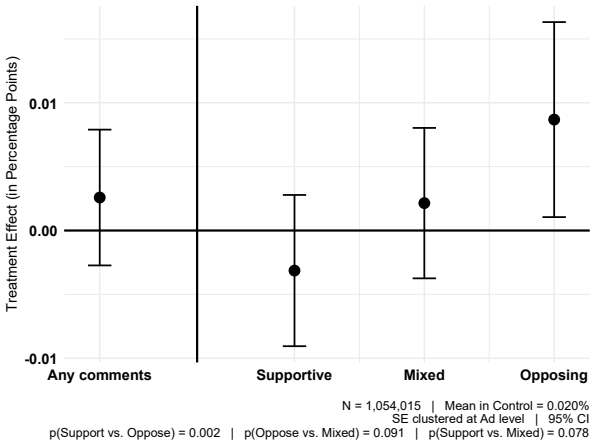
(a) All Engagement



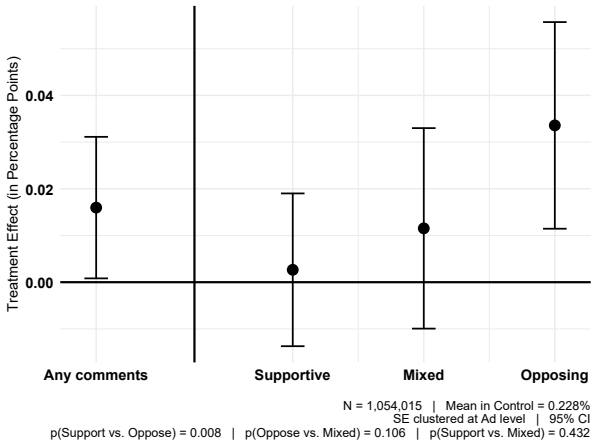
(b) Post Expansions



(c) Interactions (reactions, comments, shares)



(d) Link Clicks



tends to amplify overall user activity around the content. We next examine which specific actions drive this pattern.

Post Expansions Figure 6(b) shows that any comments increase the probability that users expand the ad panel to view the comment section by 0.049 pp ($p < 0.01$), corresponding to a 19.3 percent increase over the baseline mean of 0.254 percent. Interpreting post expansions as a proxy for attention, these findings indicate that the presence of comments per se increases attention, independent of their ideological orientation. All three stance conditions produce virtually identical effects—Opposing (0.048 pp, $p < 0.01$), Mixed (0.049 pp, $p < 0.01$), and Supportive (0.049 pp, $p < 0.01$)—and none of the pairwise differences are significant ($p > 0.90$). 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. As such, this evidence provides an additional sanity check on the successful randomization for our study, suggesting that differential ad delivery across experimental conditions is unlikely to explain our results.

Interactions Figure 6(c) 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. For Mixed comments, the coefficient is positive but statistically insignificant (0.002 pp), whereas for Supportive comments it is negative and insignificant (-0.003 pp). Differences between Opposing and Supportive conditions are highly significant ($p < 0.01$), whereas differences involving the Mixed condition are smaller and only marginally significant ($p < 0.10$). This pattern suggests that antagonistic or contentious comments elicit greater visible participation from subsequent users.

Link Clicks Figure 6(d) 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 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 the coefficients for Mixed (0.012 pp) and Supportive (0.003 pp) comments are smaller and not significant. The difference between Opposing and Supportive arms is statistically significant ($p < 0.01$), whereas comparisons involving the Mixed condition are not. These results indicate that opposing comments are particularly effective in increasing click-through activity, a key metric for organizations seeking to drive website traffic.

Discussion Across all outcomes, the presence of a comment section modestly increases on-platform user engagement. However, much of this overall effect reflects greater attention to the post itself, as captured by post expansions, and does not vary by the stance of existing comments. By contrast, comment stance shapes the composition of subsequent engagement. Opposing comments consistently generate higher rates of interactions and link clicks relative to the no-comments condition, whereas Supportive comments do not significantly outperform the control. Differences between the Opposing and Supportive conditions are substantial, while differences across the remaining conditions are generally small. Taken together, these findings suggest that opposing or contentious discourse amplifies participation and interest more than supportive commentary, pointing to a potential trade-off between engagement amplification and polarization in online comment sections (in line with Germano, Gómez and Sobbrio, 2026). In Section 5, we further examine how these engagement patterns affect attitudes and off-platform behavior in an artefactual field

experiment.

Ad Costs These engagement effects also have meaningful implications for the unit economics of digital advertising. Table 4 reports the impact of comment stance on key cost metrics. Opposing comments are associated with a \$0.30 reduction in spend per engagement (15 percent relative to control), statistically significant at conventional levels ($p < 0.05$). Mixed and Supportive comments also reduce spend per engagement ($p < 0.05$), though by smaller magnitudes. Turning to specific types of engagement, spend per interaction falls by \$32.5 (a 50 percent reduction) under Opposing comments relative to the control ($p < 0.01$), whereas Supportive and Mixed comments do not yield statistically significant differences. A similar pattern emerges for spend per link click: costs are \$0.69 lower (15 percent) under Opposing comments ($p < 0.10$), while the remaining conditions show no significant effects. Taken together, these findings indicate that comment sections featuring opposing stances not only increase engagement but also improve cost-effectiveness, particularly for higher-value actions such as interactions and link clicks, reinforcing the potential trade-off between engagement performance and off-platform outcomes.

Table 4: The Impact of the Comment Section on Costs Metrics

	Group Mean / (SD)				*t*–test *p*–value		
Variable	(1) Control	(2) Supportive	(3) Mixed	(4) Opposing	(1)–(2)	(1)–(3)	(1)–(4)
Spend per engagement	2.131 (0.337)	1.924 (0.208)	1.879 (0.283)	1.813 (0.392)	0.033	0.021	0.011
Spend per interaction	68.587 (42.266)	88.450 (48.005)	54.910 (29.021)	36.098 (13.111)	0.217	0.288	0.005
Spend per link click	4.685 (1.264)	4.497 (0.817)	4.449 (0.991)	3.996 (0.944)	0.598	0.549	0.066
Observations	18	18	18	18			

Notes: Each observation is an ad. Statistics are weighted by ad reach. P-values are based on t-tests using heteroskedasticity-robust standard errors.

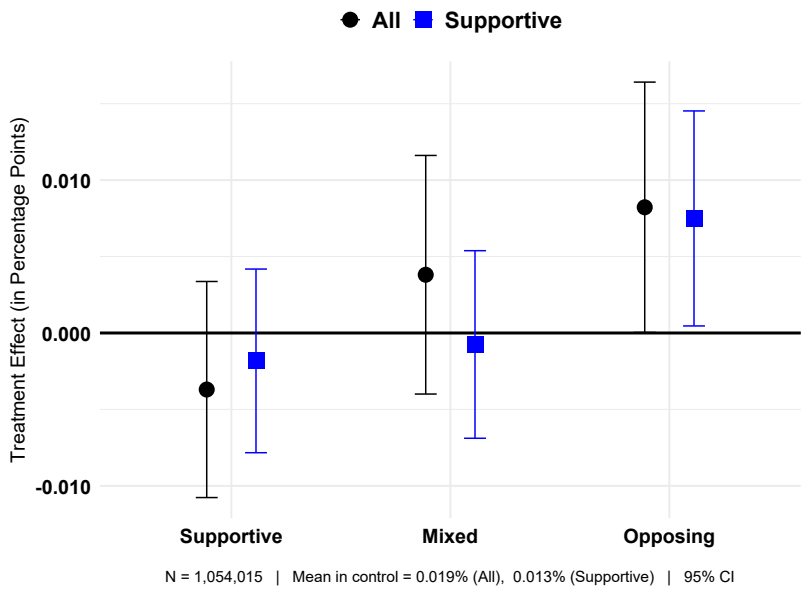
4.4.2 Valence of Subsequent Interactions

We next examine how comment stance affects the tone of subsequent engagement. Specifically, we compare effects on *all interactions*—reactions, comments, and shares—with effects on *supportive interactions*, defined as positive reactions (*likes*, *loves*, *cares*) and comments expressing agreement with the organization's message. Results for non-supportive or ambiguous reactions (e.g., *angry*, *sad*, *wow*) are reported in Appendix Table C3. The estimates are obtained from a parsimonious specification using data retrieved directly from the posts, without user-level covariates, and with heteroskedasticity-robust standard errors

Figure 7 shows that Opposing comments increase both overall and supportive interactions.

They raise all interactions by 0.0082 pp ($p < 0.05$) and supportive interactions by 0.0075 pp ($p < 0.05$), corresponding to a 56 percent increase relative to the control mean.[19] Mixed and Supportive comment sections have no statistically significant effects. Pairwise comparisons indicate that Opposing comments generate significantly more supportive follow-up interactions than either or Supportive conditions.

Figure 7: The Impact of the Comment Section on All and Supportive Interactions
Outcomes are expressed in % of total reach



Effects on non-supportive or ambiguous interactions are small and imprecisely estimated (Appendix Table C3). Taken together, these findings indicate that the additional engagement generated by Opposing comments is predominantly supportive in tone. This pattern suggests that the presence of opposing comments alters the composition of subsequent participation. One possibility is that visible disagreement increases the likelihood that users who support the original message choose to express their agreement. At the same time, the initial presence of negative comments may reduce the propensity of other users holding similar views to post additional negative responses.

4.4.3 Heterogeneous Effects Across Locations

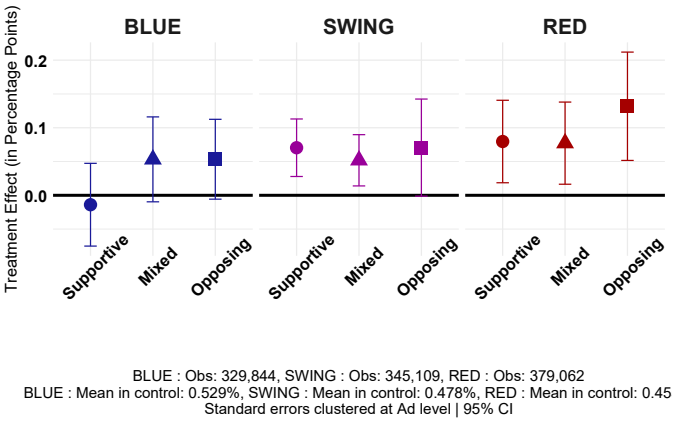
We next examine whether the impact of comment stance varies across areas with different prevailing political ideologies, distinguishing between *Blue* (mostly liberal), *Swing* (mixed), and *Red*

[19] 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.

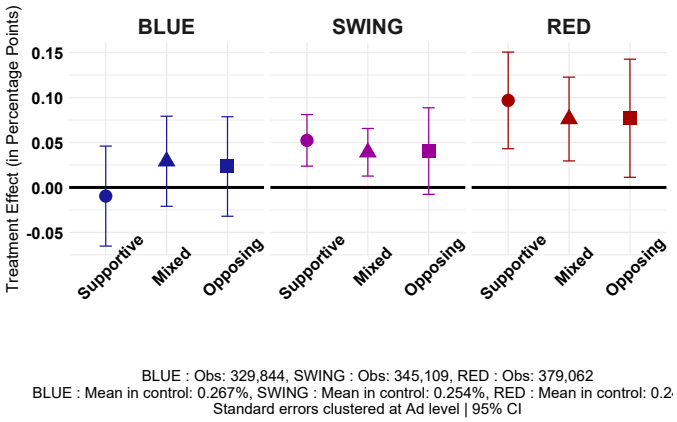
(mostly conservative) ZIP codes. Figure 8(a) displays the heterogeneous treatment effects on overall engagement, and Figures 8(b-d) summarize the corresponding effects for post expansions, interactions, and link clicks. The detailed point estimates are reported in Appendix Tables C4, C5, and C6. Across 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 8: Heterogeneous Treatment Effects across Locations
Outcomes are expressed in % of total reach

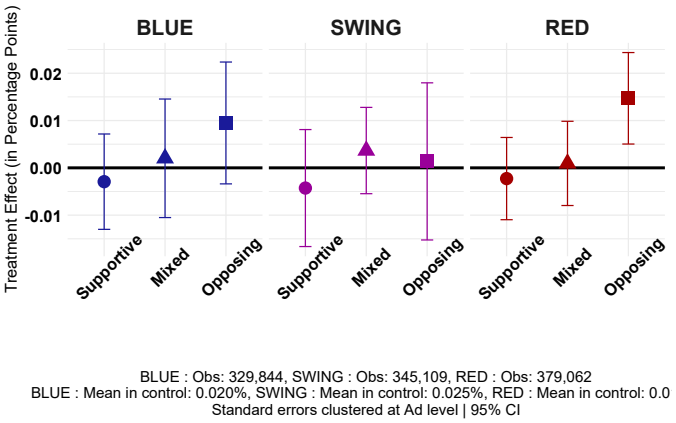
(a) All Engagement



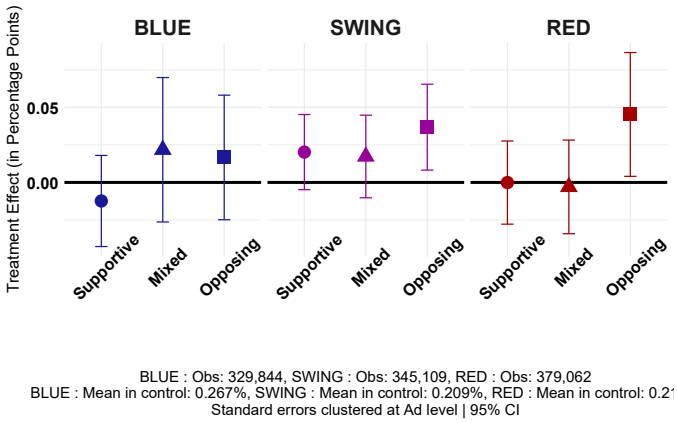
(b) Post Expansions



(c) Interactions (Reactions, Comments, Shares)



(d) Link Clicks



In *Blue areas*, effects are small and generally imprecise. Opposing and Mixed comments modestly 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 on the platform, 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 0.05-0.07 pp. 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 are both larger and more differentiated across comment stances. While post expansions are high and similar across conditions, Opposing comments generate the strongest responses in total engagement, which increases by 0.13 percentage points (pp, $p < 0.01$), and in link clicks, which rise by 0.045 pp ($p < 0.05$). These effects correspond to relative gains of roughly 20–30 percent compared with baseline means. Moreover, Opposing comments increase subsequent interactions by 0.015 pp, effectively doubling the interaction rate relative to the control group in these areas. Pairwise tests confirm that the differences between the Opposing and the Supportive and Mixed conditions are statistically significant for interactions ($p < 0.01$) and link clicks ($p < 0.05$). Overall, these results indicate that Opposing comments generate higher levels of participation and click-through activity, particularly 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, consistent with greater alignment between the campaign's message and prevailing ideology in more progressive areas.[20] Comment sections therefore appear to play a more pronounced role in conservative areas, where baseline engagement is lower. Consistent with this interpretation, our proxy for attention (i.e., post expansions) is large and statistically significant in Red areas but small and statistically insignificant in Blue areas, irrespective of comment stance. By making visible that the post has generated discussion—even when critical—the comment section may increase the salience of the content for users who might otherwise overlook it.

Taken together, these results reveal a clear ideological gradient. While engagement in more liberal and politically balanced areas responds similarly across comment tones, user activity in more conservative regions becomes markedly more sensitive to opposing narratives. The comment section thus amplifies attention and interaction most strongly when visible comments are counter-

[20]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.

attitudinal relative to the surrounding ideological environment. Two complementary mechanisms are consistent with this pattern. First, a *curiosity mechanism*: opposing comments draw attention by signaling disagreement, inducing users to click, expand, and engage to learn more. Second, an *identity-congruence mechanism*: comments that align with a user's political identity can directly motivate engagement. When progressive comments appear on progressive posts, curiosity is likely low because aligned comments are unsurprising, and the post itself already provides identity congruence, leaving limited marginal scope for comments to further increase engagement. By contrast, conservative comments on progressive posts generate higher curiosity due to their counter-attitudinal nature, while simultaneously providing identity congruence for conservative users, leading to larger observed effects on engagement.

4.4.4 Heterogeneous Effects across Genders

We next examine whether the impact of comment stance differs by gender. Figure 9 reports heterogeneous treatment effects for overall engagement, post expansions, interactions, and link clicks; detailed estimates are provided in Appendix Tables C7 and C8. Across outcomes, comment sections have little effect on female engagement, whereas effects are large and precisely estimated for male users.

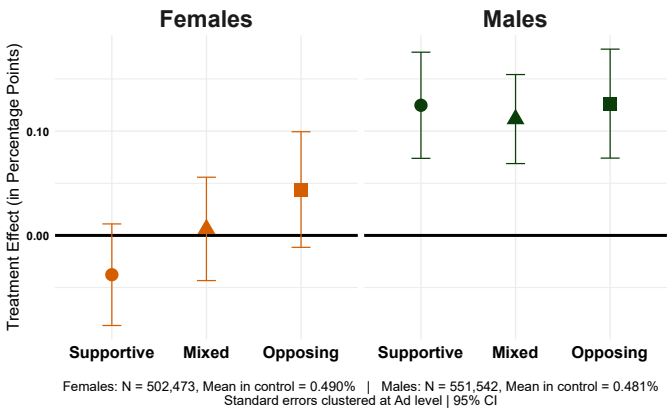
Among *female* users, treatment effects are generally small and statistically insignificant. Comment sections do not increase attention, as post expansions show no response. Opposing comments increase overall engagement by 0.044 pp (about 9 percent relative to control), but the estimate is not statistically significant. Mixed comments have no discernible effect, while Supportive comments are associated with modest declines in some outcomes, including link clicks and interactions. Overall, female engagement appears largely insensitive to both comment visibility and comment stance.

In contrast, *male* users exhibit substantial responses across outcomes. Attention to the comment section, as measured by post expansions, is statistically significant and larger in magnitude than for female users. Opposing comments increase overall engagement by 0.126 pp ($p < 0.01$), a 26 percent rise relative to control. Mixed and Supportive comments also generate sizable increases in engagement (0.112 and 0.125 pp, respectively). Opposing and Mixed comments significantly raise interaction rates, and all stances increase link clicks among men, with Opposing comments producing the largest click-through gains. These results indicate that men are considerably more responsive to both the presence and the stance of comments.

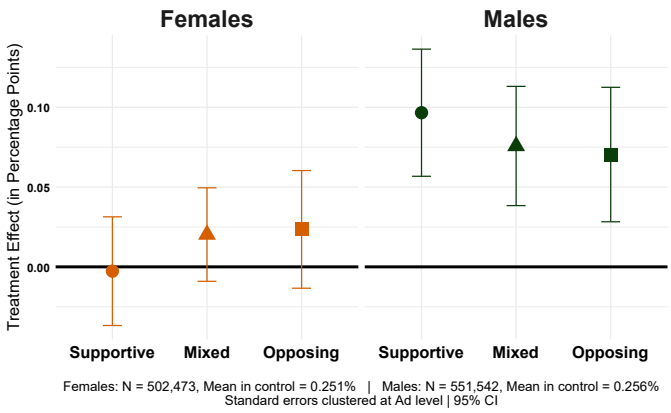
Taken together, the evidence reveals a pronounced gender gradient: the overall amplification effects of comment sections are driven primarily by male users. This heterogeneity underscores that the impact of visible discourse depends not only on comment content but also on the demographic composition of the audience.

Figure 9: Heterogeneous Treatment Effects Across Genders
Outcomes are expressed in % of total reach

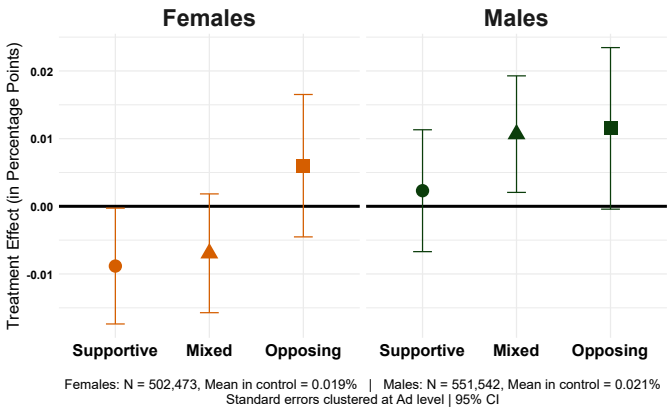
(a) All Engagement



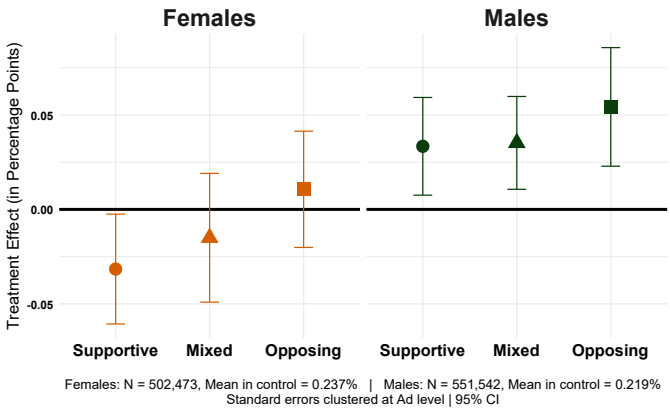
(b) Post Expansions



(c) Interactions (Reactions, Comments, Shares)



(d) Link Clicks



4.4.5 Robustness and Placebo Tests

We conduct several robustness checks to assess the sensitivity of our results. First, we re-estimate all main models using alternative combinations of control variables and fixed effects (Tables C9 and C10), and examine robustness to alternative estimators by employing a logit specification (Table C11). Second, we compute engagement rates directly at the ad level—without simulating individual-level data—and report the corresponding estimates together with heteroskedasticity-robust t-tests of equality across experimental conditions (Table C12). Finally, we assess the influence of specific geographic areas by sequentially excluding each zip code from the sample. Across all specifications, the results remain robust.

Furthermore, we test the impact of comments on landing-page views, an alternative measure of downstream behavior beyond link clicks. Appendix Table C13 suggests that the presence of any

comments increases off-platform engagement by 0.016 percentage points ($p < 0.10$), corresponding to a 9.4 percent increase relative to the control mean of 0.171 percent. Opposing comments yield the largest increase (0.019 percentage points, $p < 0.10$; 11.1 percent), followed by mixed (0.017 percentage points) and supportive comments (0.011 percentage points), both of which are not statistically significant. Differences across comment stances are small and statistically indistinguishable ($p > 0.50$), suggesting that comments modestly increase off-platform engagement regardless of stance. While landing-page views provide a useful proxy for off-platform behavior, this outcome should be interpreted with caution, as it is inferred via pixel-based tracking and is not measured at the unique user level.

One potential concern is that the higher engagement driven by opposing comments reflects not their stance per se, but other textual features correlated with stance. In particular, toxicity has been shown to drive up engagement (Beknazar-Yuzbashev et al., 2025). To address this, we classify comments using the Google Perspective API and find that average toxicity levels across the three treatment conditions are statistically indistinguishable (around 0.35), suggesting that toxicity is unlikely to confound our stance estimates. As a further robustness check, we exclude posts containing comments that exceed the Perspective API's recommended toxicity threshold of 0.7.[21] Appendix Table C16 shows that our results remain robust to this restriction, confirming that the engagement effect of opposing comments is not merely an artifact of toxic content.

5 The Impact of Comments on Attitude and Off-Platform Behavior

The field experiment provides causal evidence that comment sections influence observable engagement on the platform. However, platform data do not allow us to observe users' underlying beliefs, attitudinal shifts, or high-stakes behavioral outcomes. Nor can we directly test the psychological mechanisms, such as emotional arousal or curiosity.

To address these limitations, we conduct a complementary artefactual field experiment (Harrison and List, 2004). This design preserves key features of the organic online environment, including real social media posts, authentic comments collected from Facebook users, and behavioral incentives, while allowing us to measure beliefs, attitudes, and incentivized behaviors at the individual level. By embedding experimentally manipulated comment sections inside a controlled survey setting, we isolate how the presence and stance of comments affect attention, social perceptions, attitudes, and real monetary decisions. The artefactual field experiment is pre-registered and follows the design and empirical strategy specified in the pre-analysis plan.

The effect of comments on attitudes and behavior is theoretically ambiguous. First, through *persuasion*: exposure to opposing comments could shift attitudes in the direction of the comments themselves, as users update their beliefs in response to new arguments or information (DellaVigna

[21] `https://developers.perspectiveapi.com/s/about-the-api-score`

and Gentzkow, 2010). Alternatively, exposure to opposing views may trigger psychological reactance, leading users to entrench existing beliefs (Bail et al., 2018)—a backfire effect that would push attitudes in the opposite direction. Second, through *social norms*: if users infer from the comment section that a particular view is widely held, they may update their perceptions of the prevailing norm and conform accordingly (Bursztyn and Jensen, 2017). In this case, comments would shift attitudes in the direction of the expressed views regardless of their argumentative content. Third, through *social learning*: users may treat the volume and content of comments as a signal of the post's credibility or the organization's legitimacy, updating their assessments accordingly (Banerjee, 1992; Muchnik, Aral and Taylor, 2013). It is also possible that comments have little effect – political and social attitudes are difficult to change, and a single exposure to a comment section may be insufficient to move them (Haaland and Roth, 2023). The net effect of comments on attitudes and behavior is therefore an empirical question.

5.1 Design

5.1.1 Recruitment and Sample

We recruited approximately 5,000 participants from Prolific. Eligibility criteria required participants to be U.S. residents aged 18–64, to have completed at least 20 previous Prolific submissions, and to maintain an approval rate of at least 95 percent. Pilot participants were excluded. To enable the study of heterogeneous effects, recruitment was stratified to achieve approximate balance across self-identified political ideology (conservative, moderate, liberal) and gender.[22]

5.1.2 Survey Structure

The survey proceeded in four stages. Participants first completed baseline measures capturing political ideology, racial attitudes, beliefs about others' views, social media usage, and demographics. They were then shown multiple social media posts promoting racial justice, drawn from our collaboration with Color of Change and mirroring those used in the field experiment. Next, the visibility and stance of the comment section were randomized at the participant level. Finally, participants completed post-exposure measures capturing attention, beliefs, attitudes, behavioral choices, cognitive responses, emotional responses, and open-ended questions. To minimize experimenter demand effects and preserve ecological validity, participants were instructed to react to each post as if they encountered it naturally on their Facebook feed.

[22]Participants were informed that the survey would take approximately 10 minutes and that they could earn a bonus of up to $1 based on performance in incentivized belief elicitation, in addition to being entered into a $100 lottery.

5.1.3 Treatment Conditions

Participants were randomly assigned with equal probability to one of three conditions. In the *No Comments* control condition, posts were displayed without any visible comment section. In the *Supportive Comments* condition, each post displayed two comments supportive of racial justice. In the *Opposing Comments* condition, each post displayed two comments expressing opposition to racial justice.

The posts themselves were identical across conditions. The comments were real comments collected from Facebook users during the field experiment described earlier. To protect user privacy, names and profile images were replaced with neutral placeholders. In the comment conditions, we displayed two comments of the same stance, one using a male name and one using a female name, in order to avoid confounding stance with perceived commenter gender. Within a treatment condition, the stance of comments was consistent across all posts shown to a participant. Thus, the only dimension varying across individuals was the presence and stance of the comment section.

5.2 Outcomes

We measure a set of outcomes capturing attention, beliefs, attitudes, behavioral decisions, and psychological responses. When multiple measures capture a common construct, we aggregate them into standardized indices using inverse-covariance weighting following Anderson (2008). All component variables are re-oriented so that higher values reflect greater alignment with the organization or post.

Attention Attention is measured as the total time spent viewing posts, aggregated across exposures at the participant level, using $\log(1 + \text{time})$ to reduce the influence of outliers.

Beliefs about Others and Commenters We measure beliefs along two dimensions. First, participants estimated the percentage of individuals in the broader population holding conservative attitudes. Responses were incentivized using a quadratic scoring rule, with potential earnings of up to $1 based on accuracy. Second, we elicit perceptions of the ideological composition of the comment section. Participants were asked which group they believe is more likely to comment on the post, distinguishing between progressive and conservative individuals. These measures allow us to examine whether exposure to supportive or opposing comments shifts perceptions of the prevailing opinion environment and of who publicly engages with the content.

Attitudes toward Racial Issues and the Organization We construct a Racial Attitudes Index aggregating measures of attitudes toward the organization, perceived importance of the racial justice issues covered in the posts, willingness to engage in cross-partisan discussion, and

opinions toward Black Americans. All components are re-oriented so that higher values reflect greater alignment with the organization and are combined using inverse-covariance weighting.

Behavioral Outcomes We measure two high-stakes behavioral outcomes. Participants were given the opportunity to subscribe to the organization's mailing list by voluntarily providing their email address, reflecting willingness to engage beyond the survey environment. Participants were also enrolled in a lottery and asked to pre-commit to a donation amount. If selected as winners, the chosen amount was deducted from their prize and transferred directly to the organization. We analyze both the extensive margin, defined as an indicator for making a positive donation, and the intensive margin, defined as the committed donation amount.

Cognitive and Emotional Responses We construct a Cognitive Response Index combining participants' reported interest in seeing additional comments and their assessment of how thought-provoking they found the post. For participants in the treatment conditions, we additionally incorporate the perceived thought-provoking nature of the comments. For participants in the treatment conditions, we construct an Emotional Response Index combining Likert-scale measures of anger and annoyance triggered by the comments.

Additional Outcomes Additional outcomes include attitudes toward related organizations such as Black Lives Matter, curiosity about the topic versus curiosity about the organization, beliefs about which groups are more likely to comment by ideology and gender, and textual analysis of open-ended responses.

5.3 Empirical Strategy

Let Y_i denote an outcome for participant i. We estimate:

$$Y_i = \beta_1 Supportive_i + \beta_2 Opposing_i + \tau X_i + \varepsilon_i,$$

where the omitted category is No Comments. Covariates X_i include baseline racial attitudes, ideology, beliefs about others, and demographics. Standard errors are robust to heteroskedasticity. Moreover, the coefficient $\beta_2 - \beta_1$ captures the differential effect of opposing versus supportive comments. We pre-specify heterogeneity analyses by ideology and gender, and we discuss these in the results.

5.3.1 Descriptive Statistics

Appendix Figure D1 reports the distribution to the question of how often the respondent read or check comments on social media. Approximately 85% of survey participants report that they

sometimes, often, or very often read comments on social media, and more than 50% report doing so often or very often. This descriptive evidence confirms that comment sections are a salient feature of online consumption and motivates our focus on their effects..

Table 5 reports baseline characteristics by experimental condition and balance tests. Across all observable covariates, the treatment groups are well-balanced relative to the control group. Differences in means are small in magnitude, and the vast majority of pairwise t-tests fail to reject equality at conventional significance levels, with only a few marginal imbalances in age, ethnicity, and race. Importantly, joint orthogonality F-tests fail to reject equality across all pairwise comparisons, confirming that treatment assignment is orthogonal to observed baseline characteristics. To account for these marginal imbalances, we include age, ethnicity, and race as controls in the main specifications, as specified in the pre-analysis plan.

Table 5: Covariate Balance Across Experimental Arms

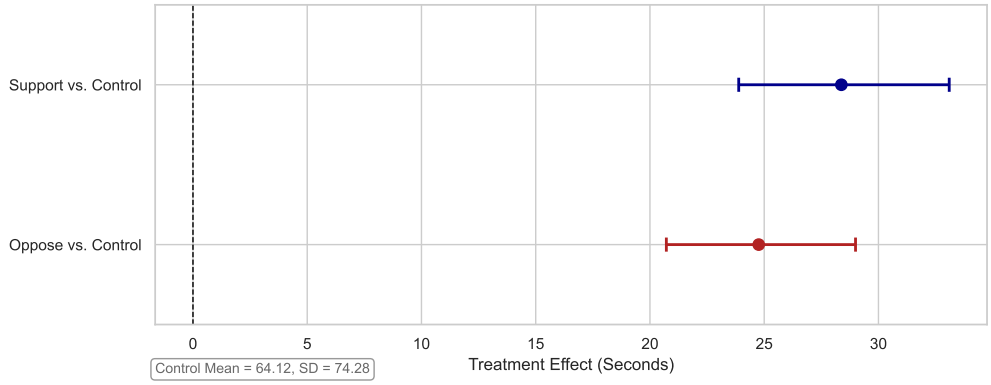
Variable	(1) Control Mean/SD	(2) Support Mean/SD	(3) Oppose Mean/SD	t-test p-value (1)-(2)	t-test p-value (1)-(3)	t-test p-value (2)-(3)
Age (Above Median)	0.48 (0.50)	0.49 (0.50)	0.51 (0.50)	0.54	0.10*	0.29
Baseline Racial Index (Above Median)	0.49 (0.50)	0.50 (0.50)	0.48 (0.50)	0.68	0.63	0.36
Education (Above Median)	0.16 (0.36)	0.15 (0.36)	0.18 (0.38)	0.84	0.17	0.11
Social Media Use (Above Median)	0.40 (0.49)	0.41 (0.49)	0.39 (0.49)	0.69	0.39	0.20
Reads Comments Frequency (Above Median)	0.18 (0.38)	0.19 (0.39)	0.18 (0.38)	0.58	0.98	0.55
Equal Rights (Above Median)	0.48 (0.50)	0.50 (0.50)	0.47 (0.50)	0.33	0.69	0.16
Race View (Above Median)	0.25 (0.43)	0.26 (0.44)	0.25 (0.43)	0.50	0.86	0.61
Female	0.51 (0.50)	0.49 (0.50)	0.49 (0.50)	0.36	0.35	0.99
White	0.78 (0.41)	0.78 (0.41)	0.81 (0.39)	0.99	0.07*	0.07*
Black / African American	0.12 (0.32)	0.12 (0.33)	0.11 (0.31)	0.69	0.36	0.17
Asian American / Pacific Islander	0.09 (0.28)	0.07 (0.26)	0.07 (0.26)	0.32	0.27	0.94
Hispanic	0.09 (0.29)	0.13 (0.33)	0.10 (0.30)	0.01***	0.59	0.03**
Democrat	0.32 (0.47)	0.31 (0.46)	0.33 (0.47)	0.91	0.44	0.36
Republican	0.30 (0.46)	0.30 (0.46)	0.28 (0.45)	0.97	0.38	0.35
N	1187	1281	1390			
F-test of joint significance (p-value)				0.623	0.560	0.271
F-test, number of observations				2468	2577	2671

Notes: This table reports covariate balance across the three treatment arms of the artefactual field experiment. Columns (1)–(3) report means with standard deviations in parentheses for the control, supportive comments, and opposing comments groups, respectively. The remaining columns report p-values from pairwise t-tests of equality of means. The final rows report F-tests of joint significance across all covariates for each pair of arms. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5.4 Results

Attention As shown in Figure 10, both Supportive and Opposing comments significantly increase time spent viewing the post by approximately 25-30 seconds, a 40-45% increase relative to the no-comments condition. These results indicate that the presence of a comment section increases the attention users pay to the post, regardless of stance.

Figure 10: Time Spent on Post



Attitudes Figure 11 shows that Opposing comments shift attitudes in a less progressive direction relative to the control condition by approximately 0.13 standard deviations. The effects are concentrated in measures of NGO favorability and willingness to engage with conservatives, both of which decline significantly. Supportive comments have comparatively small and statistically insignificant effects. Thus, exposure to Opposing comments shifts attitudes away from the organization's position, with potential implications for support of both the NGO and the cause it represents.

Figure 11: Post-Exposure Racial Attitudes

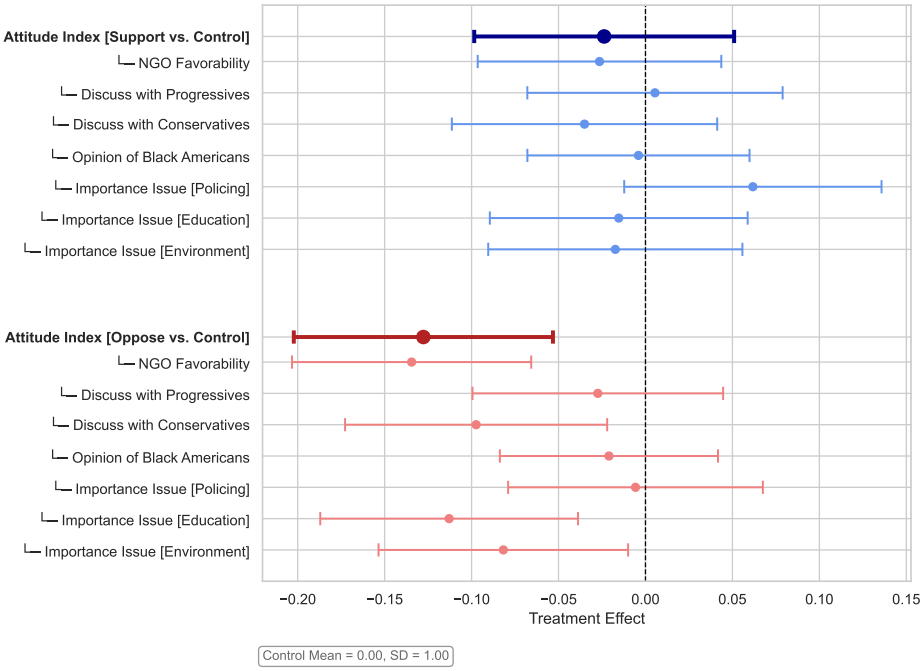
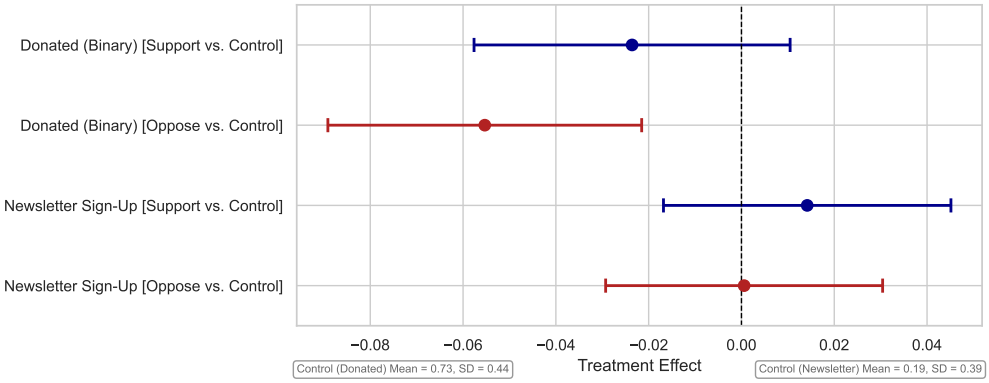
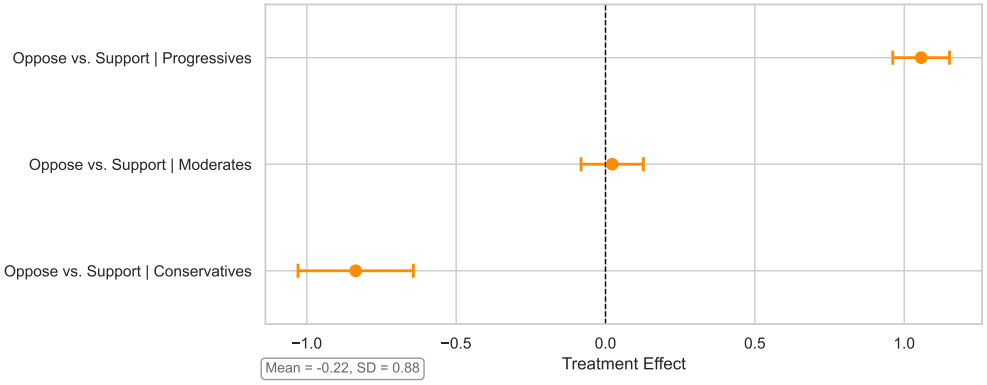


Figure 12: Donations and Newsletter Sign-Up



Off-platform Behavior As shown in Figure 12, Opposing comments significantly reduce the likelihood of making an incentivized donation relative to the control condition by 5.9 percentage points, corresponding to a 8% decline. Supportive comments have no statistically significant effect. Neither treatment meaningfully affects newsletter sign-up. Thus, while Opposing comments increase attention, they reduce high-stakes financial support to the organization.

Figure 13: Emotional Responses: Anger and Annoyance



Emotional Responses Figure 13 shows that counter-attitudinal comments generate higher levels of anger and annoyance, particularly when visible comments conflict with respondents' political identity. Relative to Supportive comments, Opposing comments increase reported anger and annoyance among progressive respondents, have no statistically significant effect among moderates, and reduce negative emotional responses among conservative respondents. These findings indicate that comment stance shapes not only engagement but also emotional reactions, and that identity congruence plays a central role in determining how users respond to visible disagreement.

Results Heterogeneity We further examine heterogeneity in both behavioral and attitudinal outcomes across baseline racial attitudes, party affiliation, gender, and age. Overall, we find limited evidence of heterogeneous effects across these dimensions. The estimated treatment effects are broadly similar across groups, suggesting that the impact of comment stance is relatively uniform across different types of users.

Discussion Taken together, the survey evidence reveals a clear tension between engagement and downstream outcomes. Comment sections—especially those featuring opposing remarks—attract attention but can shift attitudes and reduce financial support, highlighting a trade-off between amplifying engagement and preserving support for the organization and the underlying message. The field experiment reinforces this pattern: comment sections on progressive posts are disproportionately populated by conservative voices, and both their presence and stance shape subsequent on-platform behavior. Opposing comments increase clicks and interactions in the field setting, particularly in conservative areas and among male users. However, the survey evidence shows that exposure to counter-attitudinal comments generates negative emotional responses and shifts attitudes and incentivized behaviors in a less progressive direction. Taken together, these findings indicate that comment environments that maximize short-run engagement may simultaneously entail longer-run costs in terms of support for the organization and alignment with its mission.

6 Conclusion

This paper provides causal evidence that comment sections shape both on-platform engagement and downstream attitudes and behavior. In a large-scale Facebook field experiment, we show that the presence of a comment section increases engagement, and that comment stance affects how users respond to content. Opposing comments, in particular, amplify interactions (e.g., comments and reactions) and link clicks relative to supportive comments, with especially strong effects among men and in conservative areas. Because comment sections are often populated by a vocal minority, these findings imply that the visible opinions of a few users can meaningfully shape the on-platform behavior of many others.

Our complementary artefactual field experiment reveals that higher visible engagement does not necessarily imply greater support. While opposing comments increase attention, they also shift attitudes in a less progressive direction and reduce incentivized donations. Exposure to counter-attitudinal comments generates anger and annoyance among identity-incongruent users, underscoring that visible disagreement affects not only engagement behavior but also emotional responses and evaluations of the organization. Together, the evidence points to a tension between short-run on-platform engagement gains and potential off-platform longer-run costs.

These results speak to broader concerns about online discourse. Comment sections can amplify

polarized or oppositional narratives in ways that may not reflect the broader audience, potentially complicating the ideal of social media as a deliberative public sphere. At the same time, our findings do not imply that removing comment sections is a straightforward solution: comments increase attention and facilitate participation, and in some contexts may stimulate supportive expression. The challenge lies in how visible discourse is structured and moderated.

For platforms and content producers, moderation involves clear trade-offs. Tolerating contentious or opposing comments can boost engagement and reduce advertising costs, but may undermine brand safety and shift attitudes away from the content creator's objectives. Importantly, platform incentives to maximize engagement may not align with the incentives of firms, nonprofits, or political actors seeking to preserve brand integrity or policy support. Understanding these misalignments is central to current debates over platform governance and content moderation.

Methodologically, we introduce a scalable experimental pipeline that manipulates comment visibility and stance using platform-native tools while preserving ecological validity through organic user-generated content. This framework can be extended to other domains, such as commercial products or public health campaigns, to better understand how visible online discourse shapes behavior and to inform the design of comment environments that balance engagement with broader social and organizational goals.

Our study point to several promising directions for future research. First, our analysis focuses on a single platform and on sponsored posts rather than organic feed exposure, so the magnitude and nature of comment effects may vary across platforms, ranking systems, and moderation regimes. Second, although our design preserves ecological validity, the treatment involves a relatively stylized comment environment with a small number of visible comments, and repeated exposure or richer thread dynamics may generate different responses over time. Finally, we examine comment sections in a specific substantive setting—racial-justice advocacy content produced by a nonprofit organization on Facebook—and the effects we document may differ in commercial settings, where users may have different priors, stakes, and motives for engaging with posts. Future work should therefore examine whether similar engagement–welfare trade-offs arise for brands, product advertising, and other domains such as public health or political communication.

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, Michael L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association*, 103(484): 1481–1495.

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*.

Bail, Christopher A. 2021. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing.* Princeton, NJ:Princeton University Press.

Bail, Christopher A, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. "Exposure to opposing views on social media can increase political polarization." *Proceedings of the National Academy of Sciences*, 115(37): 9216–9221.

Banerjee, Abhijit V. 1992. "A Simple Model of Herd Behavior." *Quarterly Journal of Economics*, 107(3): 797–817.

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

Beknazar-Yuzbashev, George, Rafael Jiménez-Durán, and Mateusz Stalinski. 2024. "A model of harmful yet engaging content on social media." Vol. 114, 678–683, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Beknazar-Yuzbashev, George, Rafael Jiménez-Durán, Jesse McCrosky, and Mateusz Stalinski. 2025. "Toxic content and user engagement on social media: Evidence from a field experiment." CESifo Working Paper.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch. 1992. "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades." *Journal of Political Economy*, 100(5): 992–1026.

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.

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.

Bursztyn, Leonardo, and Robert Jensen. 2017. "Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure." *Annual Review of Economics*, 9: 131–153.

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.

Gauthier, Germain, Roland Hodler, Philine Widmer, and Ekaterina Zhuravskaya. 2026. "The political effects of X's feed algorithm." *Nature*, 1–8.

Germano, Fabrizio, Vicenç Gómez, and Francesco Sobbrio. 2026. "Ranking for engagement: How social media algorithms fuel misinformation and polarization." *Journal of Public Economics*, 255: 105589.

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.

Guess, Andrew M, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, et al. 2023*a*. "How do social media feed algorithms affect attitudes and behavior in an election campaign?" *Science*, 381(6656): 398–404.

Guess, Andrew M, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, et al. 2023*b*. "Reshares on social media amplify political news but do not detectably affect beliefs or opinions." *Science*, 381(6656): 404–408.

Haaland, Ingar, and Christopher Roth. 2023. "Beliefs about racial discrimination and support for pro-black policies." *Review of Economics and Statistics*, 105(1): 40–53.

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

Harrison, Glenn W, and John A List. 2004. "Field experiments." *Journal of Economic literature*, 42(4): 1009–1055.

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 4990476*.

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

Karpowitz, Christopher F, and Tali Mendelberg. 2014. *The silent sex: Gender, deliberation, and institutions.* Princeton University Press.

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

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.

Levy, Ro'ee. 2021. "Social Media, News Consumption, and Polarization: Evidence from a Field Experiment." *American Economic Review*, 111(3): 831–870.

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."

Muchnik, Lev, Sinan Aral, and Sean J Taylor. 2013. "Social Influence Bias: A Randomized Experiment." *Science*, 341(6146): 647–651.

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

Nyhan, Brendan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, et al. 2023. "Like-minded sources on Facebook are prevalent but not polarizing." *Nature*, 620(7972): 137–144.

Oswald, Lisa, William Small Schulz, and Philipp Lorenz-Spreen. 2025. "Disentangling participation in online political discussions with a collective field experiment." *Science Advances*, 11(50): eady8022.

Oswald, Lisa, W Schulz, Ralph Hertwig, David Lazer, and Sebastian Stier. 2025. "The tip of the iceberg: How the social media production-consumption gap distorts public opinion for citizens and researchers."

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."

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.

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

Social Media Comments

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.

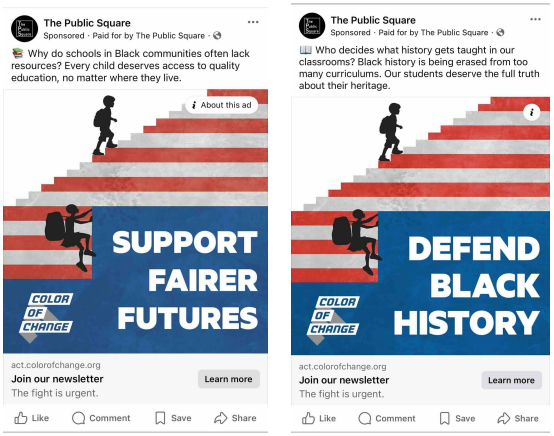
- Environmental Justice: Black communities often live near pollution sources like factories and highways, leading to higher rates of health issues such as asthma. These neighborhoods are frequently overlooked in clean-up efforts and lack green spaces. Environmental justice aims to provide Black communities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

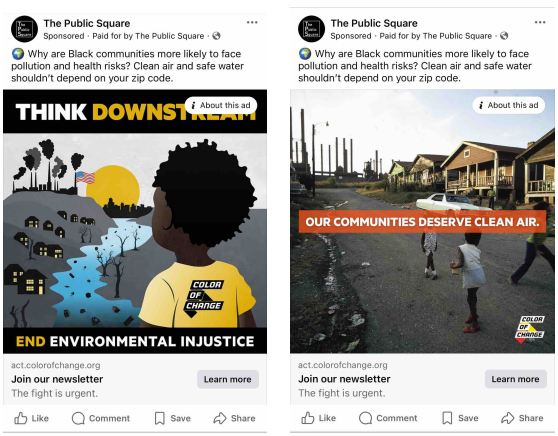
- Education Reform: Black students often attend underfunded schools with fewer resources, larger classes, and limited access to advanced courses. These disparities create achievement gaps and limit future opportunities. Education reform seeks equitable funding and support to provide Black students with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

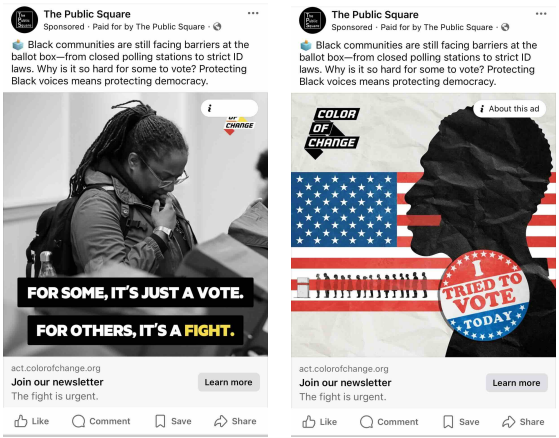
Figure A1: Ad Banners and Headlines



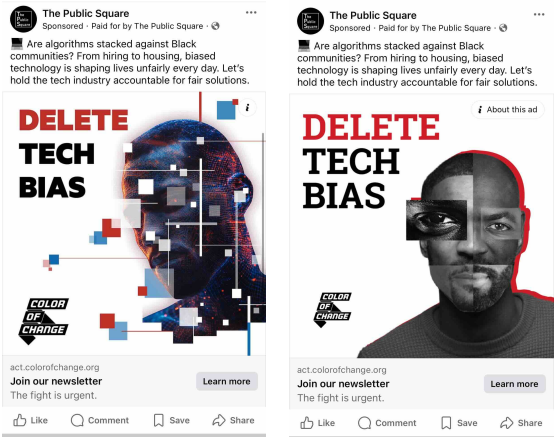
(a) Education



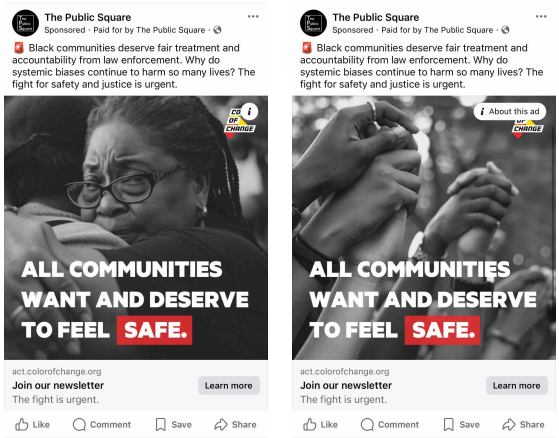
(b) Environment



(c) Voting



(d) Technology



(e) Police

A.2 Pre-tests

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.[23] 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.

Table A2 presents the aggregate results for Pre-tests B and C.[24] 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 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.

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

[23] 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.

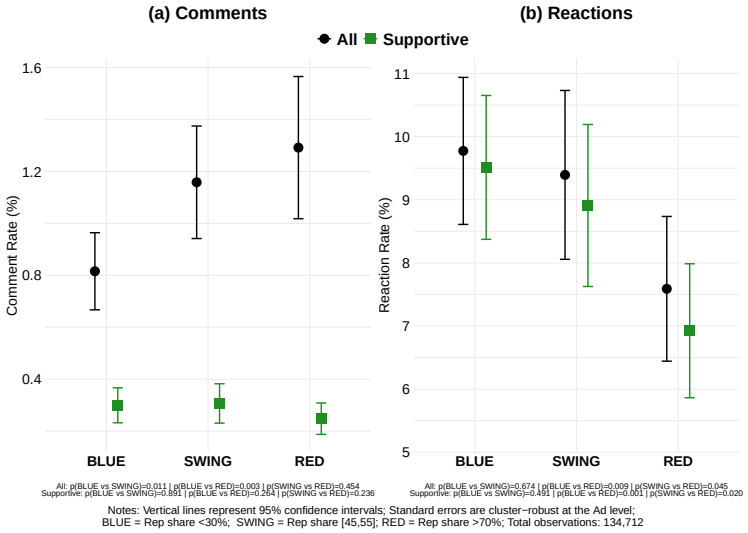
[24] The banners on education were excluded from these tests due to their low performance in Pre-test A and budgetary considerations.

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

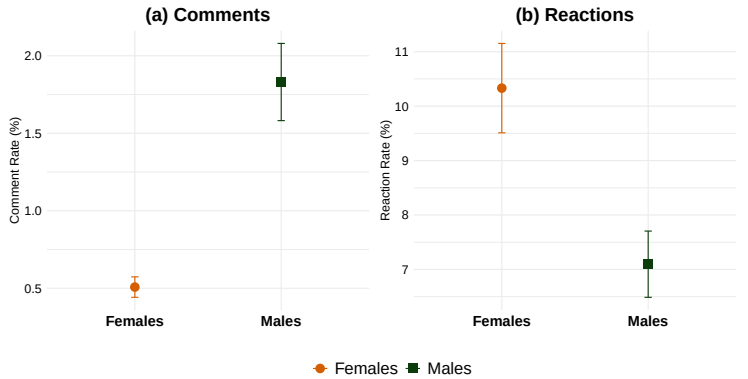
B Additional Results for Descriptive Evidence

Figure B1: Comment and Reaction Rates by Valence and Location



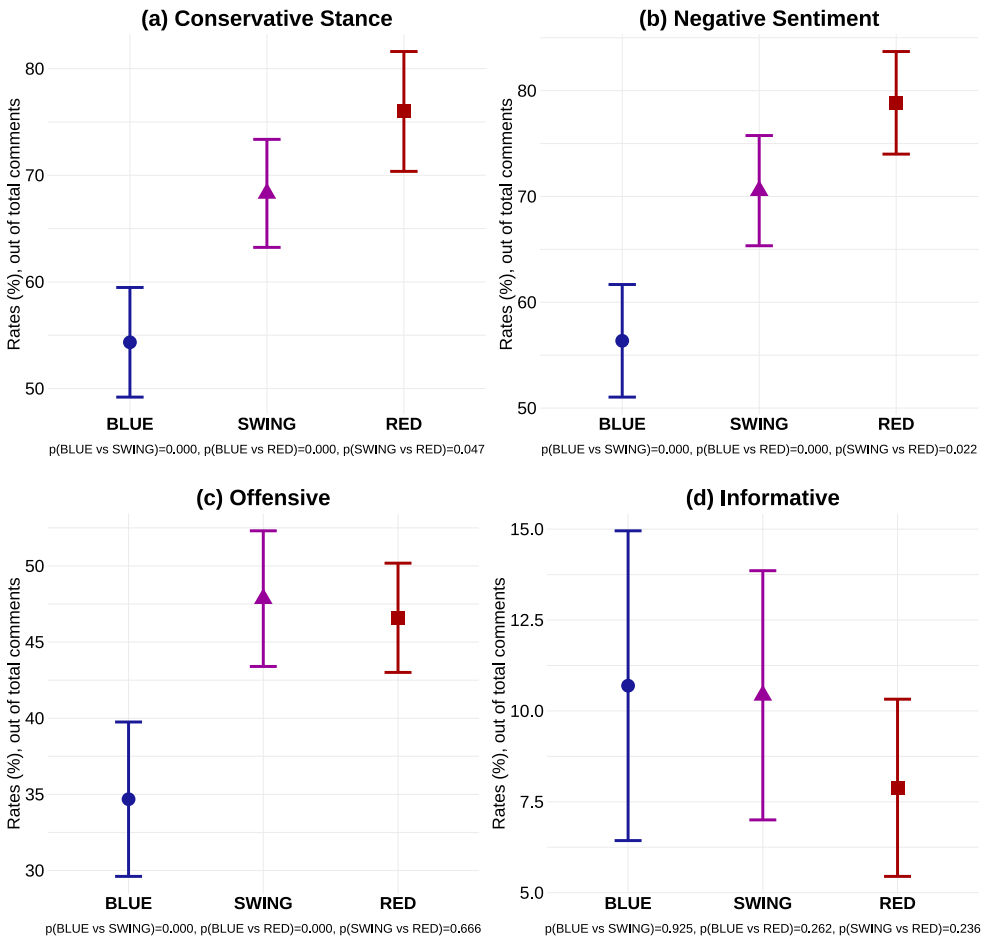
Notes: Vertical lines represent 95% confidence intervals; Standard errors are cluster−robust at the Ad level;
BLUE = Rep share <30%; SWING = Rep share [45,55]; RED = Rep share >70%; Total observations: 134,712

Figure B2: 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

Figure B3: Comments Characteristics by Location



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

C Additional Results for the Facebook Field Experiment

Table C1: Audience Saturation under Alternative Audience Size and Reach Estimates

		Campaign Reach Estimate	
		Lower Bound	Upper Bound
Audience Size Estimate		993,572	1,051,582
Lower Bound	904,700	1.098	1.162
Midpoint	975,000	1.019	1.079
Upper Bound	1,045,300	0.951	1.006

Notes: Audience size estimates correspond to Meta's "Estimated Audience Size," defined as the number of accounts meeting the specified targeting criteria. We record these estimates at the start of the campaign. Meta reports these values as potential reach ranges based on recent platform activity and targeting configuration. Estimates may fluctuate over time as platform usage and measurement systems evolve. Campaign reach estimates are measured either at the campaign level (lower bound) or at the ad level and then aggregated across ads (upper bound). Saturation is computed as the reach estimate divided by the corresponding audience estimate.

Table C2: The Impact of the Comment Section on On-platform User Engagement

	Dependent variable:							
	All Engagement		Post Expansions		Interactions		Link Clicks	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Any comments	0.065***		0.049***		0.003		0.016**	
	(0.015)		(0.013)		(0.003)		(0.008)	
Opposing		0.087***		0.048***		0.009**		0.034***
		(0.021)		(0.017)		(0.004)		(0.011)
Mixed		0.062***		0.049***		0.002		0.012
		(0.017)		(0.013)		(0.003)		(0.011)
Supportive		0.047***		0.049***		−0.003		0.003
		(0.018)		(0.015)		(0.003)		(0.008)
Constant	0.538***	0.539***	0.158**	0.158**	0.004	0.004	0.379***	0.380***
	(0.120)	(0.115)	(0.064)	(0.063)	(0.010)	(0.010)	(0.070)	(0.066)
Zipcode Set FEs	✓	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Zipcode Set × Controls	✓	✓	✓	✓	✓	✓	✓	✓
Mean Y in Control	0.485	0.485	0.254	0.254	0.020	0.020	0.228	0.228
p(Support vs. Oppose)		0.059		0.974		0.002		0.008
p(Oppose vs. Mixed)		0.216		0.935		0.091		0.106
p(Support vs. Mixed)		0.394		0.960		0.078		0.432
Observations	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015
R^2	0.0004	0.0004	0.001	0.001	0.0003	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 C3: The Impact of the Comment Section on the Valence of Subsequent Interactions (in %)

	Dependent variable:					
	All		Supporting		Non-supporting	
	(1)	(2)	(3)	(4)	(5)	(6)
Any comments	0.0028		0.0017		0.0011	
	(0.0032)		(0.0026)		(0.0017)	
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***	0.0210***	0.0143***	0.0143***	0.0067***	0.0067***
	(0.0034)	(0.0034)	(0.0027)	(0.0027)	(0.0019)	(0.0019)
Zipcode Set FEs	✓	✓	✓	✓	✓	✓
Mean Y in Control (C)	0.0190	0.0190	0.0133	0.0133	0.0057	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	1,054,015	1,054,015	1,054,015
R^2	0.00001	0.00002	0.000005	0.00001	0.000002	0.00001

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

Table C4: The Impact of the Comment Section in Blue Areas

	Dependent variable:							
	All Engagement		Post Expansions		Interactions		Link Clicks	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Any comments	0.031		0.014		0.003		0.009	
	(0.025)		(0.024)		(0.005)		(0.015)	
Opposing		0.053*		0.023		0.010		0.017
		(0.030)		(0.028)		(0.007)		(0.021)
Mixed		0.053*		0.029		0.002		0.022
		(0.032)		(0.026)		(0.006)		(0.024)
Supportive		−0.014		−0.010		−0.003		−0.012
		(0.031)		(0.028)		(0.005)		(0.016)
Constant	0.559***	0.561***	0.155**	0.156**	−0.007	−0.007	0.416***	0.416***
	(0.129)	(0.117)	(0.072)	(0.066)	(0.013)	(0.012)	(0.077)	(0.072)
Zipcode Set FEs	✓	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Zipcode Set × Controls	✓	✓	✓	✓	✓	✓	✓	✓
Mean Y in Control	0.529	0.529	0.267	0.267	0.020	0.020	0.267	0.267
p(Support vs. Oppose)		0.052		0.191		0.029		0.195
p(Oppose vs. Mixed)		0.996		0.788		0.270		0.862
p(Support vs. Mixed)		0.065		0.074		0.363		0.184
Observations	329,844	329,844	329,844	329,844	329,844	329,844	329,844	329,844
R^2	0.0004	0.0004	0.0005	0.0005	0.0003	0.0003	0.0002	0.0002

Notes: Standard errors are clustered at the advertisement level (24 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 C5: The Impact of the Comment Section in Swing Areas

	Dependent variable:							
	All Engagement		Post Expansions		Interactions		Link Clicks	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Any comments	0.064***		0.044***		0.0003		0.025**	
	(0.020)		(0.014)		(0.005)		(0.011)	
Opposing		0.071*		0.040*		0.001		0.037**
		(0.037)		(0.025)		(0.009)		(0.015)
Mixed		0.052***		0.039***		0.004		0.017
		(0.019)		(0.014)		(0.005)		(0.014)
Supportive		0.070***		0.052***		−0.004		0.020
		(0.022)		(0.015)		(0.006)		(0.013)
Constant	0.317***	0.317***	0.180***	0.180***	0.021	0.021	0.110**	0.109**
	(0.067)	(0.068)	(0.046)	(0.047)	(0.017)	(0.017)	(0.051)	(0.053)
Zipcode Set FEs	✓	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Zipcode Set × Controls	✓	✓	✓	✓	✓	✓	✓	✓
Mean Y in Control	0.478	0.478	0.254	0.254	0.025	0.025	0.209	0.209
p(Support vs. Oppose)		0.996		0.628		0.556		0.282
p(Oppose vs. Mixed)		0.604		0.956		0.789		0.237
p(Support vs. Mixed)		0.382		0.327		0.219		0.844
Observations	345,109	345,109	345,109	345,109	345,109	345,109	345,109	345,109
R^2	0.0005	0.0005	0.001	0.001	0.0003	0.0003	0.0002	0.0002

Notes: Standard errors are clustered at the advertisement level (24 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 C6: The Impact of the Comment Section in Red Areas

	Dependent variable:							
	All Engagement		Post Expansions		Interactions		Link Clicks	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Any comments	0.096***		0.083***		0.004		0.014	
	(0.030)		(0.024)		(0.004)		(0.014)	
Opposing		0.132***		0.077**		0.015***		0.045**
		(0.041)		(0.034)		(0.005)		(0.021)
Mixed		0.077**		0.076***		0.001		−0.003
		(0.031)		(0.024)		(0.004)		(0.016)
Supportive		0.080**		0.097***		−0.002		−0.0001
		(0.031)		(0.027)		(0.004)		(0.014)
Constant	0.482***	0.482***	0.286**	0.286**	−0.015**	−0.015**	0.212***	0.212***
	(0.152)	(0.145)	(0.121)	(0.120)	(0.006)	(0.007)	(0.049)	(0.041)
Zipcode Set FEs	✓	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Zipcode Set × Controls	✓	✓	✓	✓	✓	✓	✓	✓
Mean Y in Control	0.455	0.455	0.242	0.242	0.016	0.016	0.210	0.210
p(Support vs. Oppose)		0.145		0.539		0.000		0.027
p(Oppose vs. Mixed)		0.127		0.976		0.001		0.027
p(Support vs. Mixed)		0.919		0.349		0.370		0.849
Observations	379,062	379,062	379,062	379,062	379,062	379,062	379,062	379,062
R^2	0.0005	0.0005	0.001	0.001	0.0002	0.0002	0.0001	0.0002

Notes: Standard errors are clustered at the advertisement level (24 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 C7: The Impact of the Comment Section on Female Engagement

	Dependent variable:							
	All Engagement		Post Expansions		Interactions		Link Clicks	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Any comments	0.004		0.014		−0.003		−0.012	
	(0.021)		(0.015)		(0.004)		(0.013)	
Opposing		0.044		0.024		0.006		0.011
		(0.028)		(0.019)		(0.005)		(0.016)
Mixed		0.006		0.020		−0.007		−0.015
		(0.025)		(0.015)		(0.004)		(0.017)
Supportive		−0.038		−0.003		−0.009**		−0.032**
		(0.025)		(0.017)		(0.004)		(0.015)
Constant	0.558***	0.560***	0.114	0.115	0.002	0.003	0.446***	0.447***
	(0.130)	(0.116)	(0.109)	(0.107)	(0.003)	(0.004)	(0.118)	(0.110)
Zipcode Set FEs	✓	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Zipcode Set × Controls	✓	✓	✓	✓	✓	✓	✓	✓
Mean Y in Control	0.490	0.490	0.251	0.251	0.019	0.019	0.237	0.237
p(Support vs. Oppose)		0.004		0.146		0.002		0.009
p(Oppose vs. Mixed)		0.182		0.834		0.008		0.166
p(Support vs. Mixed)		0.080		0.097		0.617		0.354
Observations	502,473	502,473	502,473	502,473	502,473	502,473	502,473	502,473
R^2	0.001	0.001	0.001	0.001	0.0004	0.0005	0.0002	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, along with two-way interactions. Interactions include comments, reactions and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table C8: The Impact of the Comment Section on Male Engagement

	Dependent variable:							
	All Engagement		Post Expansions		Interactions		Link Clicks	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Any comments	0.121***		0.081***		0.008**		0.041***	
	(0.021)		(0.016)		(0.004)		(0.012)	
Opposing		0.126***		0.070***		0.012*		0.054***
		(0.027)		(0.021)		(0.006)		(0.016)
Mixed		0.112***		0.076***		0.011**		0.035***
		(0.022)		(0.019)		(0.004)		(0.013)
Supportive		0.125***		0.097***		0.002		0.033**
		(0.026)		(0.020)		(0.005)		(0.013)
Constant	0.692***	0.692***	0.295***	0.294***	−0.006*	−0.006*	0.397***	0.397**
	(0.122)	(0.124)	(0.055)	(0.059)	(0.003)	(0.004)	(0.152)	(0.154)
Zipcode Set FEs	✓	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Zipcode Set × Controls	✓	✓	✓	✓	✓	✓	✓	✓
Mean Y in Control	0.481	0.481	0.256	0.256	0.021	0.021	0.219	0.219
p(Support vs. Oppose)		0.954		0.232		0.173		0.173
p(Oppose vs. Mixed)		0.521		0.795		0.898		0.198
p(Support vs. Mixed)		0.556		0.288		0.119		0.875
Observations	551,542	551,542	551,542	551,542	551,542	551,542	551,542	551,542
R^2	0.0005	0.0005	0.001	0.001	0.0003	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, along with two-way interactions. Interactions include comments, reactions and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table C11: The Impact of the Comment Section on On-platform User Engagement (Logistic Regression)

	Dependent variable (Odds Ratios)							
	All Engagement		Post Expansions		Interactions		Link Clicks	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Any comments	1.136***		1.193***		1.131		1.070	
	(0.036)		(0.052)		(0.177)		(0.050)	
Opposing		1.181***		1.191***		1.436**		1.148**
		(0.045)		(0.063)		(0.257)		(0.064)
Mixed		1.128***		1.196***		1.110		1.051
		(0.043)		(0.063)		(0.210)		(0.060)
Supportive		1.098**		1.193***		0.848		1.012
		(0.043)		(0.063)		(0.172)		(0.058)
Constant	0.005***	0.005***	0.002***	0.002***	0.000	0.000	0.004***	0.004***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.00000)	(0.00000)	(0.001)	(0.001)
Zipcode Set FEs	✓	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Zipcode Set × Controls	✓	✓	✓	✓	✓	✓	✓	✓
Mean Y in Control (C)	0.485	0.485	0.254	0.254	0.020	0.020	0.228	0.228
Observations	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015

Notes: Coefficients reported as odds ratios. 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 C9: The Impact of Any Comment on On-platform User Engagement

	All Engagement			Post Expansions			Interactions			Link Clicks		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Any comments	0.065***	0.065***	0.066***	0.049***	0.049**	0.049***	0.003	0.003	0.003	0.016**	0.016	0.016
	(0.015)	(0.025)	(0.024)	(0.013)	(0.019)	(0.019)	(0.003)	(0.004)	(0.004)	(0.008)	(0.014)	(0.014)
Constant	0.566***	0.418***	0.485***	0.282***	0.178***	0.254***	0.034***	0.006	0.020***	0.276***	0.243***	0.228***
	(0.046)	(0.035)	(0.021)	(0.045)	(0.025)	(0.016)	(0.008)	(0.004)	(0.003)	(0.021)	(0.025)	(0.012)
Zipcode Set FEs	✓			✓			✓			✓		
Controls		✓			✓			✓			✓	
Zipcode Set × Controls												
Mean Y in Control (C)	0.485	0.485	0.485	0.254	0.254	0.254	0.020	0.020	0.020	0.228	0.228	0.228
Observations						1,054,015						
R^2	0.0001	0.0002	0.00002	0.0001	0.0002	0.00002	0.00002	0.0001	0.00000	0.0001	0.00004	0.00000

Notes: Standard errors are clustered at the advertisement level (72 ads). The unit of observation is defined at the level of the reach. Columns (1), (4), (7), (10) include zipcode fixed effects only. Columns (2), (5), (8), (11) include gender and age controls only. Columns (3), (6), (9), (12) do not include zipcode fixed effects and controls. Interactions include comments, reactions and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table C10: The Impact of the Comment Section on On-platform User Engagement

	All Engagement			Post Expansions			Interactions			Link Clicks		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Opposing	0.086***	0.087***	0.087***	0.048***	0.048*	0.048*	0.009**	0.009*	0.009*	0.033***	0.034**	0.034**
	(0.021)	(0.032)	(0.031)	(0.017)	(0.025)	(0.025)	(0.004)	(0.005)	(0.005)	(0.011)	(0.016)	(0.017)
Mixed	0.061***	0.062**	0.062**	0.049***	0.049**	0.050**	0.002	0.002	0.002	0.011	0.011	0.012
	(0.017)	(0.029)	(0.028)	(0.013)	(0.020)	(0.020)	(0.003)	(0.004)	(0.004)	(0.011)	(0.018)	(0.018)
Supportive	0.047***	0.047*	0.048*	0.049***	0.049**	0.049**	−0.003	−0.003	−0.003	0.002	0.003	0.003
	(0.017)	(0.028)	(0.028)	(0.015)	(0.023)	(0.022)	(0.003)	(0.004)	(0.004)	(0.008)	(0.014)	(0.014)
Constant	0.566***	0.418***	0.485***	0.282***	0.178***	0.254***	0.034***	0.006	0.020***	0.276***	0.243***	0.228***
	(0.046)	(0.034)	(0.020)	(0.045)	(0.025)	(0.016)	(0.006)	(0.004)	(0.003)	(0.021)	(0.025)	(0.012)
Zipcode Set FEs	✓			✓			✓			✓		
Controls		✓			✓			✓			✓	
Zipcode Set × Controls												
Mean Y in Control (C)	0.485	0.485	0.485	0.254	0.254	0.254	0.020	0.020	0.020	0.228	0.228	0.228
p(Support vs. Oppose)	0.067	0.185	0.187	0.935	0.978	0.964	0.003	0.010	0.010	0.007	0.027	0.027
p(Oppose vs. Mixed)	0.230	0.406	0.396	0.914	0.948	0.944	0.101	0.123	0.121	0.105	0.218	0.217
p(Support vs. Mixed)	0.410	0.587	0.591	0.987	0.968	0.982	0.081	0.144	0.139	0.427	0.585	0.590
Observations						1,054,015						
R^2	0.0001	0.0002	0.00002	0.0001	0.0002	0.00002	0.00003	0.0001	0.00001	0.0001	0.00005	0.00001

Notes: Standard errors are clustered at the advertisement level (72 ads). The unit of observation is defined at the level of the reach. Columns (1), (4), (7), (10) include zipcode fixed effects only. Columns (2), (5), (8), (11) include gender and age controls only. Columns (3), (6), (9), (12) do not include zipcode fixed effects and controls. Interactions include comments, reactions and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table C12: Ad-level Estimates

	Group Mean / (SD)				t–test p–value		
Variable	(1) Control	(2) Supportive	(3) Mixed	(4) Opposing	(1)–(2)	(1)–(3)	(1)–(4)
All Engagement	0.485 (0.091)	0.533 (0.080)	0.547 (0.078)	0.572 (0.098)	0.096	0.032	0.008
Post expansions	0.254 (0.073)	0.303 (0.065)	0.303 (0.046)	0.302 (0.080)	0.036	0.017	0.063
Interactions	0.020 (0.014)	0.017 (0.014)	0.022 (0.009)	0.029 (0.015)	0.506	0.557	0.079
Link Clicks	0.228 (0.050)	0.230 (0.035)	0.239 (0.058)	0.261 (0.048)	0.850	0.543	0.050
Observations	18	18	18	18			

Notes: Each observation corresponds to an ad. Statistics are weighted by ad reach. P-values are based on t-tests using heteroskedasticity-robust standard errors.

Table C13: The Impact of the Comment Section on Page Views

	Dependent variable (as % of total reach):	
	Page Views	
	(1)	(2)
Any comments	0.016*	
	(0.008)	
Opposing		0.019*
		(0.011)
Mixed		0.017
		(0.011)
Supportive		0.011
		(0.011)
Constant	0.347**	0.347***
	(0.135)	(0.134)
Zipcode Set FEs	✓	✓
Controls	✓	✓
Zipcode Set × Controls	✓	✓
Mean Y in Control (C)	0.171	0.171
p(Support vs. Oppose)		0.523
p(Oppose vs. Mixed)		0.880
p(Support vs. Mixed)		0.611
Observations	1,054,015	1,054,015
R^2	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 C14: The Impact of the Comment Section with Black Interaction

	Dependent variable:							
	All Engagement		Post Expansions		Interactions		Link Clicks	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Any comments	0.079***		0.066***		−0.0003		0.015	
	(0.022)		(0.019)		(0.004)		(0.012)	
Any comments × Black	−0.027		−0.035		0.006		0.002	
	(0.029)		(0.024)		(0.005)		(0.016)	
Opposing		0.108***		0.079***		0.005		0.028
		(0.034)		(0.025)		(0.006)		(0.020)
Opposing × Black		−0.041		−0.061*		0.007		0.011
		(0.042)		(0.032)		(0.008)		(0.023)
Mixed		0.070***		0.056***		−0.001		0.014
		(0.026)		(0.018)		(0.005)		(0.018)
Mixed × Black		−0.018		−0.014		0.006		−0.005
		(0.033)		(0.024)		(0.006)		(0.022)
Supportive		0.059**		0.064***		−0.005		0.002
		(0.025)		(0.021)		(0.004)		(0.013)
Supportive × Black		−0.023		−0.030		0.004		0.001
		(0.034)		(0.029)		(0.006)		(0.017)
Constant	0.528***	0.529***	0.145**	0.145**	0.006	0.007	0.380***	0.380***
	(0.118)	(0.113)	(0.063)	(0.064)	(0.010)	(0.009)	(0.070)	(0.066)
Zipcode Set FEs	✓	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Zipcode Set × Controls	✓	✓	✓	✓	✓	✓	✓	✓
Mean Y in Control (C)	0.485	0.485	0.254	0.254	0.020	0.020	0.228	0.228
Observations	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015
R^2	0.0004	0.0004	0.001	0.001	0.0003	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 C15: The Impact of the Comment Section with Age Heterogeneity

	Dependent variable:							
	All Engagement		Post Expansions		Interactions		Link Clicks	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Any comments	0.074***		0.057***		−0.002		0.019	
	(0.022)		(0.016)		(0.005)		(0.015)	
Any comments × Middle Aged	−0.010		−0.018		0.010		0.002	
	(0.029)		(0.023)		(0.007)		(0.018)	
Any comments × Senior	−0.053		0.017		−0.007		−0.055	
	(0.084)		(0.067)		(0.020)		(0.047)	
Opposing		0.085***		0.056***		−0.002		0.034*
		(0.028)		(0.019)		(0.005)		(0.019)
Opposing × Middle Aged		0.001		−0.025		0.018**		0.010
		(0.030)		(0.022)		(0.008)		(0.020)
Opposing × Senior		0.007		0.061		0.017		−0.074
		(0.092)		(0.073)		(0.028)		(0.063)
Mixed		0.062**		0.035*		−0.004		0.029
		(0.027)		(0.020)		(0.006)		(0.019)
Mixed × Middle Aged		0.004		0.020		0.012		−0.024
		(0.041)		(0.029)		(0.009)		(0.027)
Mixed × Senior		−0.044		0.040		−0.006		−0.066
		(0.102)		(0.088)		(0.023)		(0.057)
Supportive		0.076***		0.079***		−0.001		−0.007
		(0.028)		(0.025)		(0.006)		(0.016)
Supportive × Middle Aged		−0.037		−0.050		0.0002		0.021
		(0.036)		(0.034)		(0.009)		(0.020)
Supportive × Senior		−0.119		−0.050		−0.030		−0.025
		(0.096)		(0.083)		(0.023)		(0.048)
Constant	0.445***	0.445***	0.184***	0.184***	0.043***	0.043***	0.240***	0.240***
	(0.063)	(0.060)	(0.044)	(0.045)	(0.016)	(0.016)	(0.041)	(0.037)
Zipcode Set FEs	✓	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Zipcode Set × Controls	✓	✓	✓	✓	✓	✓	✓	✓
Mean Y in Control (C)	0.485	0.485	0.254	0.254	0.020	0.020	0.228	0.228
Observations	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015
R^2	0.0003	0.0003	0.0003	0.0003	0.0001	0.0002	0.0001	0.0001

Notes: Age groups are defined as Youth (ages 18–34), Middle Aged (ages 35–64), and Senior (age 65+), with Youth as the omitted baseline category. Standard errors are clustered at the advertisement level (72 ads). The unit of observation is defined at the level of the reach. Use gender as control. Interactions include comments, reactions and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table C16: Robustness check: Exclude the toxic comments(threshold=0.7)

	Dependent variable:			
	All Engagement	Post Expansions	Interaction	Link Clicks
	(1)	(2)	(3)	(4)
Opposing	0.100***	0.059***	0.012***	0.032**
	(0.024)	(0.021)	(0.004)	(0.014)
Mixed	0.085***	0.066***	0.005	0.017
	(0.019)	(0.014)	(0.004)	(0.015)
Supportive	0.047***	0.049***	−0.003	0.002
	(0.016)	(0.015)	(0.003)	(0.008)
Constant	0.521***	0.144**	0.002	0.378***
	(0.111)	(0.061)	(0.010)	(0.066)
Zipcode Set FEs	✓	✓	✓	✓
Controls	✓	✓	✓	✓
Zipcode Set × Controls	✓	✓	✓	✓
Mean Y in Control (C)	0.485	0.254	0.020	0.228
p(Support vs. Oppose)	0.024	0.616	0.001	0.039
p(Oppose vs. Mixed)	0.519	0.723	0.164	0.392
p(Support vs. Mixed)	0.044	0.215	0.022	0.332
Observations	878,687	878,687	878,687	878,687
R^2	0.0005	0.001	0.0003	0.0002

Notes: Standard errors are clustered at the advertisement level (60 ads). We excluded posts whose comments had toxicity scores above 0.7. 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. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

D Additional Results for the Survey Experiment

Figure D1: Frequency of Reading or Checking Comments on Social Media

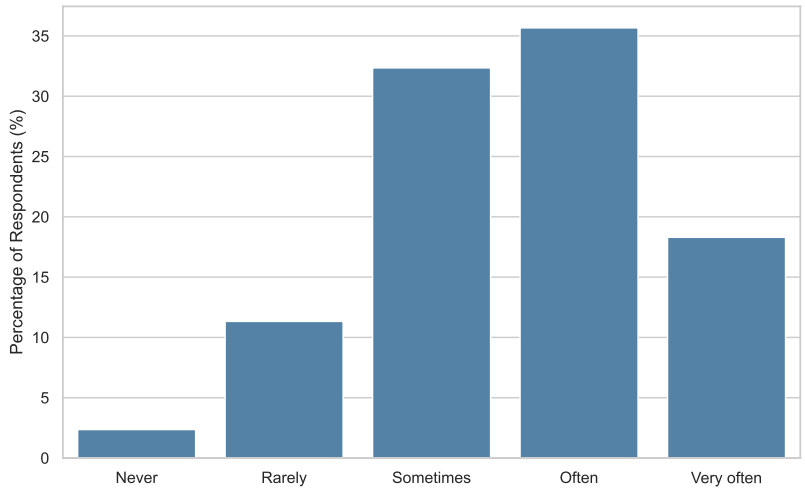
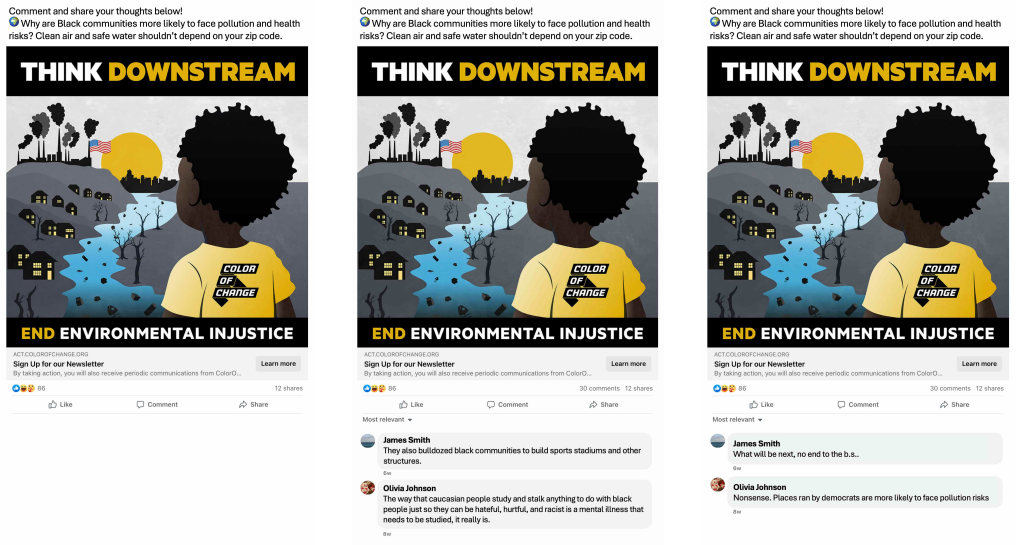


Figure D2: Survey Experiment Stimuli



(a) Control (b) Supportive (c) Opposing

D.1 Survey Instrument

Screening

Welcome! We have a few quick questions before we start.

This should take no more than 20 seconds. We will let you know if you are eligible for the study and provide details on participation payments.

1. Do you live in the United States?
   [Yes / No]

2. What is your age? *[number entry box]*

3. What is your gender?
   [Male / Female / Non-binary / other / I prefer not to answer]

4. What is your race?
   [American Indian / Alaska Native / Asian / Pacific Islander / Black / African American / White / Other / Mixed Race]

5. Are you of Hispanic or Latino origin?
   [Yes / No]

[Continue if US resident, aged 18–64.]

Consent
[Consent form]

I agree to participate, and I promise to read the questions carefully and answer honestly

I do not agree to participate, or I cannot promise to read the questions carefully and answer honestly

Baseline Opinions

Thanks for agreeing to participate! We value your opinions and are interested in hearing what you think.

1. In politics, as of today, do you consider yourself a Republican, a Democrat or an independent?
   [Republican / Democrat / Independent / Other]

2. We hear a lot of talk these days about liberals and conservatives. Which of the following best describe your political view?
   [Very liberal / Liberal / Moderate; middle of the road / Conservative / Very conservative / Haven't thought much about this/don't know]

3. When it comes to giving African Americans equal rights with white Americans, do you think our country has...
   [Gone too far / Not gone far enough / Been about right]

   [Randomize order of "Gone too far" and "Not gone far enough."]

4. Do you believe that the increased public attention to the history of slavery and racism is generally good or bad for our society?
   [Very good / Somewhat good / Neither good nor bad / Somewhat bad / Very bad]

5. **On the issues of race and racism,** my position is...
   [Very progressive / Progressive / Moderate; middle of the road / Conservative / Very conservative / Haven't thought much about this / don't know]

   [Randomly flip the order.]

6. Now we'd like to know your best guess about how people in a **representative sample of adults in the United States** answered this question in 2025.

   "When it comes to giving African Americans equal rights with white Americans, do you think our country has..."

   Please estimate what percentage of respondents chose each response. Your answers should add up to 100%.

[Constant sum question; entries for: ___% Gone too far / ___% Not gone far enough / ___% Been about right; must sum to 100.]

Social Media Use

1. How much time do you spend on social media (e.g., Facebook, Instagram, TikTok, YouTube) excluding Messenger and WhatsApp, on an average day?
   [Less than 5 minutes a day / Between 5 and 30 minutes a day / Between 30 and 60 minutes a day / Between 1 and 2 hours / Between 2 and 4 hours / More than 4 hours]

2. How often do you read or check comments on social media?
   [Never / Rarely / Sometimes / Often / Very often]

Additional Demographics

1. In what zip code do you currently live? *[text entry box; validation: US ZIP code]*

2. What is the highest degree or level of schooling that you have completed?
   [Less than a high school diploma / High school diploma or equivalent (for example: GED) / Some college but no degree / Associate's degree / Bachelor's degree / Graduate degree (for example: MA, MBA, JD, PhD)]

Attention Check

1. In order to facilitate our research, we are interested in knowing certain factors about you. Specifically, we are interested in whether you actually take the time to read the instructions; if not, then the data we collect based on your responses will be invalid. So, in order to demonstrate that you have read the instructions, please ignore the next question, and simply write "I read the instructions" in the "Any comments?" box below. Thank you very much.

   What is your marital status?
   [Single / Married / Other]

   Any comments? *[text box]*

Intervention

[Participants randomized into 3 groups: No comment, Supportive, Opposing.]

[3 posts of the same treatment type are shown, varying order of topics: education, environment, and police.]

We are interested in your reactions to social media posts about racial justice. You will be shown three posts from Color of Change, a leading US-based racial justice advocacy organization.

[Next page.]

Here is a post from Color of Change:

[Screenshot shown.]

[Next page.]

Now we will ask you some questions about this post.

[Post shown again.]

1. Which of the following emojis would you react with if you saw this post on Facebook?
   [Like / Love / Care / Haha / Wow / Sad / Angry / I would not react]

2. Would you comment on this post if you saw it on Facebook?
   [Yes / No]

3. *[If yes to previous question]* What comment would you make? *[open text]*

4. Would you click on this post to visit the website if you saw it on Facebook?
   [Yes / No]

[Screenshots of remaining posts shown; questions above repeated for each post.]

Newsletter

1. Would you like to sign up for the Color of Change newsletter?
   [Yes / No]

Donation

1. You have been automatically enrolled in a lottery to win up to $100. If you win, you have the option to donate some or all of your winnings to Color of Change.

   The payment will be made to you as a bonus, so no further action is required on your part. If you are one of the lottery winners, you will be paid, in addition to your participation payment, $100 minus the amount you donated. We will directly pay your desired donation amount to Color of Change.

   How much, if any, would you be willing to donate to Color of Change in case you won $100?
   [number entry]

Post-Exposure Opinion

1. How would you describe your overall opinion of Color of Change?
   [Very unfavorable / Somewhat unfavorable / Neutral / Somewhat favorable / Very favorable]

2. How would you describe your overall opinion of Black Lives Matter?
   [Very unfavorable / Somewhat unfavorable / Neutral / Somewhat favorable / Very favorable]

3. How willing would you be to discuss political issues with someone who has **progressive views** on racial issues?
   [Very unwilling / Somewhat unwilling / Neither willing nor unwilling / Somewhat willing / Very willing]

4. How willing would you be to discuss political issues with someone who has **conservative views** on racial issues?
   [Very unwilling / Somewhat unwilling / Neither willing nor unwilling / Somewhat willing / Very willing]

5. People differ in how important they consider different racial justice issues.

   How important is each of the following issues to you personally?
   [Matrix; rows: Voter suppression and voting rights / Criminal-justice reform (e.g., policing, sentencing, incarceration) / Education equity (e.g., school funding, achievement gaps) / Environmental justice (e.g., pollution exposure, clean air/water access) / Technology fairness (e.g., algorithmic bias, digital discrimination); 5-point scale: Not at all important / Slightly important / Moderately important / Very important / Extremely important]

6. Please tell us the extent to which you agree or disagree with the statement below.

   It's really a matter of some people not trying hard enough. Black people could be just as well off as white people if they would only try harder.
   [Strongly disagree / Disagree / Slightly disagree / Neither agree nor disagree / Slightly agree / Agree / Strongly agree]

7. What do you think is the most important issue facing Black people today? *[open text]*

Opinion about Others

Now we are going to ask you to make a guess.

You can earn a Guess Bonus of up to $1 based on the accuracy of your estimate. We will compare your estimate to the actual percentages observed in this study. The formula we use rewards you more when your estimate is closer to the true value. **The closer your guess is to the correct percentage, the larger your bonus.**

Your best strategy is to give your honest, best estimate.

(You do *not* need to know the formula to earn the bonus.)
[Next page.]

1. What percentage of participants in this U.S. adult sample do you think **agreed or strongly agreed** with the following statement?

   "It's really a matter of some people not trying hard enough. Black people could be just as well off as white people if they would only try harder."

   Please enter a number between **0 and 100**.
   You can earn a **bonus** based on how close your estimate is to the true value.

   ___**%** *[number entry]*

   [Pop-up window with quadratic scoring rule: Guess Bonus = \$1 − ((Your Answer − True Value)/100)^2. Small errors reduce your bonus slightly; larger errors reduce it much more. If your estimate is exactly correct, you will earn \$1. Your best strategy is to give your honest, best estimate.]

[Environment post shown again.]

2. How thought-provoking do you find this post?
   [Not at all / A little / Somewhat / Very / Extremely]

Opinion about Comments in Post [treatment arms only]
Now consider the **comments** below the following social media post.
[Environment post with comments shown again.]

1. To what extent do the comments make you feel:
   [Matrix; rows: Angry / Annoyed; 5-point scale: Not at all / A little / Somewhat / Very / Extremely]

2. How thought-provoking do you find these comments?
   [Not at all / A little / Somewhat / Very / Extremely]

3. To what extent do the comments make you feel:
   [Matrix; rows: Curious about the topic / Curious about the organization; 5-point scale: Not at all / A little / Somewhat / Very / Extremely]

4. How representative do you think these comments are of what people generally think about this issue?
   [Not at all representative / Slightly representative / Moderately representative / Very representative / Extremely representative]

Opinion about Comments in Post [all arms]
[Environment post shown again.]

1. How interested would you be in opening the comment section to see more comments on this post?
   [Not at all interested / Slightly interested / Moderately interested / Very interested / Extremely interested]

2. In your opinion, which group is more likely to comment on this post?
   [Men / Women / Men and women are equally likely / Not sure]

3. In your opinion, which group is more likely to comment on this post?
   [Progressives / Conservatives / Both groups are equally likely / Not sure]

Opinion about Comments in General

1. What are the main reasons you read or look at comments on social media? *[open text]*

Newsletter Sign-Up
[Shown only to participants who answered Yes to the newsletter question.]
You said that you like to sign up for the Color of Change newsletter.

1. Please enter your email address below. Your email will only be shared with Color of Change, and only for the purpose of subscribing you to their newsletter.
   [text entry box]

AI Use

1. Did you use AI at all to help you fill out this survey?
   [Yes / No]

2. *[If yes]* What question did you use AI to help you answer? *[open text]*

Final Feedback
Thank you! We really appreciate you for participating in this research!
Please let us know if you have any other feedback.
Make sure to click the next arrow to submit your survey responses.
[text box]