Gender Bias, Feedback, and Productivity

Marita Freimane*

November 17, 2025

Job Market Paper

Click here for the latest version

Abstract

I study how gender-biased feedback affects worker productivity in an online labor market. I focus on content creators, an increasingly important group of online workers, and show that female creators received more negative feedback than male creators for comparable content. This gap disappeared after YouTube removed public dislike counts. Using detailed data on videos and channels and a difference-in-differences design, I find that the reduction in negative feedback led to a lasting increase in both the productivity of female creators and demand for their content. After the change, women produced 8.4 percent more videos and saw a 15.5 percent larger increase in demand than men. The decline in negative feedback was driven by videos that had extremely high dislike counts, consistent with YouTube's goal of reducing harassment through "dislike attacks". A placebo test suggests that the effects arise from the drop in dislikes rather than their removal from public view, and I find little evidence of spillovers to comments.

Keywords: platform design, user-generated content, gender bias, labor productivity

JEL Classification: J24, J71, L82, L86

*KU Leuven and Uni Zurich, marita.freimane@business.uzh.ch. I thank Luis Aguiar, Anahid Bauer, George Beknazar-Yuzbashev, Laura Bodreau, Karol Borowiecki, Grazia Cecere, Dante Donati, Ricard Gil, Gautam Gowrisankaran, Jan Feld, Reinhold Kesler, Lavinia Kinne, Uli Laitenberger, Leonardo Madio, Suresh Naidu, Mattia Nardotto, Andrea Prat, Dominik Rehse, Imke Reimers, Michelangelo Rossi, Andrey Simonov, Pietro Tebaldi, Tommaso Valletti, and Joel Waldfogel, as well as seminar participants at Télécom Paris, ZEW Mannheim, the joint Digital Economics Seminar, BSE Summer Forum Workshop in Digital Economics, ZEW ICT Conference, PhD Consortium at Platform Strategy Research Symposium, EARIE, and doctoral seminars at the University of Zurich and Columbia University for valuable discussions and comments. Yanxi Hou provided excellent research assistance and Socialblade.com provided data on content demand. I acknowledge financial support from the Swiss National Science Foundation Grant Number 207682.

1 Introduction

Over the past century, women have entered the labor market in growing numbers, narrowing gender gaps in both employment and earnings (Goldin 2014). This shift has contributed substantially to aggregate productivity growth (Hsieh et al. 2019). In recent years, however, progress has slowed, and significant gender gaps remain (Blair and Posmanick 2023). A growing body of research examines the drivers of these remaining disparities, including differences in how men and women are treated at work. These include unequal recognition of contributions to group work (Sarsons 2017; Sarsons et al. 2021); biased subjective assessments of women's potential (Benson et al. 2024); differential assignment to *non-promotable* versus *promotable* tasks (Babcock et al. 2017; Chu et al. 2022; Bircan et al. 2024); and greater exposure of women to workplace harassment (Folke and Rickne 2022).

In this paper, I examine biased feedback as another mechanism through which men and women experience differential treatment in the workplace. While gender differences in feedback are difficult to observe in traditional offline settings, they have been documented in online environments (Botelho and Gertsberg 2022; Aguiar 2024).[1] However, little is known about the consequences of biased feedback for workers' productivity, largely because it is difficult to simultaneously observe both feedback and output in traditional work environments. I address this gap by studying online content creation, an important and rapidly growing segment of the labor market where both feedback and output are directly observable and quantifiable.[2] Content creators are directly exposed to consumer feedback and independently choose how much to work, making this an ideal setting to examine the impact of biased feedback on labor supply, performance, and potential earnings.

My empirical analysis reveals three main findings. First, I document a significant gender gap in negative feedback on YouTube, a major platform for user-generated content. This gap persists even after controlling for gender differences in the types of content produced. Second, I provide causal evidence that this gap is eliminated when the platform redesigns its reputation system. Third, I show that eliminating the gender feedback gap leads to an increase in both content supply and consumer demand. To identify these effects, I exploit a

[1] Related work on gender bias in evaluations finds disparities in diverse settings, including orchestra auditions (Goldin and Rouse 2000) and student evaluations of female faculty (Mengel et al. 2019).

[2] There are an estimated 50 million content creators worldwide, including about 2 million who work full time. https://web.archive.org/web/20250630174656/https://signalfire.com/blog/creator-economy/

design change that altered the visibility of negative feedback on the platform. In November 2021, YouTube suddenly removed the public display of dislike counts for videos. Using a unique dataset constructed from YouTube API data, I study how users disliking behavior changed when dislikes were no longer publicly observable. This variation allows me to estimate how much of the negative feedback was driven by the visibility of prior dislikes and, under certain assumptions, how much stemmed from "dislike attacks", where many users dislike a video in a short period for reasons unrelated to its content quality. I then quantify the downstream effects of such dislike attacks on content supply and audience demand.

A raw comparison of video-level dislike counts before and after the design change reveals a large gender gap in negative feedback prior to the intervention. Before dislikes were made private, videos by female content creators received 43 percent more dislikes, on average, than those by male creators. After the removal of public dislike counts, women experienced a 57 percent decline in negative feedback. Difference-in-differences estimates show that this decline reflects a causal effect of the design change: relative to men, women saw a 21 percent larger decrease in dislikes following the intervention. Next, I show that this result is not driven by gender differences in content type. To account for differences in the kinds of videos men and women produce, I train a Lasso-based machine learning model to predict how many dislikes a video would have received had it been created by a man. Including these predicted counterfactuals in my difference-in-differences regressions, I estimate a 24 percent reduction in dislikes for women, conditional on the "dislikability" the content would have had if male-authored. The larger estimate, though not statistically distinguishable, suggests that women may already have self-selected into content categories that are less exposed to excessive negative feedback. Finally, I estimate the gender gap in feedback in the pre- and post-treatment periods separately. The results show that the platform design change closed the gender gap in negative feedback.

What are the effects of biased feedback on content supply and consumer demand? To study the causal effect of reducing excessive negative feedback on content supply and demand, I estimate a difference-in-differences model. Motivated by the larger impact of public dislike counts on women than men, I compare changes in women's content produc-

tion and viewership to those for men before and after the platform change.[3] The estimates indicate that, following the removal of public dislike counts, women increased their content supply by 8.4 percent. For the median content creator in the sample, this corresponds to 0.4 additional videos per month. On the demand side, viewership of women's content rose by 15.5 percent, nearly twice the increase in supply, implying higher average views per video. A back-of-the-envelope calculation implies that the median female creator earned up to $3,100 more per month after the design change.[4]

Next, I investigate the mechanisms behind the observed effects and test whether they can plausibly be attributed to the reduction in dislike attacks. I begin by examining whether the effect is concentrated in the upper tail of the dislike distribution, consistent with targeted negative feedback. I assess the validity of this interpretation using a placebo test in countries without a pre-existing gender gap in dislikes. Finally, I examine two additional outcomes: spillovers to comment sentiment and changes in video quality.

First, if the reduction in negative feedback is primarily driven by a decline in excessive negative reactions, such as dislike attacks, then we would expect the treatment effect to be concentrated in the upper tail of the dislike distribution. In contrast, if the design change uniformly reduced disliking behavior, we would expect a similar treatment effect across the distribution. To test this, I estimate quantile regressions by gender. The results show that across most of the distribution, women experience a larger reduction in dislikes than men, except in the bottom tail. The gender gap is especially pronounced in the upper tail: at the 90th percentile, women experience a 35 percent greater decrease in dislikes than men. This pattern is consistent with YouTube's stated goal of reducing harassment through "dislike attacks".[5] "Dislike attacks" occur when large numbers of users dislike a video irrespective of its content quality, whether due to organized campaigns or herding behavior triggered by many existing dislikes.[6]

Second, I show that the observed effects on supply and demand are not driven by broader platform-wide shifts unrelated to the removal of public dislike counts, nor are they

[3] This compares a more treated group (female creators) to a less treated group (male creators). The identifying assumption is that the treatment effect has the same direction across groups, implying that the estimate represents a lower bound. I provide evidence from a placebo test to support this assumption

[4] Based on revenue estimates of 0.25˘4.00 per 1,000 views obtained from Socialblade.com.

[5] https://web.archive.org/web/20250904054805/https://blog.youtube/news-and-events/update-to-youtube/

[6] https://web.archive.org/web/20230820065708/https://www.urbandictionary.com/define.php?term=dislike-bombed

explained simply by making dislikes private. To assess this, I conduct a placebo test using data from Korea and Japan, two markets that were exposed to the same platform design change but exhibited no gender gap in dislikes prior to the intervention. In this setting, where the design change did not differentially affect feedback, I find no differential impact on supply or demand. This supports the interpretation that the treatment effects observed elsewhere are driven by the reduction in visible biased feedback.

Third, I examine whether negative feedback previously expressed through dislikes may have spilled over to comments, raising the possibility that the design change displaced toxicity rather than reducing it. Specifically, I use natural language processing tools to compare comment sentiment before and after the design change, assessing whether comments became more negative or more toxic once dislikes were no longer publicly visible. The results indicate limited spillover effects. If anything, the most negative comments became slightly less negative following the removal of public dislike counts.

Fourth, I examine whether the platform design change affected video quality. Measuring quality is inherently difficult, particularly in this context, where common proxies such as likes, dislikes, and views are themselves directly influenced by the treatment. While direct measurement is challenging, several pieces of evidence suggest that video quality did not decline, and may even have improved. First, the estimated effect on demand exceeds the effect on content supply, implying an increase in average views per video. Second, comment sentiment improves slightly after the design change. Finally, I present evidence based on predicted dislikes. Using a model trained on men's pre-treatment dislikes to estimate the counterfactual quality of women's videos, I find that predicted dislikes decrease for women following the design change. Because these predicted values are not affected by gender bias, they provide a plausibly exogenous proxy for video quality.

Related literature. This paper contributes to four strands of literature in economics. First, I contribute to the literature on gender disparities in the labor market by documenting how biased feedback affects the productivity of female workers. I focus on online content creators, a growing group of self-employed workers whose exposure to feedback and output levels are both directly observable and quantifiable. I show that gender-biased feedback reduces female content creators' supply of content and the demand for their work. My findings build on prior evidence of gendered treatment in traditional workplace settings.

For example, women are often rated lower than men in subjective performance assessments (Benson et al. 2024), less likely to be assigned promotable tasks (Bircan et al. 2024; Babcock et al. 2017), and receive less credit for group work under uncertainty (Sarsons 2017; Sarsons et al. 2021). These forms of differential treatment contribute to lower promotion rates and talent misallocation. Gender discrimination can also originate from customers: Kelley et al. (2024) show that randomized female names on an online travel agency platform lead to fewer purchases on both extensive and intensive margins. While prior work documents various forms of gender-based bias in offline workplaces, little is known about how such differential treatment affects productivity—largely because feedback and output are difficult to observe in real time. My setting overcomes these limitations: because online content creators' exposure to feedback and output decisions are both directly observable, I am able to causally estimate the effect of biased feedback on labor supply and economic outcomes for women.

Second, I contribute to the literature on the economic costs of toxic workplace behavior. I show that exposure to excessive negative feedback—specifically, through so-called "dislike attacks"—reduces content creator productivity, even when the feedback falls below legal thresholds for harassment or discrimination. This demonstrates that seemingly minor forms of toxic behavior on digital platforms can have real effects on labor supply. While professional women, including academic economists, are disproportionately targeted by online toxicity (Ederer et al. 2025; Wu 2020), establishing a causal link to output is often empirically difficult. My findings complement recent work documenting how harassment and discrimination reduce output and distort labor market outcomes. Caselli et al. (2023) find that African football players improve performance when games are played without spectators, especially those previously targeted by racial abuse. Karimli (2023) shows that reducing workplace harassment via NDA restrictions increases patenting among inventors. Other studies show that harassment contributes to occupational gender segregation and wage penalties, as workers sort into lower-paying jobs to avoid toxic environments (Folke and Rickne 2022; Batut et al. 2021). Gender-based discrimination also lowers morale and effort: Gagnon et al. (2025) find that female workers reduce labor supply when paid less than male counterparts. Finally, power asymmetries shape firm responses: Adams-Prassl et al. (2024) show that men face fewer consequences for workplace violence in male-led

firms, which reduces female retention and hiring, while perpetrators are more likely to exit female-led firms.

Third, I contribute to the literature on the production of user-generated content (UGC) by showing that toxicity—specifically, excess negative feedback—reduces both content supply and demand. Unlike most prior studies, which examine platforms where users are both content producers and consumers, I study a setting in which these roles are distinct. This separation allows me to isolate how toxicity affects creators' labor supply and how it shapes consumer demand for content. Recent work shows that toxicity can shape overall participation, but it focuses on platforms where users are both content creators and consumers—implying different incentive structures and feedback dynamics than in my setting. In field experiments, Jiménez-Durán (2023) show that reporting toxic tweets increases participation among targeted users, while Beknazar-Yuzbashev et al. (2025) find that hiding toxic content reduces both content consumption and ad engagement. These studies suggest a trade-off between protecting targeted users and maximizing platform-wide engagement. Related theoretical work highlights how ad-financed business models can disincentivize content moderation when toxicity drives user engagement (Madio and Quinn 2025; Beknazar-Yuzbashev et al. 2024; Liu et al. 2022). More broadly, my findings complement a growing literature on the drivers of UGC production. Prior work has shown that both non-monetary (likes, badges, comments) and monetary incentives (e.g., ad revenue) affect content creation across platforms including Facebook, Twitter, Reddit, and YouTube (Eckles et al. 2016; Mummalaneni et al. 2022; Burtch et al. 2022; Srinivasan 2023; Zeng et al. 2023). My results suggest that negative feedback operates in the opposite direction—deterring content production and consumption, even in the absence of financial penalties.

Fourth, I contribute to the literature on online reputation systems by providing causal evidence that gender bias in user feedback reduces content supply and demand—and that platform design can mitigate this bias. Prior work documents gender disparities in review settings, including Yelp (Botelho and Gertsberg 2022), IMDb and Rotten Tomatoes (Aguiar 2024), and Stack Exchange (Bohren et al. 2019). The first two studies find that verification mechanisms—such as elite reviewer status or verified user reviews—can partially mitigate gender bias. In contrast, Bohren et al. (2019) point to belief-based discrimina-

tion: they show that women face significant initial bias in evaluations when they lack a track record, but that this bias reverses after a sequence of positive reviews, as evaluators update their beliefs about women's abilities based on observed performance. My findings complement this work by showing how visibility-driven biases, manifested through "dislike attacks", can be amplified through feedback cascades and reduced through interface design. In doing so, I also contribute to the literature on herding and information cascades in sequential user feedback (Bikhchandani et al. 1992; Banerjee 1992; Lee et al. 2015).

The rest of the paper is organized as follows. In the second chapter I provide background information on the content creator industry and YouTube, the platform I study. In the third chapter I introduce and discuss the platform design change. In the fourth chapter I describe the data I use in my empirical analysis. In the fifth chapter I introduce my empirical framework and in the sixth chapter I present the results. In the seventh chapter I investigate the mechanisms. In the eighth chapter I conclude.

2 Background

2.1 Content Creator Economy

Content creators are individuals who produce and share original content—such as videos, podcasts, blogs, social media posts, or livestreams—primarily on digital platforms. Industry estimates value the creator economy at $250 billion,[7] and suggest that approximately 50 million individuals worldwide are engaged in content creation.[8][9] Of these, around 2 million, or roughly 4 percent, are considered full-time, professional creators for whom content creation serves as their primary occupation.

The creator industry exhibits characteristics of a superstar market, with a small number of highly successful creators accounting for a disproportionate share of audience attention

[7] This estimate, provided by Goldman Sachs, includes revenues from brand deals, affiliate links, courses, ad-sharing, and launching personal brands. https://www.goldmansachs.com/insights/articles/the-creator-economy-could-approach-half-a-trillion-dollars-by-2027.html. The creator industry is typically not tracked as a distinct sector in official labor statistics. https://web.archive.org/web/20240306223916/https://www.washingtonpost.com/technology/2023/10/26/creator-economy-influencers-youtubers-social-media/

[8] https://web.archive.org/web/20250720094555/https://www.signalfire.com/blog/creator-economy

[9] https://web.archive.org/web/20250720094555/https://www.signalfire.com/blog/creator-economy

and income, while a long tail of creators earn little and reach relatively small audiences. Revenue sources differ by creator scale. Ad revenue sharing programs offered by platforms are available to a wider range of creators and allow both large and small channels to monetize their content based on viewership and engagement. In contrast, influencer marketing, including brand sponsorships and affiliate links, is primarily concentrated among high-profile creators with large audiences.

According to industry estimates, YouTube hosts approximately 1 million full-time creators, followed by 500,000 on Instagram, 300,000 on Twitch, and 200,000 across other platforms.[10]

2.2 YouTube

YouTube is the second most visited website in the world after Google Search, with over 30 billion visits per month and average session durations of around 20 minutes: longer than most other major platforms.[11] It plays a central role in the content creator economy, hosting approximately 1 million professional and 12 million amateur creators.[12] According to company estimates, YouTube supported over 800,000 full-time equivalent jobs across six major economies in 2020: Australia, Brazil, Canada, Japan, South Korea, and the United States.[13]

The platform's content creator base is predominantly male: approximately 75 percent of major YouTube creators are estimated to be men.[14] In contrast, the consumer base is more balanced, with around 60 percent of users estimated to be male.[15] YouTube's audience skews younger, more educated, and higher-income compared to the general population.[16]

[10] Estimates for amateur content creators suggest approximately 30 million on Instagram, 12 million on YouTube, 2.7 million on Twitch, and 2 million on other platforms. While TikTok is not included in these figures, other sources suggest it supports around 1.5 million creators. https://web.archive.org/web/20250805010730/https://explodingtopics.com/blog/tiktok-creator-stats

[11] Website rankings and traffic data from SimilarWeb, July 2024. https://web.archive.org/web/20250905233000/https://www.similarweb.com/top-websites/ ; YouTube-specific traffic: https://web.archive.org/web/20240419220428/https://www.similarweb.com/website/youtube.com/

[12] Professional creators are defined as having more than 10,000 subscribers; amateur creators have between 100 and 10,000. https://web.archive.org/web/20250720094555/https://www.signalfire.com/blog/creator-economy

[13] https://web.archive.org/web/20250812080921/https://blog.youtube/inside-youtube/letter-susan-our-2022-priorities/

[14] Own calculations based on the dataset used in this study.

[15] https://web.archive.org/web/20240419220428/https://www.similarweb.com/website/youtube.com/

[16] https://web.archive.org/web/20250907103225/https://backlinko.com/youtube-users

Most consumption happens on mobile devices (63 percent), followed by television (14 percent) and desktop computers (12 percent).[17]

Advertising is YouTube's primary source of revenue.[18] In 2023, the platform generated $31.5 billion in advertising revenue (Alphabet Inc. 2023). Creators who meet eligibility thresholds—at least 1,000 subscribers and 4,000 watch hours over the past year (or 10 million Shorts views in 90 days)—can apply to join the YouTube Partner Program, which allows them to earn a share of the platform's ad revenue.[19] While YouTube has introduced additional monetization streams such as subscriptions, fan memberships, and different types of digital goods[20] advertising remains the dominant revenue source. Creators typically receive 55 percent of the revenue from ads shown alongside their content,[21] and monetization is thus closely tied to view counts.[22].

YouTube increasingly competes not only with social platforms like TikTok and Instagram, but also with streaming services such as Netflix. Long average visit durations, high TV-based viewership, and consumption of longer-form content all point to a shift toward passive, television-style viewing. Meanwhile, the digital advertising market remains opaque, with limited public data on ad pricing. Nevertheless, recent reporting suggests that ads targeting women are consistently more expensive than those targeting men, reflecting gender-based differences in advertiser demand (Mehrjoo et al. 2024).

2.3 Reputation Systems and Misuse of Online Feedback Tools

Digital platforms, including YouTube, rely heavily on user feedback tools to infer content quality, guide recommendation systems, and allocate consumer attention. However, these tools can be misused in ways that distort the perceived quality. One prominent form of misuse is large-scale negative feedback events, often referred to as "review bombs", in which many users leave negative ratings or reviews in a short period, often for reasons unrelated to content quality. Such patterns can arise through targeted attacks or through herding

[17] https://web.archive.org/web/20250907103225/https://backlinko.com/youtube-users

[18] https://www.youtube.com/howyoutubeworks/our-commitments/sharing-revenue/

[19] https://web.archive.org/web/20250901001608/https://support.google.com/youtube/answer/72851?hl=en

[20] See: https://web.archive.org/web/20250814185150/https://blog.youtube/news-and-events/10-ways-monetize-youtube/

[21] https://web.archive.org/web/20250904003425/https://blog.hootsuite.com/how-much-does-youtube-pay-per-view/

[22] A view is counted after at least 30 seconds of watch time (or the full video, if shorter)

and bandwagon behavior when users react to existing visible feedback.

Platforms have adopted various strategies to limit their effects. Goodreads imposes temporary limits on reviews during periods of suspicious activity.[23] IMDb uses weighted rating schemes and displays disclaimers when abnormal patterns are detected.[24] Steam provides full histograms of reviews over time, allowing users to detect potential manipulation themselves.[25]

These examples illustrate a broader challenge platforms face: managing large-scale, low-cost negative feedback that, when misused, can distort content signals and disproportionately affect certain creators.

3 Platform Design Change

Like other aggregator platforms, YouTube relies on engagement metrics to organize content and allocate consumer attention. Distortions in user feedback can therefore affect viewership and subsequently both platform and creator revenue.

Citing misuse of the dislike button in a manner similar to "review bombs", in what are commonly known as "dislike bombs" or "dislike attacks", where a large number of users would dislike a video in a short period, YouTube removed public dislike counts from its platform on November 10, 2021. YouTube also published results from an internal A/B test showing that dislike attacks declined when the dislike count was hidden, as a justification for the design change.[26]

While public dislike counts were removed from the YouTube website on November 10, they remained accessible via the YouTube API for an additional 34 days, until December 13. I exploit this limited period, during which users could no longer observe dislike counts, but I could still access them through the API, to identify how disliking behavior changed in response to the removal of visible feedback. Importantly, users retained the ability to

[23] https://www.theguardian.com/books/2023/dec/18/goodreads-review-bombing , https://www.nytimes.com/2023/06/26/books/goodreads-review-bombing.html , https://www.goodreads.com/review/guidelines

[24] https://help.imdb.com/article/imdb/track-movies-tv/weighted-average-ratings/GWT2DSBYVT2F25SK#

[25] https://steamcommunity.com/games/593110/announcements/detail/1448326897426987372

[26] https://web.archive.org/web/20250804185509/https://support.google.com/youtube/thread/134791097/update-to-youtube-dislike-counts

click the dislike button even after the count was hidden (Figure 1).

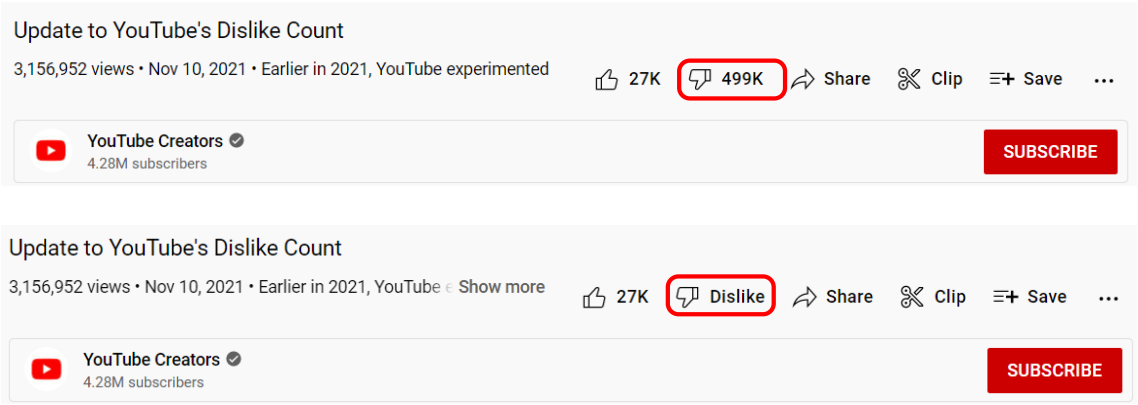


Figure 1: Above: dislikes still visible. Below: dislikes no longer visible

3.1 Incentives to Dislike with Public Dislike Counts

In addition to consuming video content, consumers can engage with a video by clicking the like or the dislike button, or by leaving one or more comments.[27] Consumers have several incentives to leave feedback on YouTube videos. First, feedback on content consumed informs YouTube's algorithm for future *personalized* recommendations.[28] Second, viewers may want to inform content creators about their preferences, to enable them to improve their content in the future. Third, viewers may want to inform other potential viewers about the quality of the content. Other viewers can then use this feedback to themselves assess their interest in the video. Fourth, viewers can choose to engage in dislike attacks described above.

3.2 Incentives to Dislike With Private Dislike Counts

The removal of public dislike counts alters the incentives for disliking videos. Some motivations remain unchanged, as they do not depend on the visibility of dislike counts—such as disliking a video to influence personalized recommendations or to provide feedback to creators, who still have access to dislike information.

However, the design change removes the incentive for consumers to inform others about a video's quality, as dislike counts are no longer visible. Importantly, it also prevents dislike

[27] When signed into their Google account, a consumer can leave one like *or* dislike, and multiple comments on a video.

[28] https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/

attacks, because consumers can no longer observe how many others have disliked a video. Previously, visible dislike counts served as an implicit coordination device that enabled herding behavior. Since dislike attacks involve unusually large numbers of dislikes, the design change is therefore likely to have its strongest effects in the right tail of the dislike distribution.

From a theoretical perspective, engagement in harassment through dislike attacks could be understood as a form of *self-signalling* (Bodner and Prelec 2003). Self-signalling describes the signalling of one's own identity, values, or beliefs to ones-self. This mechanism can rationalize engagement in specific behaviors, such as dislike attacks, even though one's personal contribution can not be identified by others. In this specific case, the utility of self-signalling also depends on the number of other people engaging in the dislike attack. The incentives to dislike a video are thus higher when many other people do so too, leading consumers to engage in herding.

3.3 Alternative Sources of Dislike Bias

Biased feedback can arise from unconscious biases rather than deliberate actions, as users may unknowingly exhibit these biases when using the platform's reputation tools. Psychological evidence indicates that much discrimination is subconscious and unintentional; Bertrand et al. (2005) refer to this as *implicit* discrimination, noting that factors like time pressure, stress, and inattentiveness can exacerbate these biases. In the context of YouTube videos, inattentiveness is likely prevalent, as users are unlikely devote much of their mental resources to the decision of whether or not to dislike a video. Botelho and Gertsberg (2022) document that Yelp reviewers rate restaurants 0.13 stars *lower* when served by a woman, but awarding an "Elite" reviewer status reduces this gap by over half, mainly by decreasing 1-star ratings. They hypothesize that this occurs due to a heightened sense of responsibility among reviewers, suggesting that the bias against women may be unintentional and that increasing cognitive resources for feedback could help mitigate it. Similarly, restricting negative feedback to a more costly form of feedback — comments — on YouTube may encourage consumers to reflect more before providing their assessments.

3.4 Effects on Supply and Demand

Biased reviews likely influence both demand and supply of content. Eliminating these might reduce the cognitive burden on content creators who were previously affected,[29] and lead to an increase in the amount of content they produce by reducing the perceived costs of video production. In turn, the removal of public dislike counts is unlikely to affect the content supply by creators who were not subject to dislike attacks and other sources of bias in dislikes before the design change.

The removal of public dislike counts may affect demand for content because it alters the signal that consumers receive about video quality. Potential consumers often rely on engagement metrics — likes, dislikes, comments, and views — to assess the perceived quality of a video and determine whether a video aligns with their tastes and is worth watching. The platform design change affects the signal in two ways. First, it might increase demand by removing downward-bias in the signal of video quality for those content creators who were previously subject to dislike attacks and whose videos therefore received more negative feedback than warranted by their level of quality. Second, it might decrease demand by increasing uncertainty about quality overall because dislikes are no longer observable for any video. This way, the design change may adversely affect videos that were not previously subjected to dislike attacks. At the same time, engagement metrics are highly correlated for most videos, and removing dislikes might thus not matter too much.

4 Data

This study combines three datasets. The first is a repeated cross-section of trending videos from Canada and the United States, capturing daily snapshots of engagement metrics—likes, dislikes, comments, and views—for videos featured on the trending video page. From this dataset, I compile a list of channels. The second dataset is a balanced panel of all videos uploaded by these channels, allowing me to track changes in content supply over time. The third is a balanced panel of channel-level daily views, which I use to measure changes in demand.

[29] Literature in psychology has shown that stereotyping (e.g. through lower feedback for female content creators, in particular in topics where women are a minority) can reduce performance, with decrease in working memory as one potential mechanism (Spencer et al. 2016).

4.1 Dislikes and Trending Videos Data

To measure changes in dislikes following the platform design change, I use a repeated cross-section of trending videos spanning the period from August 12, 2020, to October 23, 2022. Dislikes, however, are only recorded for 34 days after the policy change. I therefore restrict the sample for the analysis of dislikes to videos posted before December 13, 2021. As 99 percent of channels in my data appear among trending videos more than once, I construct an unbalanced panel from the repeated cross-sections with minimal information loss.[30]

I focus on Canada and the United States because (i) dislike attacks, the type of harassment that YouTube aimed to address with the policy, have been most frequently discussed in the North American context; and (ii) these are the only countries for which I observe both content supply and demand responses.[31]

For each video, I observe a snapshot of engagement metrics at the time when the video was trending: the number of likes, dislikes, comments, and views. I also observe the video title, description, tags (keywords that describe the video and help viewers discover content), the video's rank on the trending list, publication time, trending time, trending country, and its assignment to one of 15 video categories.

For each channel, I observe the channel name and description, profile photo, date of creation, country of origin, verification and monetization status, and assignment to one or more of 7 possible categories and 62 subcategories. I also estimate the content creator's gender and age from the profile photo see Section 4.2.

I compare the observable characteristics of trending videos by male and female content creators in Table 1. While these videos are broadly similar in terms of video age, likes, and views, three notable differences emerge. First, videos by women receive, on average, 1,231 more dislikes (a 43 percent increase) and 3,338 more comments (a 38 percent increase) than those by men. Second, female creators whose videos appear among trending videos

[30] This procedure excludes 34 channels. Estimated effects using the repeated cross-section are comparable to those obtained from fixed-effects regressions (not reported).

[31] The original dataset also includes France, Germany, and the United Kingdom, which I use in robustness checks presented in Appendix A.3 to show that the observed effects extend beyond North America. Additionally, it includes Korea and Japan, where I do not find a gender gap in dislikes; I use these countries in a placebo test to demonstrate that, in the absence of a change in dislike behavior, there is no corresponding effect on supply or demand. This supports the interpretation that the observed downstream effects are driven by the reduction in negative feedback, rather than by the removal of public dislike counts per se.

operate channels that are, on average, 10.3 months older than those of male creators. Third, there are pronounced differences in content categories. Male creators produce 10.4 percentage points more content in entertainment, 10 percentage points more in sports, and 7.8 percentage points more in gaming, whereas female creators produce 30 percentage points more content in lifestyle.

Table 1: Balance table of pre-treatment characteristics for trending videos by content creator gender

	Male		Female		
	mean	sd	mean	sd	Diff
Age (years)					
Video age (days)	4.04	2.45	4.01	2.28	-0.035
Channel age	7.26	4.02	8.12	3.96	0.858**
Creator age	31.11	4.98	30.83	3.41	-0.273
Trending country (share)					
Canada	0.51	0.50	0.49	0.50	-0.015
United States	0.49	0.50	0.51	0.50	0.015
Channel country (share)					
United States	0.55	0.50	0.61	0.49	0.066
Canada	0.05	0.22	0.07	0.25	0.020
Missing	0.20	0.40	0.21	0.41	0.013
Other	0.20	0.40	0.11	0.31	-0.098***
Channel category (share)					
Lifestyle	0.45	0.50	0.75	0.43	0.297***
Entertainment	0.31	0.46	0.20	0.40	-0.104**
Music	0.34	0.47	0.29	0.45	-0.050
Gaming	0.13	0.33	0.03	0.16	-0.100***
Sports	0.08	0.27	0.00	0.06	-0.078***
Society	0.03	0.17	0.06	0.24	0.034
Knowledge	0.02	0.14	0.02	0.15	0.004
Engagement ('000)					
Likes	134.64	241.22	162.68	355.31	28.044
Dislikes	2.84	9.29	4.07	11.13	1.231*
Comments	8.89	19.03	12.22	26.91	3.338*
Views	2,292.78	5,281.43	2,670.22	6,488.19	377.436
Other					
Rank (out of 200)	100.49	56.52	99.82	56.48	-0.672
Verified (share)	0.81	0.40	0.84	0.37	0.030
Monetized (share)	0.93	0.26	0.95	0.22	0.020

Note: The different groups each contain the following number of observations: $n_{male} = 34,328$; $n_{female} = 12,056$.

Whether a video appears in the trending list on YouTube is determined by the platform's trending algorithm. Trending status increases a video's visibility and reach. The algorithm is non-personalized and considers several factors, including total view count, the rate of view accumulation, the sources of views (with external referrals weighted more heavily), the video's age, and its performance relative to other recent uploads from the same channel.[32] YouTube also claims to prioritize what it calls "safe content", excluding videos with severe profanity, mature themes, violence, or content that disparages others in the YouTube community.

Importantly, dislike counts do not feature in the trending algorithm. As a result, variation in dislikes does not influence whether a video appears in the trending list. An independent study by the Mozilla Foundation, based on data from over 20,000 users, confirms that dislikes are not part of the selection process.[33] Moreover, there are no indications that the trending algorithm changed around the time YouTube removed public dislike counts on November 10, 2021.[34] Based on this evidence, I assume that the trending videos before and after the policy change are comparable.[35]

4.2 Channel Classification

To examine the individual impact of gendered feedback, I limit the sample to channels that are clearly identifiable as run by individuals, and for which the individual's gender can be inferred from the profile photo. While this approach may not perfectly isolate creators who are directly exposed to feedback, as some may work with a sizable team, it serves as a reasonable proxy for identifying channels that are, on average, more likely to respond to such feedback.

Many YouTube creators use a personal photograph as their channel profile image. I use DeepFace, a Python package that wraps several state-of-the-art facial recognition models, to analyze these images and infer gender (Serengil and Ozpinar 2020).[36] The tool first

[32] https://web.archive.org/web/20220129085738/https://support.google.com/youtube/answer/7239739?hl=en

[33] https://www.wired.co.uk/article/youtube-dislike-button-mozilla-research

[34] Based on a review of https://blog.youtube/

[35] Moreover, because I estimate a TWFE OLS regression, identification comes from channels that appear in the sample in both the pre- and post-treatment periods.

[36] https://github.com/serengil/deepface. I specify the SSD model for face detection and VGG-Face for attribute estimation.

detects the number of faces in a profile image, then predicts gender for each detected face. For each face, it outputs both the probability that the object is a face and the probability that the face is male or female.[37]

Based on this classification, I identify 2,016 channels (37.5 percent) as belonging to individuals (a single face identified in the profile photo), 474 channels (8.8 percent) as group accounts (multiple faces), and 2,884 channels (53.7 percent) as corporate or other accounts (no face detected; Table 2). Since my aim is to study feedback effects on identifiable individuals, I prioritize minimizing false positives, i.e., cases where non-individual or corporate channels are incorrectly classified as individuals. I therefore set a high threshold of 90 percent facial-recognition certainty, which leaves 1,828 channels where an individual face can be confidently identified.[38]

Table 2: Category: (1) Individual, (2) Group, (3) Other

	Count	Pct
Individual	2,016	37.5
Group (2+ faces)	474	8.8
Other (No face(s) found)	2,884	53.7
Total	5,374	100.0

Of these 1,282 channels, 429 faces (23.5 percent) are identified as female and 1,399 (76.5 percent) as male (Table 3). To further reduce the risk of gender misclassification, I drop observations where gender is inferred with less than 70 percent certainty, resulting in a final sample of 1,808 individual-run channels.

Table 3: Gender

	Count	Pct
Female	429	23.5
Male	1,399	76.5
Total	1,828	100.0

To evaluate the accuracy of this classification approach, I manually code a random subsample of 339 channels (approximately 18 percent of the full sample). In this subsample, 72 channels (22 percent) were identified as female-run and 253 (78 percent) as male-run. Treating this manual coding as the ground truth, the DeepFace classifier correctly identifies

[37] I also use the tool's age estimates to verify that the average estimated age does not differ significantly by gender (Table 1).

[38] This drops less than 10 percent of the channels with a detected face.

248 of the 253 male-run channels (98 percent) and 63 of the 72 female-run channels (88 percent). While the tool performs well overall, it is more likely to misclassify women (the treated group) as men (the control group) than vice versa, thus, if anything, biasing my treatment effect estimates downward.

4.3 Supply Data

I then construct a balanced panel of monthly video produced for all channels that appear in the trending video sample at least once. Using the YouTube API, I retrieve the complete upload history for each channel. Crucially, the data includes deleted videos and their upload dates, allowing me to capture the full production history of each content creator.[39] Including deleted videos is important, as the likelihood of deletion may depend on the number of dislikes a video receives, and whether those dislikes are visible.[40]

The final sample includes 1,583 channels: 1,277 by male creators and 306 by female creators. I restrict the analysis to the period from October 2020 to November 2022 and aggregate video output at the monthly level.[41] Before the policy change, male creators produced an average of 13 videos per month, compared to 9 for female creators (Table 4), a statistically significant difference. The distribution is skewed: the median male creator produced 5 videos per month, and the median female creator produced 4.

Table 4: Balance table of pre-treatment content supply by content creator gender

	Male			Female			
	mean	sd	p50	mean	sd	p50	Diff
Monthly supply							
Video count	12.65	23.86	5.00	8.64	18.77	4.00	-4.009***

Note: The different groups each contain the following number of observations: $n_{male} = 16,575$; $n_{female} = 3,978$.

[39] The date a video is added to the upload list serves as an approximate publication date, although creators sometimes upload videos before making them public.

[40] While deleted videos are included, those made private before I accessed the API are not. Since the decision to unlist a video may itself be affected by the design change, I test for discontinuities in video unlisting behavior around the policy change. No such pattern is observed for either male or female creators (Figure A1).

[41] Although many creators post on a weekly schedule, I aggregate video output monthly to match the temporal resolution of the demand data. I restrict the supply sample to end in November 2022—slightly earlier than the demand data—to use the period from December 2022 to February 2023 to assess whether a channel remains active.

4.4 Demand Data

I construct a balanced panel of monthly views for each channel using data from the Socialblade API, which provides daily snapshots of aggregate channel views and subscriber counts.[42] As my main measure of demand, I calculate the daily change in aggregate views. These aggregate view counts exclude videos that are later deleted or made private, which means the change in views across two consecutive dates can be negative. To address this, I aggregate the data to the monthly level and drop any channels that exhibit negative month-over-month changes in views.[43]

After these adjustments, the final sample includes 1,749 channels, of which 1,317 belong to male content creators and 432 to female content creators. I restrict the sample period to October 2020 through February 2023.

There are no statistically significant gender differences in average monthly views. On average, male creators receive 28.2 million views per month, while female creators receive 33.7 million, or 16 percent more, but the difference is not statistically significant (Table 5). The distribution is highly skewed, however: at the median, male creators receive 5.4 million views per month, compared to just 3.1 million for female creators, a 56 percent difference that is statistically significant ($p < 0.001$).

Subscriber growth shows a similar pattern. On average, female channels gain 9,000 fewer subscribers per month than male channels—a difference of 18 percent ($p = 0.06$). At the median, the gap is larger: female creators gain 10,000 fewer subscribers per month than male creators—a 50 percent difference that is statistically significant ($p < 0.001$).

4.5 Comments

To test for potential spillovers from the removal of public dislikes to other feedback channels, I analyze the content of video comments. I focus on comments that directly address the video, excluding replies to avoid potential bias from sentiment directed at other users rather than the content creator.

[42] Socialblade retrieves this information daily from the YouTube API. https://socialblade.com/business-api

[43] I test whether the frequency of these negative view changes—reflecting when videos are deleted or made private—changes discontinuously after the policy change. Using the density discontinuity test developed by Cattaneo et al. (2018), I find no evidence of sorting around the treatment date (Figure A1).

Table 5: Balance table of pre-treatment demand by content creator gender

	Male			Female			
	mean	sd	p50	mean	sd	p50	Diff
Monthly demand ('000)							
Views	28,236	301,664	5,395	33,734	443,710	3,050	5,498
Subscribers	51	122	20	42	102	10	-9*

Note: The different groups each contain the following number of observations: $n_{male} = 25,023$; $n_{female} = 8,208$.

Trending videos primarily feature highly popular channels that receive a large volume of comments. For this analysis, I limit the sample to comments posted within 45 days before and after the platform design change.[44] I further restrict the analysis to comments posted within two weeks of the video's publication, which captures about 60 percent of all comments. To exclude partially treated videos, I omit the two weeks immediately preceding the policy change and define the pre-treatment period as the 45 days prior to that.

The resulting sample includes over 26 million comments on more than 50,000 videos, posted by 1,256 channels, 940 operated by male content creators and 316 by female content creators. Before the platform design change, videos by male creators received an average of 515 comments from 219 unique users, while those by female creators received 607 comments from 262 unique users, although these differences are not statistically significant (Table 6).

I use two natural language processing tools to analyze changes in the tone of comments following the policy change.[45] First, I use VADER (Valence Aware Dictionary and sEntiment Reasoner), a rule-based sentiment analysis tool designed specifically for social media text (Hutto and Gilbert 2014).[46] VADER combines a lexicon of sentiment-labeled terms with heuristics to account for punctuation, capitalization, and word order. It also handles social-media-specific content such as emojis and slang, which are common in YouTube comments. Sentiment is scored on a continuous scale from –1 (very negative) to 1 (very positive).

[44] I collected comment data approximately two years after compiling the initial dataset. During this period, some videos were deleted, limiting my ability to retrieve comments. To reduce potential bias from selectively deleted videos, I focus on a narrower pre- and post-treatment window. This yields a sample that covers 93 percent of videos in the original dataset. (As explained in Section 4.3, deleted videos still appear in the supply analysis, albeit with limited metadata.)

[45] Both tools have been used in previous academic research. For example, Aguiar (2024) use VADER, while Beknazar-Yuzbashev et al. (2025) use Perspective API. Perspective API is also widely used in industry by platforms such as The New York Times, Reddit, and YouTube. Source: https://perspectiveapi.com

[46] https://github.com/cjhutto/vaderSentiment

Second, I use Perspective API, a machine learning model developed for content moderation.[47] Trained on millions of human-rated examples, the model assigns toxicity scores based on the probability that a comment will be perceived as rude, disrespectful, or likely to drive someone out of a discussion.[48]

The main strength of Perspective API is that it measures toxicity directly, rather than general sentiment. However, unlike VADER, it does not recognize internet slang or emojis, which are common in my dataset. In addition, API quota limits restrict me to processing only a subset of comments.[49] I therefore report results from VADER in the main text and present toxicity-based results using Perspective API in Appendix A.5.

Before the platform design change, videos by male creators received comments with less positive sentiment than those by female creators. On average, comments on male creators' videos had a VADER score of 0.18, while those on female creators' videos scored 0.33. This pattern persists across the distribution: at the 10th percentile, comments score –0.31 for men and –0.15 for women; at the 90th percentile, the scores are 0.70 and 0.82, respectively (Table 6).

Table 6: Balance table of pre-treatment characteristics for trending videos by content creator gender

	Male		Female		
	mean	sd	mean	sd	Diff
Count					
Comments	515.23	2,624.26	606.82	1,573.87	91.597
Commentators	219.32	1,239.49	261.56	812.49	42.246
Sentiment [-1,1]					
Mean	0.18	0.18	0.33	0.20	0.148***
10th percentile	-0.31	0.28	-0.15	0.29	0.157***
50th percentile	0.14	0.23	0.34	0.29	0.196***
90th percentile	0.70	0.22	0.82	0.20	0.118***

Note: The different groups each contain the following number of observations: $n_{male} = 19,784$; $n_{female} = 4,306$. 11,933 comments (less than half of a percent) were not classified, for example due to being too short.

[47] https://perspectiveapi.com

[48] For more details, see https://developers.perspectiveapi.com/s/about-the-api-training-data. For example, if three out of ten raters label a comment as toxic, the model assigns a toxicity score of 0.3.

[49] I analyze a random sample of 100 comments per video using the Perspective API.

5 Empirical Approach

5.1 Effect on Dislikes for Female Content Creators

I estimate the average treatment effect of the platform design change in a two-way fixed effects OLS regression specification:

$$Y_{ict} = \alpha + \beta Post_t * Female_i + \gamma' \mathbf{X}_{ict} + \delta_{ic} + \lambda_t + \epsilon_{ict} \tag{1}$$

where Y_{ict} is the logged daily mean count of dislikes per channel-country observation, $Post_t$ is an indicator for the post-treatment period, $Female_i$ is an indicator for female content creators, X_{ict} is a vector of covariates, δ_{ic} denotes channel-country fixed effects that account for time-invariant channel and country specific characteristics, and λ_t denotes daily fixed effects that account for time-specific factors. Covariates include video characteristics: number of views, video age (in hours), and fixed effects for 15 video categories.[50] The coefficient β captures the average treatment effect on women's channels.[51]

5.2 Controlling for Content Differences with Counterfactual Dislikes

As Table 1 illustrates, male and female content creators produce different types of content. To assess how these differences might affect my estimates, I take two complementary approaches. First, I include channel-country and video-topic fixed effects in all regressions. These fixed effects control for systematic differences in content categories, so that gender effects are identified from within-channel and within-topic variation. Second, I predict the counterfactual number of dislikes that a given video by a female creator would have received had it been produced by a male creator. I then include this predicted value as an additional control in the regression. This allows me to account for more nuanced differences in content style or genre that may not be fully captured by broader topic categories or fixed effects.

[50] While channel categories are fixed within each channel, video categories can vary over time.

[51] Throughout the paper, I interpret coefficient estimates as percentage effects when reporting results. While the coefficients deviate from the precise percentage effects, the magnitude of deviations is at most 4 percent. The precise effect can be calculated as: $exp(\beta) - 1$.

To generate these predictions, I train a machine-learning model on data describing videos produced by men.[52] I restrict the training sample to pre-treatment observations, since men also experienced some degree of treatment. To implement this approach, I estimate a Lasso (Least Absolute Shrinkage and Selection Operator, L1) regularized regression (James et al. 2023):

$$y_i = X_i\theta + \epsilon_i \tag{2}$$

where y_i is the logged number of dislikes, X_i is the vector of predictors and θ is the coefficient vector to be estimated. The Lasso estimator minimizes the following cost function:

$$\hat{\theta} = \arg\min_{\theta} \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - X_i\theta)^2 + \alpha \sum_{j=1}^{p} |\theta_j| \right) \tag{3}$$

In this equation, n is the number of observations, y_i is the true target number of dislikes for the i-th observation, $\hat{y}_i$ is the predicted target number of dislikes for the i-th observation, $\hat{y}_i = X_i\theta$ is the coefficient vector for the predictors, p is the number of predictors, α is the regularization parameter controlling the strength of the L1 penalty, and $\sum_{j=1}^{p} |\theta_j|$ is the L1 regularization term encouraging sparsity in the coefficients.

Intuitively, the Lasso estimator shrinks coefficient estimates toward zero, with the L1 penalty driving some coefficients to be exactly zero. That is, Lasso effectively performs variable selection, improving interpretability. I use Lasso over Ridge regression, which includes all predictors and therefore lacks this sparsity property.[53]

Predictors include the number of likes, views, comments, and their ratios, which capture popularity and engagement as proxies for audience interest. Video and channel categories control for differences in content type and target audience. The video's trending country and country of origin capture geographic and demographic patterns. Timing variables,

[52] This approach is similar to that of Iaria et al. (2024), who document a gender gap in scientific citations and face the challenge that papers by men and women may differ systematically. They train a model to predict citations using only papers authored by men, and show that the estimated gender gap remains robust even when including these counterfactuals in the regression.

[53] Ridge regression uses an L2 penalty ($\sum_{j=1}^{p} \theta_j^2$), which discourages large coefficients but does not shrink any to zero. As a result, Ridge includes all predictors in the final model. While this can improve prediction accuracy, it comes at the cost of reduced interpretability. In my setting, Ridge regression performs comparably well, explaining 96 percent of the variation in the test sample.

such as the publication time, the time at which the video began trending, and the elapsed time between these events, account for temporal variation in engagement. I also include indicators for whether a channel is monetized or verified, which proxy for status or professionalism. Finally, interactions and second-degree polynomials of these variables allow the model to capture potential nonlinear relationships.

I split the sample into training and testing sets (80/20 percent) and standardize all predictors by removing the mean and scaling to unit variance. Using 10-fold cross-validation, I select a regularization parameter of $\alpha = 7.7$. The resulting Lasso model explains 96 percent of the variation in the test set. This high R^2 reflects the strong correlation between dislikes and other engagement metrics such as likes, views, and comments. It confirms that, aside from dislike attacks, dislikes are highly predictable based on audience engagement and video characteristics.

I then use the trained model to predict the *counterfactual* number of dislikes for the entire sample and re-estimate Equation 1, including the predicted dislikes as an additional control. This variable can be interpreted as the *predicted* logged daily mean number of dislikes per channel–country observation, had the video been produced by a male creator in the pre-treatment period.[54] Controlling for these predicted dislikes allows me to condition the analysis on the underlying dislikability of the content, i.e. the level of negative feedback the video would have received absent gender-biased responses.

5.3 Additional Specifications

To assess the baseline effect of the design change on all creators, I also estimate a version of Equation 1 that excludes time fixed effects, allowing the post-period dummy to capture the overall change in dislikes following the policy. This coefficient reflects the average reduction in dislikes experienced by creators irrespective of gender.

Next, to examine how the gender gap in dislikes evolved, I estimate Equation 1 separately for the pre- and post-treatment periods, omitting channel–country fixed effects so that the coefficient on the female dummy measures the average difference in dislikes between men and women. Comparing these estimates across periods provides a direct

[54] Negative predicted dislikes are replaced with zero, and then 1 is added to all dislike counts before taking the logarithm. This adjustment affects 3 percent of the total predicted dislikes and 12 percent of the predicted dislikes for videos by female creators.

measure of how the platform design change changed the gender gap in feedback.

5.4 Effect on Supply and Demand of Content

I estimate the impact of the policy on content supply and demand using a difference-in-differences design, with women as the treated group and men as the control group. This approach is motivated by TWFE OLS estimates showing that the platform design change reduced dislikes for both male and female content creators, with a substantially larger effect for women.

$$Y_{it} = \alpha + \beta' Post_t * Female_i + \gamma' \mathbf{X}_{it} + \delta_i + \lambda_t + \epsilon_{it} \tag{4}$$

where Y_{it} is either the (i) logged video count or the (ii) logged video views for channel i and month t, $Post_t$ is an indicator for the post-treatment period, $Female_i$ is an indicator for female content creators, X_{it} is a vector of covariates, δ_i denotes channel fixed effects that account for time-invariant channel specific characteristics, and λ_t denotes time fixed effects that account for time-specific factors. The coefficient β estimates the difference-in-differences effect for women, capturing the change in outcomes before and after the policy relative to male creators. This estimate can be interpreted as a lower bound on the average treatment effect, under two key assumptions: (i) that male and female creators would have followed parallel trends in the absence of treatment, and (ii) because men were also exposed to the policy change, that both groups responded in the same direction.

To assess the parallel trends assumption and examine the persistence of effects over time, I estimate and plot an event-study version of the regression (Figures 4 and 5).

To support the second assumption, I conduct a placebo test using data from Japan and Korea, where no gender gap in dislikes is observed (Table 17). In this context, I find no differential response to the policy in either content supply or demand (Tables 18 and 19).

5.5 Mechanisms

5.5.1 Heterogeneous Effects on Dislikes

To examine how the effect of the platform design change varies across the distribution of dislikes, I estimate quantile treatment effects. However, because my preferred specifications include panel fixed effects, I cannot directly estimate a standard quantile regression: conventional methods for removing unobserved fixed effects (such as demeaning or differencing) rely on the linearity of expectations, which does not hold for quantiles.

Instead, I implement the two-step estimator proposed by Canay (2011). In the first step, I estimate the fixed effects regression in Equation 1 and retrieve the estimated channel-country fixed effects. I then demean the outcome variable using these fixed effects. In the second step, I estimate a quantile regression on the demeaned data:

$$Quant_{\tau}(Y_{ict}) = \alpha(\tau) + \beta' Post_t * Female_i(\tau) + \gamma' \mathbf{X}_{ict} + \epsilon_{ict} \tag{5}$$

where Y_{ict} is the demeaned logged daily mean number of dislikes per channel-country observation, $Female_i$ is the dummy for female creators, $Post_t$ is the treatment dummy, and X_{ict} is a vector of covariates. All coefficients vary with the quantile τ. The coefficient $\beta(\tau)$ captures the treatment effect for female creators at quantile τ, relative to male creators at the same quantile.

The key assumption underlying this approach is that, at each quantile, the outcome variable is a linear function of the covariates (Cameron and Trivedi 2022). Following Canay (2011), I bootstrap the standard errors for correct inference.

To examine how the gender gap in feedback varies across the distribution of dislikes before and after the design change, I split the data into pre- and post-treatment samples and estimate a pooled quantile regression for each period. In this specification, I exclude the channel-country fixed effects and include a female dummy variable.

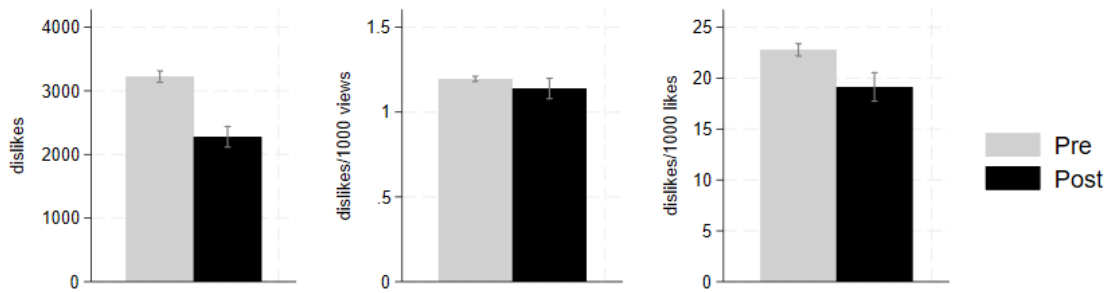
6 Results

6.1 Dislikes Before and After the Design Change: Descriptive Evidence

To illustrate the effect of the platform design change on dislikes, I compare average dislikes per video in the 454 days before and the 34 days after the policy was implemented. Following the removal of public dislike counts, the number of dislikes declines: the average video receives 937 fewer dislikes in the post-treatment period: a 30 percent decrease (Figure 2).

Part of this decline may reflect a faster media cycle, with trending videos being replaced more quickly over time. When normalizing by exposure, the reduction in dislikes is smaller. Per 1,000 views, the average video receives 0.05 fewer dislikes in the post period, a 4 percent decrease that is only marginally significant (p = 0.xx). Using likes as the normalization metric instead, the average video receives 3 fewer dislikes per 1,000 likes, corresponding to a 14 percent reduction after the design change.

Figure 2: Overall Dislikes



To assess the impact of the platform design change on dislikes by gender, I compare the average number of dislikes per video before and after the change, separately for channels operated by male and female content creators. This comparison reveals substantial pre-treatment differences. Prior to the design change, videos by women received significantly

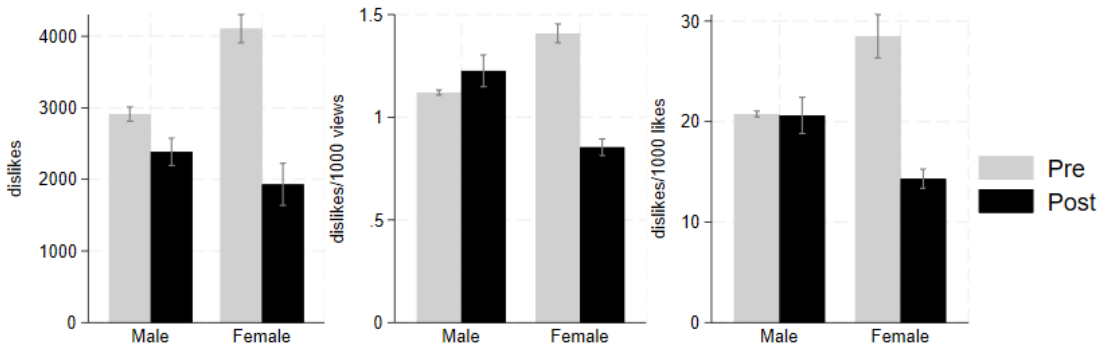
more dislikes than those by men: an average of 4,067 versus 2,836. This disparity also holds when normalizing by exposure: women received 1.4 dislikes per 1,000 views and 29 dislikes per 1,000 likes, compared to 1.1 and 21, respectively, for men (Figure 3).

Following the design change, women experienced a substantial decline in dislikes. On average, their videos received 1,574 dislikes, compared to 2,354 for men. This pattern persists in relative terms: women's videos received 0.9 dislikes per 1,000 views and 15 per 1,000 likes, while men's received 1.2 and 21, respectively.

For women, this corresponds to a reduction of 2,313 dislikes per video (–57 percent), 0.55 fewer dislikes per 1,000 views (–39 percent), and 14 fewer per 1,000 likes (–49 percent) in the post period. In contrast, videos by men saw a more modest reduction of 482 dislikes per video (–17 percent). When normalized by views, men's videos received 0.1 more dislikes per 1,000 views—a change that is not statistically significant ($p = 0.102$). Dislikes per 1,000 likes remained stable at 21.

Taken together, these patterns suggest a gendered response to the platform design change, with women experiencing a substantially larger reduction in negative feedback.

Figure 3: Dislikes by Gender



6.2 Causal Effect of the Design Change on Dislikes

I estimate the causal effect of the platform design change on dislikes using a difference-in-differences regression (Table 7). The results show that, relative to men, female content creators experienced a 30 percent reduction in dislikes following the design change (Column 1). This effect is robust to the inclusion of additional controls: conditioning on the number of views per video and video age, the estimated effect decreases to 21 percent but remains statistically significant (Column 2). Adding fixed effects for video topic categories leaves the estimate unchanged, suggesting that the gender-differentiated effect is not driven by differences in content type (Column 3).

Table 7: Effect of platform design change on dislikes

	(1) ln(Dislikes)	(2) ln(Dislikes)	(3) ln(Dislikes)
Post × Female	-0.30**	-0.21**	-0.21**
	(0.12)	(0.10)	(0.10)
Observations	49,759	49,759	49,759
R-squared	0.76	0.83	0.83
YT Channel-Country IDs	2864	2864	2864
YT Channels	1808	1808	1808

Standard errors clustered by channel
*** p<0.01, ** p<0.05, * p<0.1

Note: Two-way fixed-effects OLS estimates with date and channel-country fixed effects. Column (2) additionally controls for number of views and video age, column (3) additionally controls for video category fixed effects. The outcome is ln(Dislikes). Standard errors clustered by channel. *** p<0.01, ** p<0.05, * p<0.1.

To better account for potential differences in content between male and female creators, I condition the analysis on predicted counterfactual dislikes, that is, the number of dislikes a video by a woman would have received if it had been created by a man prior to the design change. The estimated coefficient on predicted dislikes is positive and statistically significant across all specifications (Table 8): a 1 percent increase in predicted dislikes is associated with a 0.11 to 0.15 percent increase in actual dislikes, even after controlling for time and channel fixed effects.

After accounting for predicted dislikes, the estimated treatment effect increases: women experience a 31 percent reduction in dislikes following the design change, relative to men (Column 1). When additionally controlling for views and video age, the estimated effect

falls slightly to 23 percent (Column 2), and increases to 24 percent after further including fixed effects for the video's topic category (Column 3). That the estimated treatment effect becomes larger when conditioning on dislikability may suggest that, on average, women produce content that is less likely to elicit excessive negative feedback.

Table 8: Controlling for gender differences in types of content produced

	(1) ln(Dislikes)	(2) ln(Dislikes)	(3) ln(Dislikes)
Post × Female	-0.31***	-0.23***	-0.24***
	(0.09)	(0.09)	(0.09)
ln(Predicted dislikes)	0.15***	0.11***	0.11***
	(0.01)	(0.01)	(0.01)
Observations	49,759	49,759	49,759
R-squared	0.80	0.85	0.85
YT Channel-Country IDs	2864	2864	2864
YT Channels	1808	1808	1808

Standard errors clustered by channel
*** p<0.01, ** p<0.05, * p<0.1

Note: Two-way fixed-effects OLS estimates with date and channel-country fixed effects. Column (2) additionally controls for number of views and video age, column (3) additionally controls for video category fixed effects. The outcome is ln(Dislikes). Standard errors clustered by channel. *** p<0.01, ** p<0.05, * p<0.1.

The platform design change significantly reduced dislikes on women's content. But what was its effect on male content creators? To estimate the impact on men, I exclude time fixed effects from the specification. This allows me to identify a baseline coefficient for the post-treatment period, capturing changes in the number of dislikes received by men after the design change (Table 9).

Controlling only for channel fixed effects (Column 1), I find that men experienced a 12 percent reduction in dislikes following the policy change ($p = 0.029$). However, this effect becomes smaller and statistically insignificant once I control for video-level characteristics such as views and video age (Column 2), add topic category fixed effects (Column 3), or condition on predicted counterfactual dislikes (Column 4). These results suggest that the design change had only a limited impact on male content creators.

Next, I assess whether the platform design change closed the gender gap in negative feedback by estimating split-sample regressions without channel-country fixed effects, al-

Table 9: Treatment effects for men

	(1) ln(Dislikes)	(2) ln(Dislikes)	(3) ln(Dislikes)	(4) ln(Dislikes)
Post	-0.12**	-0.05	-0.04	-0.06
	(0.06)	(0.05)	(0.05)	(0.05)
Post × Female	-0.32***	-0.23**	-0.23**	-0.25***
	(0.12)	(0.10)	(0.10)	(0.09)
ln(Predicted dislikes)				0.12***
				(0.01)
Observations	49,759	49,759	49,759	49,759
R-squared	0.74	0.81	0.82	0.84
YT Channel-Country IDs	2864	2864	2864	2864
YT Channels	1808	1808	1808	1808

Standard errors clustered by channel
*** p<0.01, ** p<0.05, * p<0.1

Note: OLS estimates with channel-country fixed effects. Column (2) additionally controls for number of views and video age, columns (3) and (4) additionally controls for video category fixed effects. The outcome is ln(Dislikes). Standard errors clustered by channel. *** p<0.01, ** p<0.05, * p<0.1.

lowing identification of a coefficient on a female dummy variable (Table 10).

In the pre-treatment sample, the coefficient on the female dummy is 13 percent, though marginally statistically insignificant ($p = 0.106$) when not controlling for predicted dislikes (Column 1). When I control for predicted dislikes, the estimated gap increases slightly to 14 percent and becomes statistically significant at the 5 percent level ($p = 0.041$; Column 2).

In the post-treatment sample, the coefficient becomes negative and statistically insignificant: –8 percent without controlling for predicted dislikes (Column 3) and –4 percent when doing so (Column 4).[55]

Taken together, these results suggest that the platform design change substantially reduced, and likely closed, the gender gap in negative feedback. After the design change, women no longer receive significantly more dislikes than men for similarly "dislikable" content.

[55]Because men are also partially treated, I separately predict dislikes based on videos by male creators in the post-treatment sample and use these predictions as controls in the corresponding regressions. I follow the same procedure as for the pre-treatment sample. The model performs similarly well: using 10-fold cross-validation, I set the regularization parameter $\alpha = 10.5$, and the Lasso explains 95 percent of the variation in the test sample.

Table 10: Gender gap in feedback before and after treatment

	(1) ln(Dislikes)	(2) ln(Dislikes)	(3) ln(Dislikes)	(4) ln(Dislikes)
Female	0.13	0.14**	-0.08	-0.04
	(0.08)	(0.07)	(0.13)	(0.11)
ln(Predicted dislikes)		0.21***		0.23***
		(0.01)		(0.02)
Observations	46,407	46,407	3,386	3,386
R-squared	0.45	0.57	0.56	0.67
YT Channels	1706	1706	408	408

Standard errors clustered by channel
*** p<0.01, ** p<0.05, * p<0.1

Note: OLS estimates with date fixed effects. All columns additionally control for number of views, video age and video category fixed effects. The outcome is ln(Dislikes). Standard errors clustered by channel. *** p<0.01, ** p<0.05, * p<0.1.

6.3 Effect on Supply

I estimate the causal effect of the platform design change on female content creators' supply of content in a difference-in-differences regression (Table 11). The estimates indicate that, relative to men, women increase their monthly supply of videos by 8.4 percent after the design change. This result corresponds to an additional 0.4 videos uploaded per month for the median content creator in my sample. The effect is statistically significant at the 5 percent level, with a 95 percent confidence interval of (1.1, 18.0) percent.

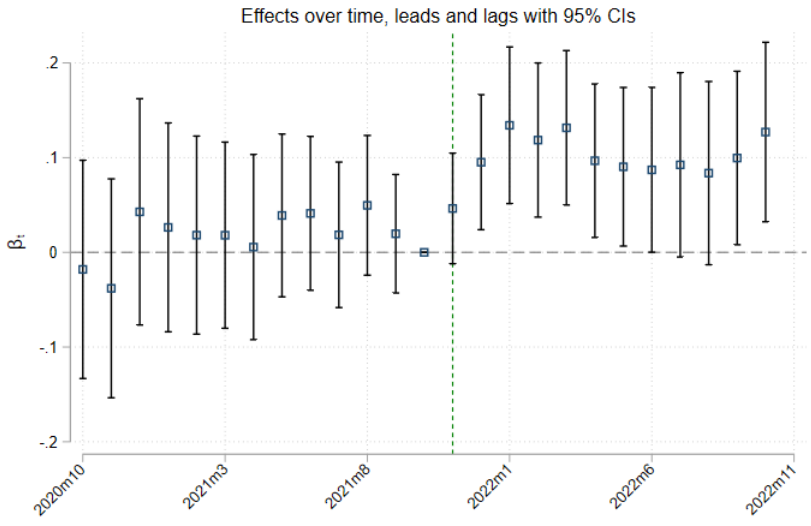
Table 11: Effect on Supply

	(1) ln(Videos)
Post × Female	0.084
	[.011,.180]
Observations	37,992
Channels	1,583
R-squared	0.68
Median number of videos	5

Note: Two-way fixed-effects difference-in-differences estimates with month and channel fixed effects. Standard errors calculated using wild cluster bootstrap method with 200 replications. 95 percent confidence intervals reported.

Event study estimates show that female and male content creators followed parallel trends in content supply prior to the platform design change (Figure 4). Following the change, women's content supply increases relative to men's and remains elevated throughout the remainder of the sample period, suggesting a persistent treatment effect.

Figure 4: Parallel trends in monthly supply



Note: Event-study estimates with month and channel fixed effects. Standard errors calculated using wild cluster bootstrap method with 200 replications. 95 percent confidence intervals reported.

6.4 Effect on Demand

I estimate the causal effect of the platform design change on demand for content by female content creators using a difference-in-differences regression (Table 12). The estimates show that, relative to men, women experience a 15.5 percent increase in monthly views following the design change. The effect is statistically significant at the 5 percent level, with a 95 percent confidence interval of [0.4, 31.3] percent. A back-of-the-envelope calculation suggests that this corresponds to an increase in monthly revenue of approximately $193 to 3{,}100$ for the median channel in my sample.[56]

Event study estimates show that demand for content by male and female creators followed parallel trends prior to the platform design change (Figure 5). Following the change, demand for women's content increases and stays higher for the rest of the sample period, suggesting a persistent treatment effect similar to that observed for content supply.

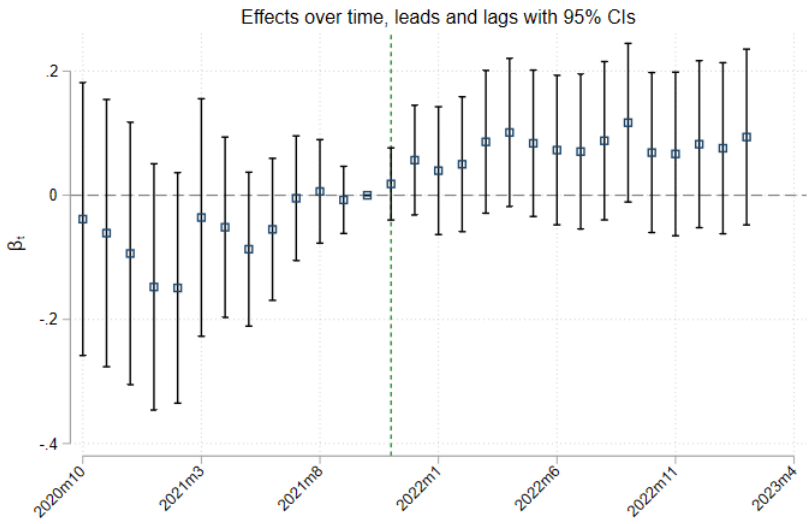
[56] Based on revenue estimates of \$0.25–\$4.00 per 1,000 views by Socialblade.com.

Table 12: Effect on Demand

	(1) ln(Views)
Post × Female	0.155 [0.004,0.313]
Observations	48,972
Channels	1,749
R-squared	0.74
Median views	4,990,216

Note: Two-way fixed-effects difference-in-differences estimates with month and channel fixed effects. Standard errors calculated using wild cluster bootstrap method with 200 replications. 95 percent confidence intervals reported.

Figure 5: Parallel trends in monthly demand



Note: Event-study estimates with month and channel fixed effects. Standard errors calculated using wild cluster bootstrap method with 200 replications. 95 percent confidence intervals reported.

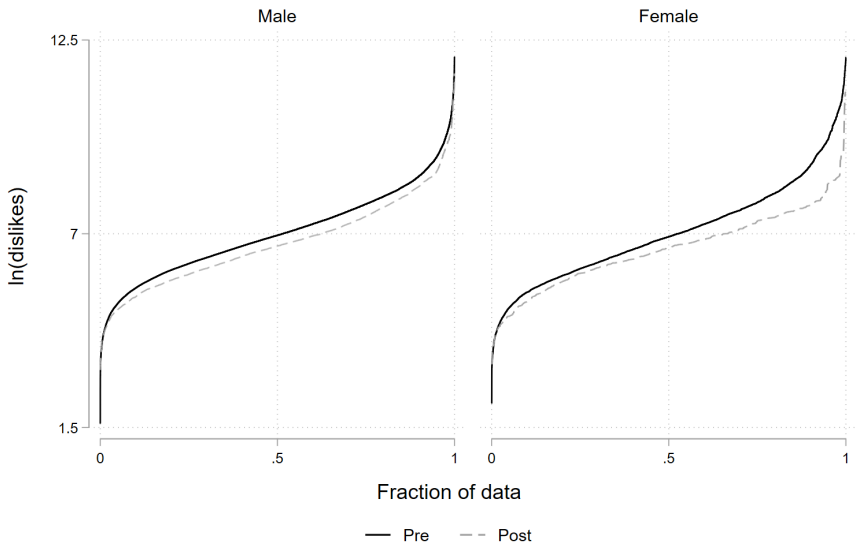
7 Mechanisms

I document a gender gap in feedback on YouTube: Female content creators receive significantly more negative feedback but removal of public dislikes eliminates this gap. Additional analyses show that the closure of the gender feedback gap coincides with a significant increase in the productivity of female content creators and in the demand for their content. In this section, I investigate the mechanisms behind these results.

7.1 Heterogeneous Effects on Dislikes

A comparison of the distribution of dislikes by gender before and after the platform design change reveals distinctly different responses towards male and female content creators (Figure 6). For male content creators (left), the dislike curve shifts downward, indicating an overall reduction in dislikes that appears symmetrical across the distribution. In contrast, the distribution of dislikes towards female content creators (right) rotates. Specifically, dislikes decrease significantly more in the right tail of the distribution. Intuitively, this rotation suggests that the removal of public dislike counts eliminated dislike attacks, which by YouTube's definition are associated with exceptionally high dislike counts for videos that would have otherwise received few.

Figure 6: Comparison of dislike distribution pre- and post-treatment by gender

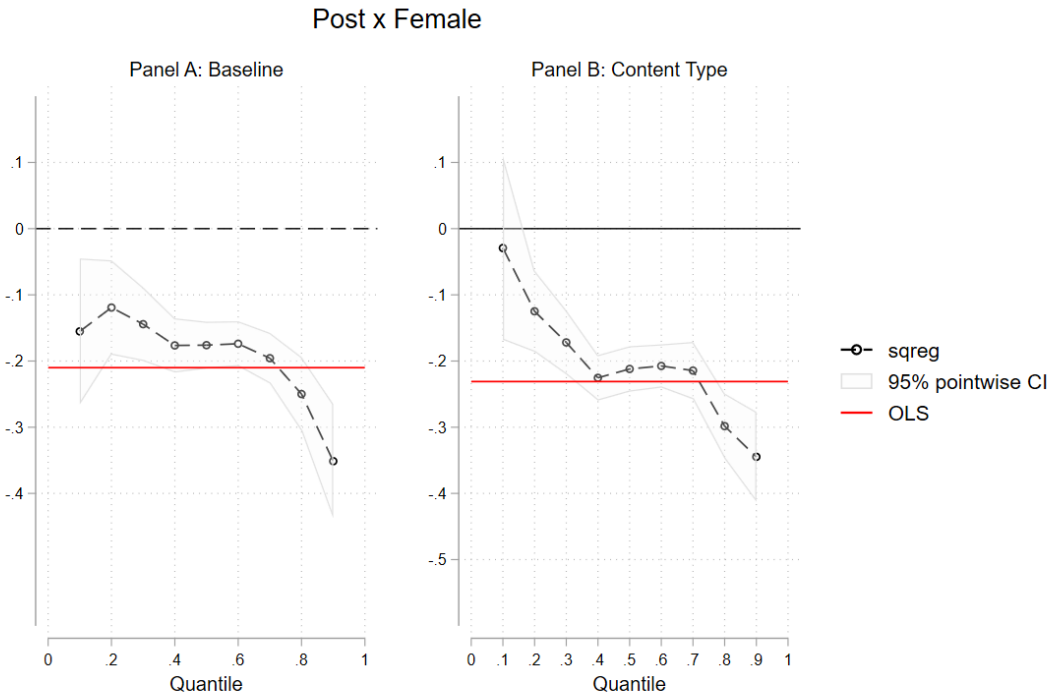


I estimate a quantile regression using the Canay (2011) two-step estimator to analyze the heterogeneous effects of making dislikes private on the number of dislikes (Panel A: Baseline, Figure 7). This specification is equivalent to the OLS difference-in-differences model that I use to obtain average treatment effects for the treated channels but allows me to capture changes in the treatment effect along the dislike distribution.[57] The treatment effect on dislikes towards women's content is negative throughout the distribution. The

[57] Specification in Panel A: Baseline in Figure 7 corresponds to the OLS difference-in-differences regression in Column 2 of Table 7 and accounts for channel-country fixed effects, daily fixed effects, video age, and views per video. See Table A.1 in the Appendix for detailed results.

effect, however, shows significant levels of heterogeneity. It is lowest in the left tail of the distribution (between 12 and 16 percent for quantiles 10 to 30) and significantly higher in the right tail (between 25 and 35 percent for quantiles 80 to 90). The regression estimates align closely with the descriptive evidence, indicating that the removal of public dislikes (i) induced a downward shift in the distribution of dislikes towards women's content and (ii) lead to a rotation in the dislike distribution with larger effects in the right tail. This finding is consistent with a reduction in dislike attacks, as evidenced by a substantial decrease in the number of videos with exceptionally high dislike counts after the design change. Estimates remain comparable conditional on predicted dislikes (Panel B: Type of Content, Figure 7).[58]

Figure 7: Quantile regression estimates of platform design change on dislikes



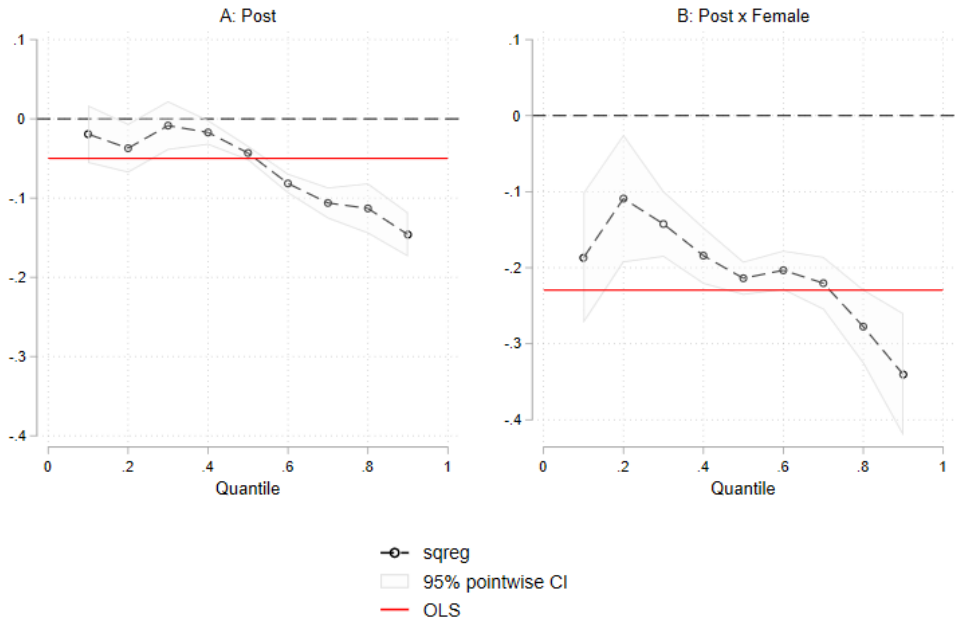
To estimate the heterogeneous effects of the platform design change on male content creators, I exclude the time fixed-effects from the specification to estimate a baseline coefficient for the post-period (Panel A in Figure 8).[59] The baseline estimate (Panel A: Post),

[58] Specification in Panel B: Content Type 7 corresponds to the OLS difference-in-differences regression in Column 2 of Table 8 and accounts for predicted dislikes, channel-country fixed effects, daily fixed effects, video age, and views per video. See Table A.2 in the Appendix for detailed results.

[59] Specification in Figure 8 corresponds to the OLS regression in Column 2 of Table 9 and accounts for channel-country fixed effects, video age, and views per video. See Table A.4 in the Appendix for detailed results.

like the additional treatment effect female content creators (Panel B: Post x Female), leads to a rotation in the dislike distribution with larger effects in the right, i.e. in very large dislike counts (largest decrease of 16.7 percent in q90). At the same time, the baseline effect is not significantly different from zero in left tail.

Figure 8: Quantile regression estimates: Treatment effect for men and women



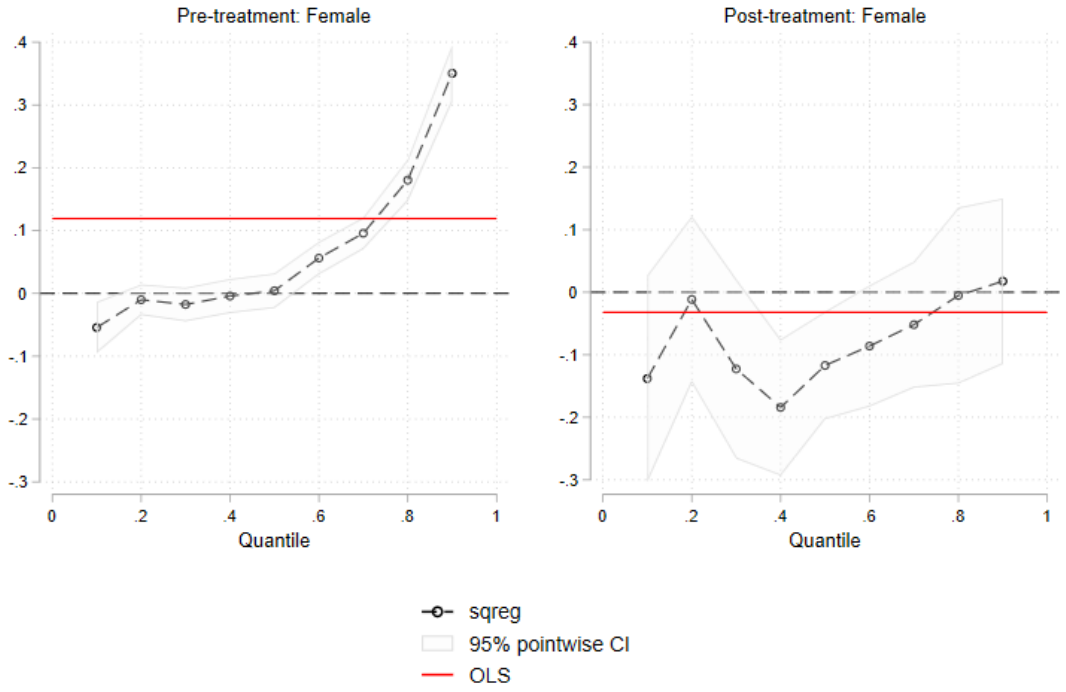
Next, I assess whether the platform design change closed the gender gap in feedback across the entire dislike distribution rather than 'just' on average (as shown in Table 10). To implement this analysis, I first divide the dataset into pre-treatment and post-treatment subsamples. I then estimate a pooled quantile regression for each sample separately. This specification allows me to separately identify the gender feedback gap throughout the dislike dislike distribution before and after the design change.[60]

The pre-treatment estimates indicate that women receive significantly more dislikes than men at and above the 60th percentile of the distribution, with the effect monotonically increasing from lower to higher percentiles. Specifically, at the 10th percentile, women receive 5.4 percent fewer dislikes than men, while at the 90th percentile, they receive 35 percent more dislikes than men. In contrast, the post-treatment estimates no longer display

[60]Specification in Figure 9 corresponds to the OLS regression in Columns 2 and 4 of Table 10 and accounts for channel-country fixed effects, monthly time-fixed effects, video age, views per video, channel category, and predicted dislikes. Monthly time fixed effects instead of daily time fixed effects are included to ensure convergence of the estimates given the highly unbalanced panel structure and the inclusion of a large number of channel category dummies. See Table A.5 in the Appendix for detailed results.

consistent variation in the gender gap across the distribution. Instead, point estimates are almost all (8 out of 9 percentiles) negative and almost all (7 out of 9 percentiles) insignificant at 5 percent significance level, and no longer depict any discernible pattern. This result suggests that the gender gap in feedback was eliminated both on average and throughout the entire dislike distribution.

Figure 9: Quantile regression estimates: Gender gap in feedback before and after treatment



7.2 Spillover Effects on Comments

Next, I use natural language processing to analyze whether the removal of public discounts led to negative feedback spilling over into video comments. I examine changes in comments at the extensive and intensive margins. At the extensive margin, I assess three outcomes: (i) the total number of comments (Column 1), (ii) the number of unique commentators (Column 2), and (iii) the average number of comments per commentator (Column 3) (Table 13). The treatment seems to have had no effect on the extensive margin, with the coefficients being small and statistically insignificant across all three outcomes.

Table 13: Effect on comment quantity

VARIABLES	(1) Comments	(2) Commentators	(3) Comments per commentator
Post × Female	16.99	-7.56	0.00
	(65.94)	(28.85)	(0.00)
Constant	509.86***	217.97***	1.05***
	(7.15)	(3.13)	(0.00)
Observations	50,816	50,816	50,816
R-squared	0.44	0.35	0.42
Channels	1176	1176	1176
Comment publ. date FE	✓	✓	✓
Channel FE	✓	✓	✓
Level of analysis	Video	Video	Video

Standard errors clustered by channel
*** p<0.01, ** p<0.05, * p<0.1

To evaluate the effects at the intensive margin, I examine the mean sentiment of comments across three levels of data aggregation: (i) disaggregated at the comment level (Column 1), (ii) averaged at the video level (Column 2), and (iii) averaged at the channel-day level (Column 3) (see Table A.11). At the comment level, the average sentiment increases slightly (Column 1). In addition, this specification reveals that comments on women's videos, on average have a more positive sentiment. At the video level, the effect is close to zero and statistically insignificant (Column 2).The combination of these two findings suggest that there may have been changes in the composition of comments per video. At the channel-day level, I observe a slight increase in the average sentiment per channel (Column 3).[61] Taken together, these results suggest that, if anything, the platform design change increased, rather than decreased, the average sentiment of comments.

[61] One can think of the measure in Column 3 as the stream of comments a creator is exposed to per day.

Table 14: Effect on the mean comment sentiment

VARIABLES	(1) Mean sentiment	(2) Mean sentiment	(3) Mean sentiment
Female	0.12***		
	(0.01)		
Post × Female	0.02*	-0.00	0.01**
	(0.01)	(0.01)	(0.01)
Constant	0.16***	0.21***	0.24***
	(0.00)	(0.00)	(0.00)
Observations	26,162,217	50,813	95,397
R-squared	0.02	0.45	0.42
Channels		1176	1253
Video upload date FE	✓	✓	
Comment publishing date FE	✓		✓
Channel FE		✓	✓
Level of analysis	Comment	Video	Channel

Standard errors clustered by channel
*** p<0.01, ** p<0.05, * p<0.1

To understand the origins of the change in mean sentiment and to identify any shifts in the composition of comments, I also estimate the effect of the platform design change on the tails of the sentiment distribution (Table A.12). The effect on the upper tail (at the 90th percentile) is small and insignificant at both the video level (Column 1) and channel level (Column 2). Similarly, the effect on the lower tail (at the 10th percentile) is small and insignificant at the video level (Column 1), but positive and significant at the channel level (Column 2). This finding suggests that the increase in average comment sentiment primarily stems from a reduction in negative comments.

Table 15: Effect on comment sentiment at the tails

	(1)	(2)	(3)	(4)
VARIABLES	p90 (posit.)	sentiment	p10 (neg.)	sentiment
Post × Female	-0.01	0.01	-0.00	0.02*
	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.72***	0.73***	-0.28***	-0.26***
	(0.00)	(0.00)	(0.00)	(0.00)
Observations	50,813	95,397	50,813	95,397
R-squared	0.44	0.31	0.36	0.32
Channels	1176	1253	1176	1253
Upload date FE	✓		✓	
Channel FE	✓	✓	✓	✓
Comment publ. date FE		✓		✓
Level of analysis	Video	Channel	Video	Channel

Standard errors clustered by channel
*** p<0.01, ** p<0.05, * p<0.1

7.3 Quality Effects

The previous sections offer indirect evidence on the effects of the platform design change on video quality. First, the observed increase in video demand significantly exceeded the observed effect on video supply, suggesting an increase in per-video views. Second, video comments became slightly more positive, on average, after the design change. I conduct further analysis to study the effect of the design change on quality.

As an additional measure of video quality, I estimate the effect of the treatment on predicted dislikes (Table 16). These dislikes are predicted based on dislikes for videos by male content creators prior to the treatment and thus are a measure of video quality that excludes gender bias. The results suggest that predicted dislikes decrease considerably over time: TWFE DiD estimates show a 48 percent decrease (Column 1), estimates with additional controls for the number of views per video and video age show a 39 percent decrease (Column 2), and estimates including controls for the video's topic category show a 38 percent decrease (Column 3).

Table 16: Effect of platform design change on predicted dislikes

VARIABLES	(1) ln(pred. dislikes)	(2) ln(pred. dislikes)	(3) ln(pred. dislikes)
Post × Female	-0.48***	-0.39***	-0.38***
	(0.15)	(0.14)	(0.14)
Observations	83,683	83,683	83,683
R-squared	0.49	0.52	0.52
YT Channel-Country IDs	4082	4082	4082
YT Channels	2567	2567	2567
Channel-Country FE	✓	✓	✓
Daily FE	✓	✓	✓
1000 views		✓	✓
Video age in hours		✓	✓
Video cat. FE			✓

Standard errors clustered by channel
*** p<0.01, ** p<0.05, * p<0.1

7.4 Placebo test: Impact on Content Supply and Demand Among Japanese & Korean Content Creators

Finally, I conduct a placebo test to show that the supply and demand responses I observe are indeed driven by changing dislike behavior and not a result of aggregate shifts in content production. Specifically, I show that demand and supply effects for female content creators can only be observed where I also observe a reduction in the gender feedback gap. To implement this placebo test, I begin by estimating the causal effect of the platform design change on dislikes towards female content creators in Korea and Japan (Table 17). The estimated effect is insignificant in TWFE DiD specifications as well as after including additional controls. This finding allows me to test whether demand and supply effects can also be observed if there is no reduction in the feedback gender gap.

Table 17: Effect of platform design change on dislikes for Korean and Japanese channels

VARIABLES	(1) ln(Dislikes)	(2) ln(Dislikes)	(3) ln(Dislikes)
Post × Female	0.09	0.06	0.07
	(0.11)	(0.10)	(0.10)
Observations	31,582	31,582	31,582
R-squared	0.71	0.76	0.76
YT Channel-Country IDs	1042	1042	1042
YT Channels	1029	1029	1029
Channel-Country FE	✓	✓	✓
Daily FE	✓	✓	✓
1000 views		✓	✓
Video age in hours		✓	✓
Video cat. FE			✓

Standard errors clustered by channel
*** p<0.01, ** p<0.05, * p<0.1

Estimates of the effect of the platform design change on video supply are insignificant and close to zero (Table 18). This suggests that, in absence of a treatment effect on dislikes, creator productivity does not change and that the reduction in dislikes drives the observed productivity results.

Table 18: Effect on productivity of Korean and Japanese channels

	(1) ln(Videos)
Post × Female	-0.060
	[-.134,0.023]
Observations	21,000
Channels	874
R-squared	0.68
Median number of videos	9

Note: Standard errors calculated using wild cluster bootstrap method with 200 replications. 95 percent confidence intervals reported.

Estimates of the effect of the platform design change on video demand are likewise insignificant and negative (Table 19). This suggests that, in absence of a treatment effect

on dislikes, demand for creator content does not change and that the reduction in dislikes drives the observed productivity results.

Table 19: Effect on demand for content from Korean and Japanese channels

	(1) ln(Views)
Post × Female	-0.086
	[-0.214,0.041]
Observations	25,004
Channels	893
R-squared	0.77
Median views	2,981,584

Note: Standard errors calculated using wild cluster bootstrap method with 200 replications. 95 percent confidence intervals reported.

8 Conclusion

I document a large gender gap in feedback among content creators on YouTube and demonstrate that platform design changes can close this gap. The removal of public dislike counts reduced the number of dislikes on videos by women by 21 percent, while the number of dislikes on videos by men remained unchanged. Alternative specifications that control for predicted counterfactual dislikes women's videos would have received if they were produced by men, suggest that women responded to the gender gap by selecting into content categories that were less likely to attract dislikes.

I then show that the closure of the gender feedback gap is driven by a decrease in particularly large dislike counts (-35 percent at the 90th quantile). This result is consistent with a reduction in 'dislike attacks', the form of harassment that YouTube sought to remove with the platform design change.

The reduction in dislikes also improved content creator productivity and demand for their content. After the design change, women increase their video production by 8.4 percent and generate 15.5 more demand for their content.

My findings have important implications for our understanding of both platform design and gender bias in labor market settings. The findings show that it is possible to eliminate gender-biased feedback through design changes, and that this leads to positive

supply and demand effects. This indicates that other user-generated content platforms could benefit from adopting similar policies. Causal evidence on the link between the gender feedback gap and productivity highlights the importance and economic potential of addressing similar issues in different contexts.

References

Abi Adams-Prassl, Kristiina Huttunen, Emily Nix, and Ning Zhang. Violence against women at work. *The Quarterly Journal of Economics*, 139(2):937–991, 2024.

Luis Aguiar. Bad apples on rotten tomatoes: Critics, crowds, and gender bias in product ratings. *Marketing Science*, 2024. Articles in Advance, published online November 12, 2024.

Alphabet Inc. Alphabet inc. 10-k report 2023. https://abc.xyz/assets/43/44/675b83d7455885c4615d848d52a4/goog-10-k-2023.pdf, 2023. Accessed: 2024-11-06.

Linda Babcock, Maria P Recalde, Lise Vesterlund, and Laurie Weingart. Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review*, 107(3):714–747, 2017.

Abhijit V. Banerjee. A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3):797–817, 1992.

Cyprien Batut, Caroline Coly, and Sarah Schneider-Strawczynski. It's a man's world: Culture of abuse, #metoo and worker flows. 2021. URL https://eprints.lse.ac.uk/127729/1/67d405f1023e1.pdf.

George Beknazar-Yuzbashev, Rafael Jiménez-Durán, and Mateusz Stalinski. A model of harmful yet engaging content on social media. In *AEA Papers and Proceedings*, volume 114, pages 678–683, 2024.

George Beknazar-Yuzbashev, Rafael Jiménez-Durán, Jesse McCrosky, and Mateusz Stalinski. Toxic content and user engagement on social media: Evidence from a field experiment. Technical Report 11644, CESifo Working Paper, 2025.

Alan Benson, Danielle Li, and Kelly Shue. Potential and the gender promotions gap. Available at SSRN 4747175, 2024. URL https://papers.ssrn.com/abstract_id=4747175.

Marianne Bertrand, Dolly Chugh, and Sendhil Mullainathan. Implicit discrimination. *American Economic Review*, 95(2):94–98, 2005.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5): 992–1026, 1992.

Cagatay Bircan, Guido Friebel, and Tristan Stahl. Gender promotion gaps in knowledge work: The role of task assignment in teams. Technical Report 291, EBRD Working Paper, 2024. URL https://www.ebrd.com/content/dam/ebrd_dxp/assets/pdfs/office-of-the-chief-economist/working-papers/working-papers-2024/WP-291.pdf.

Peter Q. Blair and Benjamin Posmanick. Why did gender wage convergence in the united states stall? Technical Report 30821, National Bureau of Economic Research, 2023.

Ronit Bodner and Drazen Prelec. Self-signaling and diagnostic utility in everyday decision making. In Isabelle Brodscholl, George Loewenstein, and Daniel Read, editors, *The Psychology of Economic Decisions: Rationality and Well-Being*, pages 105–126. Oxford University Press, 2003.

J. Aislinn Bohren, Alex Imas, and Michael Rosenberg. The dynamics of discrimination: Theory and evidence. *American Economic Review*, 109(10):3395–3436, 2019.

Tristan L. Botelho and Marina Gertsberg. The disciplining effect of status: Evaluator status awards and observed gender bias in evaluations. *Management Science*, 68(7): 5311–5329, 2022.

Gordon Burtch, Qinglai He, Yili Hong, and Dokyun Lee. How do peer awards motivate creative content? experimental evidence from reddit. *Management Science*, 68(5):3488–3506, 2022.

Adrian Colin Cameron and Pravin K Trivedi. *Microeconometrics Using Stata*. Stata Press, 2nd edition, 2022.

Ivan A. Canay. A simple approach to quantile regression for panel data. *The Econometrics Journal*, 14(3):368–386, 2011.

Mauro Caselli, Paolo Falco, and Gianpiero Mattera. When the stadium goes silent: How crowds affect the performance of discriminated groups. *Journal of Labor Economics*, 41 (2):431–451, 2023.

Matias D Cattaneo, Michael Jansson, and Xinwei Ma. Manipulation testing based on density discontinuity. *The Stata Journal*, 18(1):234–261, 2018.

Bryan Chu, Ben Handel, Jon Kolstad, and Ulrike Malmendier. Gender differences in non-promotable tasks: The case of clinical note-taking. 2022. URL https://eml.berkeley.edu/~ulrike/Papers/Gender_Paper.pdf.

Dean Eckles, René F Kizilcec, and Eytan Bakshy. Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences*, 113(27): 7316–7322, 2016.

Florian Ederer, Paul Goldsmith-Pinkham, and Kyle Jensen. Anonymity and identity online. Available at SSRN 5095583, 2025. URL https://papers.ssrn.com/abstract_id=5095583.

Olle Folke and Johanna Rickne. Sexual harassment and gender inequality in the labor market. *The Quarterly Journal of Economics*, 137(4):2163–2212, 2022.

Nickolas Gagnon, Kristof Bosmans, and Arno Riedl. The effect of gender discrimination on labor supply. *Journal of Political Economy*, 133(3):1047–1081, 2025.

Claudia Goldin. A grand gender convergence: Its last chapter. *American Economic Review*, 104(4):1091–1119, 2014.

Claudia Goldin and Cecilia Rouse. Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, 90(4):715–741, 2000.

Chang-Tai Hsieh, Erik Hurst, Charles I. Jones, and Peter J. Klenow. The allocation of talent and us economic growth. *Econometrica*, 87(5):1439–1474, 2019.

C.J. Hutto and Eric E. Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, pages 216–225, Ann Arbor, MI, June 2014.

Alessandro Iaria, Carlo Schwarz, and Fabian Waldinger. Gender gaps in academia: Global evidence over the twentieth century. Available at SSRN 4150221, 2024. URL https://ssrn.com/abstract=4150221.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. *An Introduction to Statistical Learning: With Applications in Python*. Springer, 2023.

Rafael Jiménez-Durán. The economics of content moderation: Theory and experimental evidence from hate speech on twitter. Technical Report 324, George J. Stigler Center for the Study of the Economy & the State, 2023.

Tural Karimli. The costs of hidden workplace harassment. Available at SSRN 4618312, 2023. URL https://ssrn.com/abstract=4618312.

Erin M Kelley, Gregory V Lane, Matthew Pecenco, and Edward A Rubin. Customer discrimination in the workplace: Evidence from online sales. Technical Report w31998, National Bureau of Economic Research, 2024.

Young-Jin Lee, Kartik Hosanagar, and Yong Tan. Do i follow my friends or the crowd? information cascades in online movie ratings. *Management Science*, 61(9):2241–2258, 2015.

Yi Liu, Pinar Yildirim, and Z. John Zhang. Implications of revenue models and technology for content moderation strategies. *Marketing Science*, 41(4):831–847, 2022.

Leonardo Madio and Martin Quinn. Content moderation and advertising in social media platforms. *Journal of Economics & Management Strategy*, 34(2):342–369, 2025.

Amir Mehrjoo, Rubén Cuevas, and Ángel Cuevas. Online advertisement in a pink-colored market. *EPJ Data Science*, 13(1):36, 2024.

Friederike Mengel, Jan Sauermann, and Ulf Zölitz. Gender bias in teaching evaluations. *Journal of the European Economic Association*, 17(2):535–566, 2019.

Simha Mummalaneni, Hema Yoganarasimhan, and Varad Pathak. How do content producers respond to engagement on social media platforms? Available at SSRN 4173537, 2022. URL https://ssrn.com/abstract=4173537.

Heather Sarsons. Recognition for group work: Gender differences in academia. *American Economic Review*, 107(5):141–145, 2017.

Heather Sarsons, Klarita Gërxhani, Ernesto Reuben, and Arthur Schram. Gender differences in recognition for group work. *Journal of Political Economy*, 129(1):101–147, 2021.

Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27, 2020. doi: 10.1109/ASYU50717.2020.9259802. URL https://doi.org/10.1109/ASYU50717.2020.9259802.

Steven J Spencer, Christine Logel, and Paul G Davies. Stereotype threat. *Annual review of psychology*, 67(1):415–437, 2016.

Karthik Srinivasan. Paying attention. 2023. URL https://karthikecon.github.io/karthiksrinivasan.org/paying_attention.pdf.

Alice H. Wu. Gender bias among professionals: An identity-based interpretation. *Review of Economics and Statistics*, 102(5):867–880, 2020.

Zhiyu Zeng, Hengchen Dai, Dennis J Zhang, Heng Zhang, Renyu Zhang, Zhiwei Xu, and Zuo-Jun Max Shen. The impact of social nudges on user-generated content for social network platforms. *Management Science*, 69(9):5189–5208, 2023.

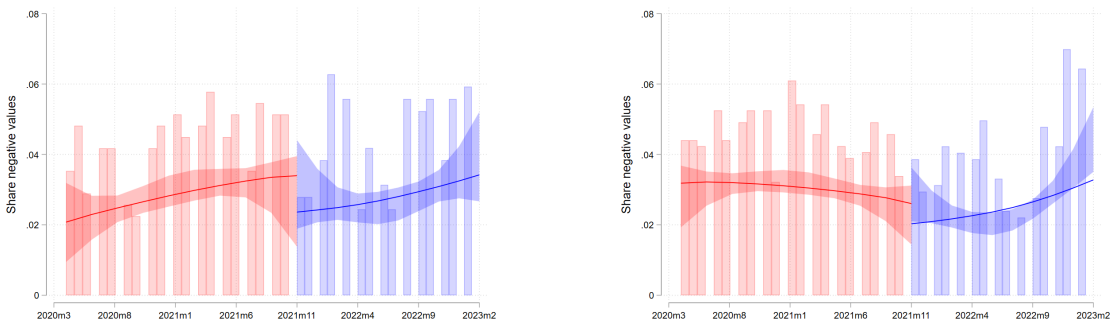
A Appendix

A.1 Data

A.1.1 Testing for Discontinuity in Video Unlisting Patterns

The panel of videos per channel over time is based on the upload playlist of each channel. This playlist contains all public and private videos over time, as well as those deleted by the channel or removed by YouTube (typically because of violating terms of service or copyright).[62] This dataset does not contain unlisted videos, which are only accessible via a link. Unlisted videos should only represent a small share of total video count and therefore generally not be a cause for concern. However, it is possible that the probability with which videos get unlisted changes over time. In particular, the removal of public dislike counts might have reduced the probability with which videos get unlisted. I am able to test this by leveraging a feature of the dataset of views per channel over time to my advantage. Specifically, videos that get unlisted/ made private / deleted / removed are removed from the view count. Thus, I can test whether the share of observations over time with negative view counts (implying removed videos) changes. I do this using a density test typically used for falsification tests as well as detection of self-selection / sorting in a regression discontinuity designs (Cattaneo et al. 2018). I run the test for women and men separately and plot the test results in Figure A1. The test suggests that there is no discontinuity around the treatment, as shown by overlapping confidence intervals.

Figure A1: Testing for discontinuity in video unlisting patterns



Female Male

[62] The information on videos that are no longer public is limited to the upload date but this suffices for most of my analysis

A.2 Heterogeneous Effects on Dislikes

Table A.1: Quantile regression estimates of platform design change on dislike behavior

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
VARIABLES	q10	q20	q30	q40	q50	q60	q70	q80	q90
Post × Female	-0.16***	-0.12***	-0.14***	-0.18***	-0.18***	-0.17***	-0.20***	-0.25***	-0.35***
	(0.05)	(0.04)	(0.03)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)
Observations	49,793	49,793	49,793	49,793	49,793	49,793	49,793	49,793	49,793

Bootstrapped standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: This specification controls for channel-country fixed effects, daily fixed effects, video age, and views per video.

Table A.2: Quantile regression estimates of platform design change on dislike behavior - controlling for gender differences in types of content produced

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
VARIABLES	q10	q20	q30	q40	q50	q60	q70	q80	q90
Post × Female	-0.03	-0.12***	-0.17***	-0.23***	-0.21***	-0.21***	-0.21***	-0.30***	-0.34***
	(0.06)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)
ln(Predicted dislikes)	0.12***	0.12***	0.11***	0.10***	0.10***	0.10***	0.10***	0.09***	0.09***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Observations	49,793	49,793	49,793	49,793	49,793	49,793	49,793	49,793	49,793

Bootstrapped standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: This specification controls for predicted dislikes, channel-country fixed effects, daily fixed effects, video age, and views per video.

Table A.3: Quantile regression estimates of platform design change on dislike behavior - accounting for selection into trending videos

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
VARIABLES	q10	q20	q30	q40	q50	q60	q70	q80	q90
Post × Female	-0.13**	-0.10***	-0.12***	-0.16***	-0.19***	-0.21***	-0.27***	-0.34***	-0.29***
	(0.06)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.03)	(0.03)	(0.05)
ln(Predicted dislikes)	0.16***	0.14***	0.13***	0.12***	0.11***	0.10***	0.10***	0.09***	0.09***
	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Observations	18,263	18,263	18,263	18,263	18,263	18,263	18,263	18,263	18,263

Bootstrapped standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: This specification controls for predicted dislikes, channel-country fixed effects, daily fixed effects, video age, and views per video.

Table A.4: Quantile regression estimates of platform design change on dislike behavior - treatment effects for men

VARIABLES	(1) q10	(2) q20	(3) q30	(4) q40	(5) q50	(6) q60	(7) q70	(8) q80	(9) q90
Post	-0.111***	-0.0722***	-0.0595***	-0.0664***	-0.0841***	-0.109***	-0.130***	-0.157***	-0.167***
	(0.0193)	(0.0155)	(0.00960)	(0.00741)	(0.00707)	(0.00761)	(0.00889)	(0.00962)	(0.0185)
Post × Female	-0.0813	-0.117***	-0.147***	-0.184***	-0.194***	-0.209***	-0.215***	-0.272***	-0.293***
	(0.0549)	(0.0359)	(0.0245)	(0.0178)	(0.0189)	(0.0162)	(0.0202)	(0.0255)	(0.0407)
ln(Predicted dislikes)	0.121***	0.115***	0.108***	0.105***	0.106***	0.103***	0.0990***	0.0968***	0.0966***
	(0.00358)	(0.00175)	(0.00167)	(0.00121)	(0.00108)	(0.00119)	(0.00140)	(0.00127)	(0.00202)
Observations	49,793	49,793	49,793	49,793	49,793	49,793	49,793	49,793	49,793

Bootstrapped standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: This specification controls for predicted dislikes, channel-country fixed effects, video age, and views per video.

Table A.5: Quantile regression estimates of platform design change on dislike behavior: Gender gap in feedback before treatment

VARIABLES	(1) q10	(2) q20	(3) q30	(4) q40	(5) q50	(6) q60	(7) q70	(8) q80	(9) q90
Female	-0.0542**	-0.0100	-0.0176	-0.00417	0.00433	0.0563***	0.0958***	0.180***	0.350***
	(0.0237)	(0.0185)	(0.0139)	(0.0147)	(0.0120)	(0.0129)	(0.0130)	(0.0158)	(0.0235)
Observations	46,407	46,407	46,407	46,407	46,407	46,407	46,407	46,407	46,407

Bootstrapped standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: This specification controls for predicted dislikes, monthly fixed effects, video age, views per video, and category fixed effects.

Table A.6: Quantile regression estimates of platform design change on dislike behavior: Gender gap in feedback after treatment

VARIABLES	(1) q10	(2) q20	(3) q30	(4) q40	(5) q50	(6) q60	(7) q70	(8) q80	(9) q90
Female	-0.138*	-0.0115	-0.123*	-0.184***	-0.117**	-0.0861	-0.0518	-0.00529	0.0176
	(0.0820)	(0.0712)	(0.0642)	(0.0525)	(0.0457)	(0.0532)	(0.0543)	(0.0588)	(0.0635)
Observations	3,386	3,386	3,386	3,386	3,386	3,386	3,386	3,386	3,386

Bootstrapped standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: This specification controls for predicted dislikes, monthly fixed effects, video age, views per video, and category fixed effects.

A.3 Effect on Dislikes in 5 Country Sample

Table A.7: Effect of platform design change on dislikes - 5 country sample

VARIABLES	(1) ln(Dislikes)	(2) ln(Dislikes)	(3) ln(Dislikes)
Post × Female	-0.33***	-0.25***	-0.25***
	(0.09)	(0.07)	(0.07)
Constant	6.30***	5.73***	5.73***
	(0.00)	(0.01)	(0.01)
Observations	135,135	135,135	135,135
R-squared	0.82	0.86	0.86
YT Channel-Country IDs	6974	6974	6974
YT Channels	3986	3986	3986
Channel-Country FE	✓	✓	✓
Daily FE	✓	✓	✓
1000 views		✓	✓
Video age in hours		✓	✓
Video cat. FE			✓

Standard errors clustered by channel
*** p<0.01, ** p<0.05, * p<0.1

Table A.8: Accounting for selection into trending videos - 5 country sample

VARIABLES	(1) ln(Dislikes)	(2) ln(Dislikes)	(3) ln(Dislikes)
Post × Female	-0.32***	-0.24***	-0.24***
	(0.09)	(0.06)	(0.06)
Constant	6.28***	5.72***	5.72***
	(0.00)	(0.02)	(0.02)
Observations	58,054	58,054	58,054
R-squared	0.78	0.83	0.84
YT Channel-Country IDs	1173	1173	1173
YT Channels	804	804	804
Channel-Country FE	✓	✓	✓
Daily FE	✓	✓	✓
1000 views		✓	✓
Video age in hours		✓	✓
Video cat. FE			✓

Standard errors clustered by channel
*** p<0.01, ** p<0.05, * p<0.1

Table A.9: Treatment effects for men - 5 country sample

VARIABLES	(1) ln(Dislikes)	(2) ln(Dislikes)	(3) ln(Dislikes)
Post × Female	-0.33***	-0.26***	-0.25***
	(0.08)	(0.07)	(0.07)
Constant	6.31***	5.74***	5.74***
	(0.00)	(0.01)	(0.01)
Observations	135,135	135,135	135,135
R-squared	0.81	0.86	0.86
YT Channel-Country IDs	6974	6974	6974
YT Channels	3986	3986	3986
Channel-trending country FE	✓	✓	✓
1000 views		✓	✓
Video age in hours		✓	✓
Video cat. FE			✓
Channel cat. FE			✓

Standard errors clustered by channel
*** p<0.01, ** p<0.05, * p<0.1

Table A.10: Gender gap in feedback before and after treatment - 5 country sample

VARIABLES	(1) ln(Dislikes)	(2) ln(Dislikes)
Female	0.19**	0.03
	(0.08)	(0.10)
Observations	125,851	9,244
R-squared	0.39	0.51
YT Channels	3810	997
Daily FE	✓	✓
1000 views	✓	✓
Video age in hours	✓	✓
Video cat. FE	✓	✓
Channel cat. FE	✓	✓

Standard errors clustered by channel
*** p<0.01, ** p<0.05, * p<0.1

A.4 Heterogeneous Effects on Dislikes: 5 country sample

Figure A2: Comparison of dislike distribution pre- and post-treatment by gender: 5 country sample

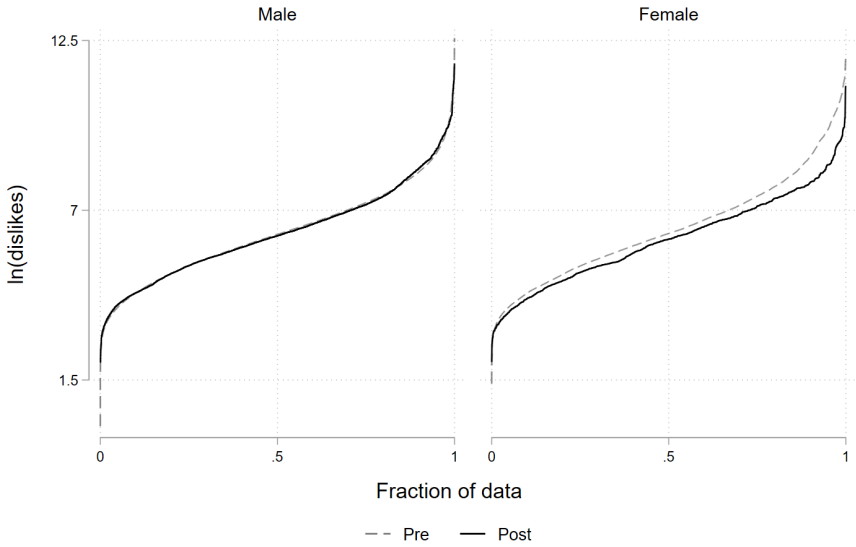


Figure A3: Quantile regression estimates of platform design change on dislikes: 5 country sample

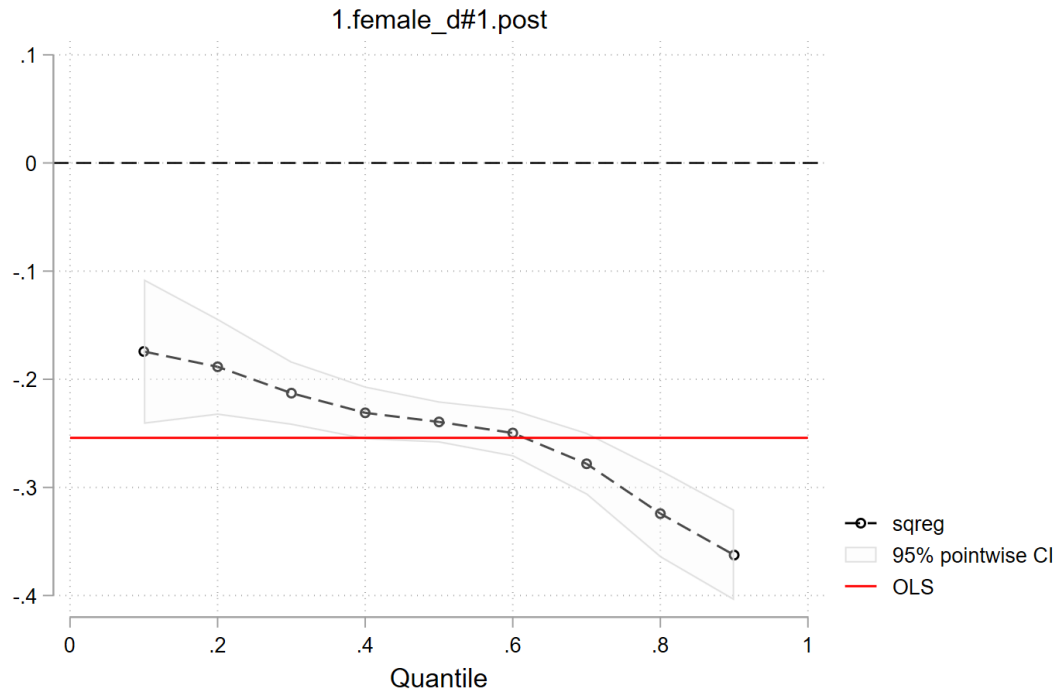


Figure A4: Quantile regression estimates: Treatment effect for men and women in 5 country sample

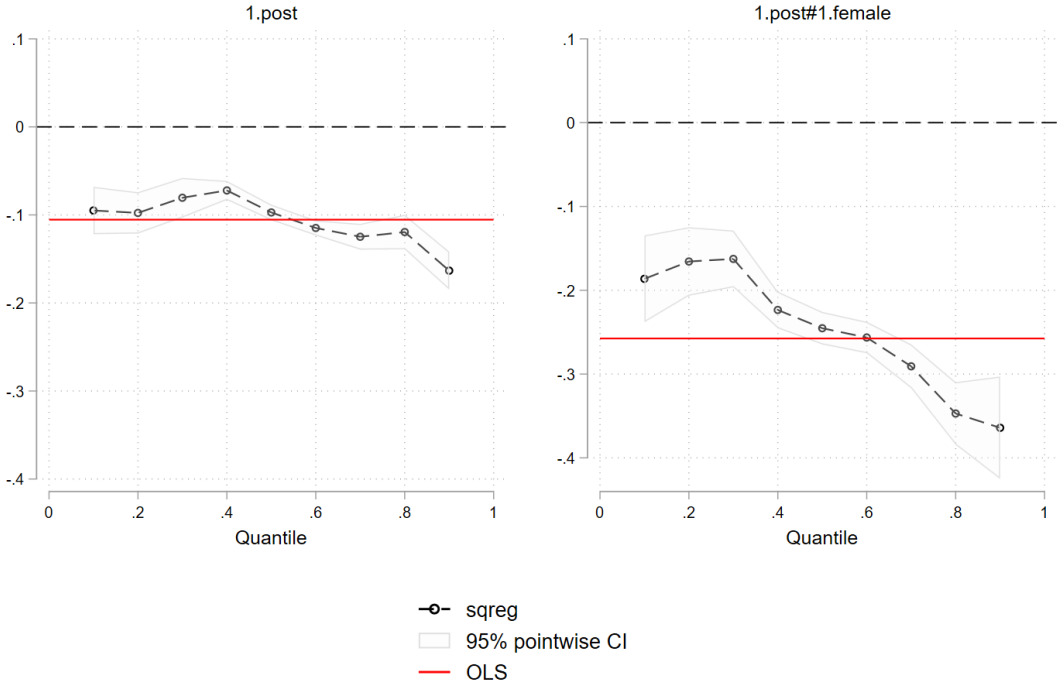
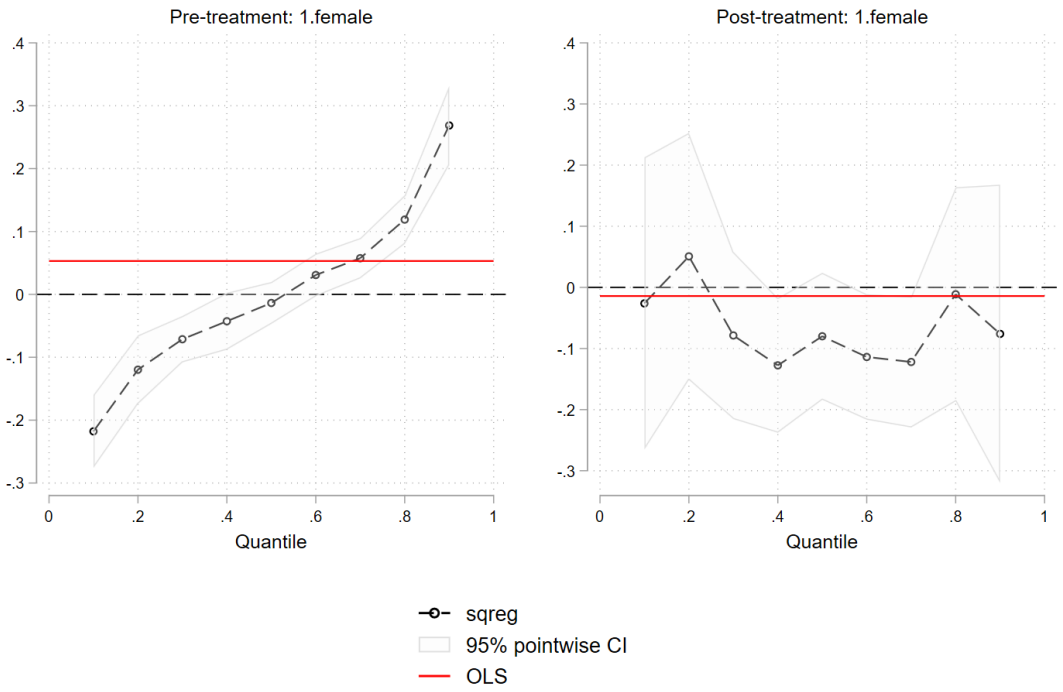


Figure A5: Quantile regression estimates: Gender gap in feedback before and after treatment in 5 country sample



A.5 Spillover Effects on Comments: Comment Toxicity

Table A.11: Effect on the mean comment toxicity $[0, 1]$

	(1)	(2)	(3)
VARIABLES	Mean toxicity	Mean toxicity	Mean toxicity
Female	-0.04***		
	(0.01)		
Post × Female	0.01	0.00	-0.01
	(0.01)	(0.01)	(0.01)
Observations	1,979,192	9,657	14,794
R-squared	0.02	0.27	0.33
Channels		317	358
Video upload date FE	✓	✓	
Comment publishing date FE	✓		✓
Channel FE		✓	✓
Level of analysis	Comment	Video	Channel

Standard errors clustered by channel
*** p<0.01, ** p<0.05, * p<0.1

Table A.12: Effect on comment sentiment at the tails: comment toxicity $[0, 1]$

	(1)	(2)	(3)	(4)
VARIABLES	p90 (neg.)	sentiment	p10 (posit.)	sentiment
Post × Female	-0.00	-0.01	0.00	-0.01
	(0.02)	(0.01)	(0.01)	(0.00)
Constant	0.32***	0.33***	0.03***	0.03***
	(0.00)	(0.00)	(0.00)	(0.00)
Observations	9,657	14,794	9,657	14,794
R-squared	0.32	0.35	0.12	0.12
Channels	317	358	317	358
Upload date FE	✓		✓	
Channel FE	✓	✓	✓	✓
Comment publ. date FE		✓		✓
Level of analysis	Video	Channel	Video	Channel

Standard errors clustered by channel
*** p<0.01, ** p<0.05, * p<0.1

A.6 Other Engagement Metrics

Table A.13: Effect of platform design change on likes

VARIABLES	(1) ln(Likes)	(2) ln(Likes)	(3) ln(Likes)
Post × Female	-0.22*	-0.14	-0.14
	(0.12)	(0.09)	(0.09)
Observations	49,759	49,759	49,759
R-squared	0.81	0.87	0.87
YT Channel-Country IDs	2864	2864	2864
YT Channels	1808	1808	1808
Channel-Country FE	✓	✓	✓
Daily FE	✓	✓	✓
1000 views		✓	✓
Video age in hours		✓	✓
Video cat. FE			✓

Standard errors clustered by channel
*** p<0.01, ** p<0.05, * p<0.1

Table A.14: Effect of platform design change on comments

VARIABLES	(1) ln(Comments)	(2) ln(Comments)	(3) ln(Comments)
Post × Female	-0.29**	-0.21*	-0.22**
	(0.13)	(0.11)	(0.11)
Observations	49,737	49,737	49,737
R-squared	0.79	0.83	0.83
YT Channel-Country IDs	2863	2863	2863
YT Channels	1807	1807	1807
Channel-Country FE	✓	✓	✓
Daily FE	✓	✓	✓
1000 views		✓	✓
Video age in hours		✓	✓
Video cat. FE			✓

Standard errors clustered by channel
*** p<0.01, ** p<0.05, * p<0.1