Gender Bias, Feedback, and Productivity

Marita Freimane*

November 19, 2024

Job Market Paper

Click here for the latest version

Abstract

I explore how gender biased feedback affects the productivity of workers in an online labor market. Using a design change on YouTube where the platform removed public displays of how often a video has been disliked, I show that — while dislike counts were public — female content creators received significantly more negative feedback on comparable content than male content creators. This gender gap in negative feedback is eliminated after the design change. Using detailed video- and channel- level data and a fuzzy difference-in-differences identification strategy, I show that the removal of excess negative feedback significantly and persistently increased the productivity of female content creators and consumer demand for their content. Relative to men, women produce 8.4 percent more videos after the platform design change. The increase in productivity coincides with an even larger increase of 15.5 percent in demand for content produced by women. Investigating mechanisms, I show that the reduction in negative feedback is primarily driven by changes in the upper tail of the distribution of dislikes and is consistent with the platform's objective of reducing harassment through 'dislike attacks'. Finally, I show that there are limited spillover effects on toxicity in other feedback channels and provide evidence from a placebo-test to confirm that productivity effects are indeed driven by the reduction in dislikes.

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, 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 last century, women have entered the labor market at increasing rates, narrowing gender employment and wage gaps (Goldin 2014). This shift has driven substantial gains in aggregate productivity (Hsieh et al. 2019). Yet progress has slowed, and gender gaps persist (Blair and Posmanick 2023). A growing body of literature studies the drivers behind remaining disparities, including differences in how men and women are treated at work: Differences in the recognition of contributions to group work (Sarsons 2017; Sarsons et al. 2021), bias in subjective assessments of women's potential (Benson et al. 2024), differences in assignments to 'non-promotable' tasks (Babcock et al. 2017; Chu et al. 2022) versus 'promotable' tasks (Bircan et al. 2024), and higher exposure of women to workplace harassment (Folke and Rickne 2022).

In this paper, I explore another aspect of how men and women are treated differently in the workplace: biased feedback. Gender differences in feedback are difficult to observe in traditional offline workplaces but they have been documented in online settings (Botelho and Gertsberg 2021; Aguiar 2024).[1] I study the impact of biased feedback on productivity. Specifically, I investigate online work as a setting where both feedback and productivity is directly observable and quantifiable. My paper focuses on content creators, an important and growing segment of the online labor market.[2] Online content creators are directly exposed to consumer feedback and make independent decisions about their labor supply. This provides an ideal setting to examine the effects of biased feedback on performance and potential earnings.

My empirical analysis reveals three major findings: First, I document a significant gender gap in negative feedback on an online user-generated content platform and this gap remains significant conditional on gender differences in the types of content produced. Second, I provide causal evidence that this gender gap can be eliminated by redesigning the platform's reputation system. Third, I show that the elimination of the gender feedback gap causes an increase in both content supply and demand on the platform. To identify these effects, I leverage a platform design change on YouTube. Consumers on YouTube can

[1] Related literature on gender bias in evaluations identifies gaps in various contexts, including orchestra auditions (Goldin and Rouse 2000) and teaching evaluations of female faculty (Mengel et al. 2019).

[2] There are an estimated 50 million content creators worldwide, including about 2 million engaged in full-time work as content creators. https://signalfire.com/blog/creator-economy/

provide positive (like) and negative (dislike) feedback on video content. In 2021, YouTube suddenly removed the public display of how many times a video had been disliked. I generate a unique dataset based on information from the YouTube API that allows me to observe how consumers adjusted their dislike behavior after the number of dislikes is no longer displayed on the platform. This data allows me to identify the share of dislikes that were attributable to dislikes being visible to other users of the platform.

A comparison of per-video dislike counts before and after the dislikes were made private indicates that there was a large gender gap in dislikes before the design change. Female content creators received 43 percent more dislikes than male content creators prior to the design change and experienced a 57 percent decrease in negative feedback once dislike counts are removed from public display. Difference-in-differences estimates confirm this effect: relative to men, women experience a 21 percent decrease in dislikes after the design change.[3] Next, I show that this result is not driven by gender differences in the type of content created. To account for gender differences in content production, I train a Lasso-based machine learning model that predicts how many dislikes women's videos would have received if they were produced by men instead. Including these predicted dislikes in my difference-in-differences regressions, I estimate a 24 percent decrease in dislikes for women's content caused by the design change. The increased point estimate, conditional on how 'dislikable' content would be if produced by men, indicates that women might already have selected into content categories less susceptible to excessive negative feedback. To estimate the total effect on the gender gap in feedback, I estimate the gender gap separately for my pre-treatment and post-treatment sample. The results indicate that the platform design change eliminated the gender gap in feedback.

What are the effects of biased feedback on productivity? To study the causal effect of removing excess negative feedback on content supply and demand, I estimate a fuzzy difference-in-differences regression (De Chaisemartin and d'Haultfoeuille 2018) that compares the changes in content production and demand for women's content to those for men's content following the platform design change.[4] Difference-in-differences estimates of

[3] Event study estimates show that the channels of both female and male content creators followed parallel trends before the design change, making them comparable.

[4] A fuzzy difference-in-differences estimation compares a more treated group (female content creators) to a less treated group (male content creators). The main identifying assumption of a fuzzy difference-in-differences specification is that the effect of the treatment is the same in the treatment and in the control group.

content creator productivity show that women increase their content supply by 8.4 percent after the design change. The median content creator in my sample generates 0.4 additional videos per month. Corresponding difference-in-differences estimates on video viewership show that user demand for women's content increases by 15.5 percent. This effect is notably larger than the increase in supply and implies an average increase in views per video. A back-of-the-envelope calculation suggests that the median female content creator in my sample experiences a revenue increase of up to $3,100 per month after the design change.[5]

Next, I investigate the mechanisms driving the changes in content creator productivity. First, I explore the heterogeneity of the treatment effect on dislikes along the dislike distribution. To implement this analysis, I estimate quantile regressions. I find that, along the entire dislike distribution, women experience a larger reduction in dislikes than men. This effect, however, is particularly pronounced in the upper tail of the dislike distribution. At the 90th percentile, women see a 35 percent greater decrease in dislikes than men. This result is consistent with YouTube's objective of implementing the design change to reduce harassment through 'dislike attacks' on the platform.[6] Dislike attacks are commonly defined as a form of harassment where users purposefully dislike a content creator's video, independently of its quality, after observing how many others have disliked the video before.[7]

Second, I examine the extent to which negative feedback, previously contained in dislikes, may have spilled over to comments. Specifically, I use natural language processing algorithms to compare the text sentiment of comments before and after the design change to analyze whether comments have become more toxic. My analysis indicates limited spillover effects on comments. If anything, after the removal of public dislike counts, the most negative comments become slightly less negative.

Third, I investigate whether the platform design change affected video quality. My findings suggest that the treatment effect on demand exceeds the effect on productivity in relative magnitude. Similarly, improvements in comment sentiment suggest that video quality did not decrease, and may even have increased. I present an additional piece of evidence: predicted dislikes, which are based on dislikes by men prior to treatment and

[5] Based on revenue estimates of $0.25 - $4.00 per 1000 views by Socialblade.com
[6] https://blog.youtube/news-and-events/update-to-youtube/
[7] https://www.urbandictionary.com/define.php?term=dislike-bombed

hence are a measure of video quality not affected by gender bias, decrease for women following the design change.

Fourth, I demonstrate that the observed results are not driven by general supply and demand shifts that are unrelated to negative feedback through dislikes. Specifically, I conduct a placebo test based on YouTube content creators in Korea and Japan, where I observe no gender gap in dislike counts. I find that in this setting, where the platform design change does not affect dislike behavior, there is also no detectable impact on video supply and demand.

Related literature. This paper contributes to four strands of literature in economics. First, I study gender disparities in the labor market, focusing on how the differential treatment of male and female workers impacts their productivity. Specifically, I show that gender biased feedback directly impacts worker productivity in a setting focused on online content creators, a group of workers whose feedback exposure and productivity levels are readily observable and quantifiable. In an offline workplace setting, Benson et al. (2024) document that women are rated lower than men in subjective assessments of their future potential as employees, in spite of better current performance. The authors show that differences in estimated potential account for half the gender promotion gap and induces talent misallocation. Bircan et al. (2024) study a similar bias in task allocation: They show that women supervised by male directors are significantly less likely to be assigned to promotable team leadership roles, having to wait half a year longer for their first leadership role than their male peers. Conversely, Babcock et al. (2017) show that women are more likely to receive and accept requests for 'non-promotable' tasks. Sarsons (2017); Sarsons et al. (2021) document a gender disparity in academic tenure outcomes related to co-authoring. Women are less likely to receive tenure when they co-author, especially with men, while men's co-authoring choices do not affect their tenure rates. Experiments in Sarsons et al. (2021) show that this is driven by women receiving less credit for their contributions to group work under uncertainty about relative contribution of group members. Kelley et al. (2024) show that gender discrimination at the workplace can also originate from customers. In a field experiment, they randomize worker names in an online travel agency and find that female names lead to reduced customer purchases, both on the extensive and intensive margins.

Second, I contribute to literature on the economic cost of toxic behavior at workplace. Specifically, I demonstrate that exposure to toxicity in the form of excess negative feedback from 'dislike attacks' can reduce worker productivity, even when this feedback does not meet the salience levels defined by legal thresholds for harassment or discrimination. Caselli et al. (2023) examine the effect of harassment by football spectators on football players performance using variation from COVID-19 lock-downs in Italy. The authors find that African players improve their performance when spectators are banned during lock-downs, with an 11 percent improvement for those previously directly targeted by verbal harassment and racial slurs from football crowds. Karimli (2023) finds that one standard deviation reduction in workplace harassment results in 7 percent increase in patenting among inventors, using evidence from a staggered roll-out of changes to the rules governing the use of non-disclosure agreements (NDAs) in workplace harassment cases for US firms. Folke and Rickne (2022) find that both women and men face more sexual harassment in occupations dominated by the opposite gender, and that these effects contribute to occupational gender segregation. The sorting response to harassment entails a 10 percent wage penalty for women as they sort into lower-paying jobs. Batut et al. (2021) examine the labor market impact of increased salience of workplace harassment due to the #MeToo movement. They find that salience leads to an increased quit-rate among female workers and that these women sort into lower-paying jobs where they incur a 2 percent wage penalty. Gagnon et al. (Forthcoming) show that gender inequality reduces work morale among female workers: in their experimental setting female workers reduce their labor supply by 0.27 standard deviations when a offered a lower piece-rate wage than men. Adams-Prassl et al. (2022) find that men are more often perpetrators of workplace violence and face fewer repercussions for attacks on female colleagues due to economic power imbalances, which reduces female retention and hiring, especially in firms with male managers. Firms with female managers differ in one key way: they are much less likely to retain perpetrators after such incidents.

Third, I contribute to literature on the production of user-generated content[8]. Specifically, I show that toxicity, in form of excess negative feedback, negatively impacts content supply and demand. My study investigates a setting where content producers and con-

[8]Aridor et al. (2024) provide a recent review of literature on the economics of social media, the main source of user-generated content.

sumers have distinct roles, complementing previous empirical studies that primarily focus on platforms with a single user type that both creates and consumes content.[9] In a field experiment on Twitter, Jiménez-Durán (2023) finds that reporting a toxic post increases the likelihood of the post's removal and increases participation rates among users targeted by the toxic content. In another field experiment, Beknazar-Yuzbashev et al. (2022) hide toxic content on both Twitter and Facebook, and find that this manipulation leads to a drop in content consumption on both platforms, and a decrease in ad consumption on Twitter (where this metric was available).[10] Their finding suggests a potential trade-off between increasing participation from those targeted by toxic content and maximizing overall content consumption and ad revenues. Theoretical approaches further highlights these trade-offs: Madio and Quinn (2023) explicitly model trade-offs between brand safety concerns for advertisers and the platforms objective to maximize consumer participation. Beknazar-Yuzbashev et al. (2024) highlight that advertising-based business models can misalign incentives between social media platforms and users when they incentivize platforms to display toxic content because the toxic content is sufficiently engaging and motivates users to engage and spend more time online. Liu et al. (2022) show that a platform's optimal degree of moderation depends on its revenue model. Ad-financed platforms are more likely to moderate content than subscription-financed platforms, but do so at a lower rate to maximize consumer engagement.

More broadly, studies suggest that both non-monetary and monetary incentives are important drivers of content creation. Empirical studies have documented that positive feedback (including likes, comments, nudges and badges) increase content production on Facebook (Eckles et al. 2016), Twitter (Mummalaneni et al. 2022), Reddit (Burtch et al. 2022; Srinivasan 2023), and a Chinese video-sharing platform (Zeng et al. 2023). Monetary incentives, such as ad-revenue sharing programs on YouTube and other ad-financed platforms, similarly increase content creation. Kerkhof (2024) shows that increased ad intensity on YouTube leads to more content differentiation as ads, considered a nuisance by most consumers, act as a *de facto* price on content consumption. El-Komboz et al. (2023) show that access restrictions to YouTube's Partnership Program reduced content supply by marginal content creators just below the new size thresholds for monetization. Andres

[9] Furthermore, watching a video is not a pre-requisite for disliking it.

[10] Beknazar-Yuzbashev et al. (2022) also conducted the experiment on YouTube but found no effect, likely because the intervention targeted comments rather than videos themselves.

et al. (2023) show that stricter moderation policies for monetizable content on YouTube lead to an increase in the amount of subscriber-only content on Patreon, and subsequently raised the average toxicity level of the posted content.[11] Johnson et al. (2023) demonstrate that the removal of content personalization settings for children resulted in a decline in both the production and consumption of children's content, reflecting a loss of personalized revenues and poorer matches between consumers and relevant content.

Fourth, I contribute to literature on online reputation systems[12] by estimating the causal effects of gender bias in ratings on content supply and demand, and by demonstrating how platform design can mitigate these biases. Botelho and Gertsberg (2021) document that reviewers on Yelp rate restaurants 0.13 stars *lower* when served by a woman. Receiving an "Elite" reviewer status reduces this gap by more than half, and this effect is driven by a decrease in extreme 1-star ratings. Aguiar (2024) examines gender bias in crowd-sourced movie ratings on IMDb and Rotten Tomatoes and finds that ratings are significantly lower for movies with a female lead.[13] These lower ratings are primarily driven by a specific group of reviewers: non-verified and predominantly male, suggesting that verification successfully mitigates gender bias.[14] In an experiment on Stack Exchange, Bohren et al. (2019) investigate dynamic gender discrimination in reviews. Exogenously assigning gender and review histories to accounts, they find that women without prior reviews face significant discrimination but that after a sequence of positive reviews, the direction of discrimination reverses. The authors interpret their findings as evidence of *belief-based* discrimination, which can be alleviated through the updating of beliefs about women's abilities. Finally, by showing that higher dislike counts for women are driven by 'dislike attacks,' my paper contributes to the literature on review bias caused by information cascades (Bikhchandani et al. 1992) and herding (Banerjee 1992), that shows how biases in sequential decision-making can lead to an inefficient equilibrium. Lee et al. (2015) studies herding in movie reviews, and finds that reviewers are less likely to deviate from the crowd for more popular

[11] In 2017, YouTube restricted access to its partner program and introduced stricter moderation policies to increase the oversight of monetized content, in response to an exodus of advertisers following reports that their ads were being shown along xenophobic and extremist content. This episode is known as YouTube "Adpocalypse". https://www.washingtonpost.com/technology/interactive/2023/influencers-timeline-money-blogs-tiktok-youtube/

[12] Pocchiari et al. (2024) provide a recent review that integrates research on ratings and reviews from economics, marketing, and related fields.

[13] Aguiar (2024) compares crowd-sourced and expert movie reviews, finding no differences by gender in the latter.

[14] Rotten Tomatoes enables verified reviews through the upload of a movie ticket.

movies. Deng et al. (2022) illustrate how editorial reviews influence subsequent user reviews and Park et al. (2021) show that the first review of a product has a disproportionate impact on quantity and valence of future reviews.

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 photos, videos, blogs, podcasts, and social media posts — primarily on digital platforms. Industry estimates value the creator economy at $250 billion,[15] encompassing around 50 million creators worldwide.[16] Of these creators around 2 million — approximately 4 percent — are professional content creators for whom content creation serves as their full-time job. The creator economy exemplifies a superstar economy, characterized by a few large and dominant creators alongside a long tail of smaller contributors. The primary income sources for content creators are influencer marketing and ad revenue sharing programs. Influencer marketing, which includes brand deals and affiliate content, primarily targets established creators with larger subscriber bases. In contrast, platform ad revenue sharing programs are available to both large and small creators, enabling them to monetize their content based on viewership and engagement levels.[17] According to industry estimates, there are approximately 1 million full-time content creators on YouTube, 500,000 on Instagram,

[15] This estimate is provided by Goldman Sachs and suggests ad-sharing, brand deals or starting an own brand, affiliated links, courses as some of the revenue sources. https://www.goldmansachs.com/intelligence/pages/the-creator-economy-could-approach-half-a-trillion-dollars-by-2027.html The creator economy as a separate industry is generally missing from official statistics https://www.washingtonpost.com/technology/2023/10/26/creator-economy-influencers-youtubers-social-media/

[16] https://signalfire.com/blog/creator-economy/

[17] https://www.goldmansachs.com/intelligence/pages/the-creator-economy-could-approach-half-a-trillion-dollars-by-2027.html

300,000 on Twitch, and 200,000 across other content-hosting platforms.[18]

2.2 YouTube

After Google Search, YouTube is the world's second most visited website, receiving more than 30 billion visits each month.[19] It is also at the core of the growing content-creator industry. YouTube hosts 1 million professional and around 12 million amateur content creators.[20] By its own estimates, YouTube has created over 800,000 full-time equivalent jobs across Australia, Brazil, Canada, Japan, South Korea, and the United States in 2020.[21]

80 percent of content creators on YouTube are male.[22] The gender distribution of content consumers is more balanced, with 60 percent being male.[23] YouTube's content consumers are younger, more educated, and have higher incomes compared with the general population.[24] The main competitors for YouTube, according to its CEO, are streaming platforms like Netflix.[25][26] Online advertising markets are intransparent, with little publicly available information on cost of ads. Interestingly, however, research suggests ads targeting women are more expensive than those targeting men (Mehrjoo et al. 2024).

Advertising is YouTube's main source of revenues.[27] In 2023, YouTube earned $31.5 billion from ads (Inc. 2023). Creators with at least 1,000 subscribers and 4,000 watch hours in the past 12 months can apply to join the YouTube Partner Program, allowing

[18] Corresponding estimates for amateur content creators indicate there are approximately 30 million on Instagram, 12 million on YouTube, 2.7 million on Twitch, and 2 million on other platforms. While TikTok has a substantial user base, it is not included in these estimates; other sources suggest there are around 1 million creators on TikTok. https://explodingtopics.com/blog/tiktok-creator-stats

[19] Ranking: https://www.similarweb.com/top-websites/ Visits to YouTube: www.similarweb.com/website/youtube.com/#traffic (July 2024). The average visit lasts around 20 minutes, longer than visits to other large websites. https://www.similarweb.com/top-websites/

[20] https://signalfire.com/blog/creator-economy/ Amateur content creators are defined as those with between 100 and 10,000 subscribers; professional content creators are defined as those having more than 10,000 subscribers.

[21] https://blog.youtube/inside-youtube/letter-susan-our-2022-priorities/

[22] This is an estimate based on the sample of content creators used in this paper.

[23] www.similarweb.com/website/youtube.com/#geography (July 2024).

[24] https://backlinko.com/youtube-users

[25] https://www.washingtonpost.com/technology/2024/02/06/you-tube-creators-earnings/

[26] This is also supported by statistics showing that, similar to Netflix and other streaming platforms, mobile devices are the leading source of traffic at 63 percent, followed by TV at 14 percent, and desktop only coming in the third place at 12 percent. Source: https://backlinko.com/youtube-users

[27] https://www.youtube.com/howyoutubeworks/our-commitments/sharing-revenue/ While YouTube has introduced several new revenue streams in recent years—most prominently subscriptions—advertising remains its primary source of revenue. Other revenue sources include: paid digital goods such as "Super Chat," "Super Stickers," and "Super Thanks" (used for engagement with creators during live streams); memberships, merchandise, ticketing, and various direct funds from YouTube. See: https://blog.youtube/news-and-events/10-ways-monetize-youtube/

them to share in YouTube's revenue.[28] Advertising revenue depends on the number of ad impressions, with creators usually receiving 55 percent of the revenue.[29] Since ads are shown alongside content, ad impressions and revenue closely correlate with the number of views each video receives.[30]

3 Platform Design Change

3.1 Removal of Public Dislike Counts

YouTube announced an immediate removal of public dislike counts for videos from the platform on November 10th, 2021. While YouTube removed public dislike counts on the website from that day, dislike counts were still available through the YouTube API for another 34 days, until December 13th, 2021. I exploit this data to identify how dislike behavior changed once dislike counts were no longer public. While consumers were no longer able to see how often a video had already been disliked, they were still able to click on the dislike button (Figure 1).

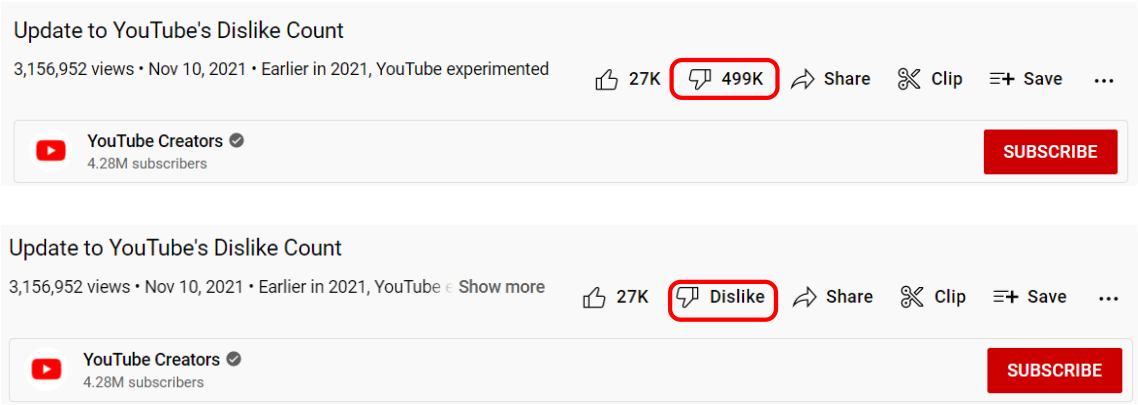


Figure 1: Above: dislikes still visible. Below: dislikes no longer visible

YouTube motivated the removal of public dislike counts with misuse of dislike button for a form of harassment referred to as 'dislike attacks' or 'dislike bombs'. YouTube de-

[28] Alternatively, creators could qualify by achieving 10 million views on YouTube Shorts within the past 90 days. https://support.google.com/youtube/answer/72851?hl=en. However, while YouTube Shorts have been rapidly growing since their introduction in September 2020, they represented a relatively small portion of total platform views during the period considered in this study. https://photutorial.com/youtube-shorts-statistics/#you-tube-shorts-receive-over-70-billion-daily-views

[29] https://blog.hootsuite.com/how-much-does-youtube-pay-per-view

[30] Views are counted when a video has been watched for at least 30 seconds, or for the entire duration if the video is shorter than that.

fines this harassment as a large number of people driving up the number of dislikes for a video unrelated to the inherent quality of the video.[31] Together with the announcement, YouTube also reported on the results of an A/B test in which they found that dislikes attacks decreased following the hiding of the dislike count, thus justifying the design change.

Dislike attacks are closely related to review bombs, a similar misuse of review systems where consumers leave a large number of very negative reviews. Like dislike attacks, this misuse of feedback tools is not unique to YouTube; other platforms also implement measures to limit such behavior. Goodreads, a book review platform, introduced temporary limits on ratings and reviews for books during times of unusual activity that suggests review bombing.[32] IMDb, a movie review platform, applies an alternate weighting to reviews and introduced a disclaimer to highlight suspicious activity that might constitute review bombing.[33] Steam, a video game platform, rather than hiding or altering the review score, displays a full histogram of positive to negative reviews over time.[34]

3.2 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.[35] Consumers have several incentives to leave feedback on YouTube videos. First, feedback on content consumed informs YouTube's algorithm for future *personalized* recommendations.[36] 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.

[31] According to Urban Dictionary, 'dislike bombs' typically occur within a short period, leading to dislikes exceeding likes by a significant margin. https://www.urbandictionary.com/define.php?term=dislike-bombed

[32] 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

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

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

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

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

3.3 Incentives to Dislike Without Public 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, the design change also prevents dislike attacks, as consumers can no longer see how many others have disliked a video. This is because the dislike count previously served as an implicit coordination tool that enabled this form of herding. Because dislike attacks are characterized by extreme numbers of dislikes, this change is likely to disproportionately affect 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.

3.4 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 (2021) 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.

3.5 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,[37] 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.

[37] 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 Data

This study uses three primary datasets. The first dataset is a repeated cross-section of trending videos from Canada and the United States, providing a daily snapshot of engagement metrics—including the number of likes, dislikes, comments, and views—for these trending videos. From this dataset, I compile a list of channels, which I then use to retrieve further information. The second dataset is a balanced panel of all uploaded videos over time for these channels, allowing me to measure changes in content supply. The third dataset is a balanced panel of channel views over time, allowing me to measure changes in demand.

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 in Canada and the United States, spanning the period from August 12, 2020, to October 23, 2022.[38] Dislikes, however, are only recorded for 34 days after the platform design change. I therefore end the sample for my analysis of effects on dislikes on December 13, 2021. As 99 percent of channels in my data appear among trending videos more than once, I generate an unbalanced panel based on the cross-sectional data with limited information loss.[39]

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. In addition, I observe the video title, description, tags—keywords that describe the video and help viewers find content more easily—, the video's rank within the trending videos list, publication time, trending time, trending country, and assignment to one of the 15 video categories.

For each channel to which these videos belong, I observe the channel's name and description, a profile photo, the date when the channel was established, the channel's country

[38] I focus on Canada and the United States for two reasons: (i) dislike attacks, the form of harassment that YouTube aimed to address with this policy, have been most frequently discussed in the North American context, and (ii) these are the two countries for which I have supply and demand information. The original dataset also includes France, Germany, and the United Kingdom, which I use in robustness checks presented in Appendix A.3 to demonstrate that the observed effects apply more broadly. Additionally, it includes Korea and Japan, where I do not find a gender gap; I use these countries in a placebo test to show that, in the absence of an effect on dislikes, there is no effect on content supply or demand. This finding demonstrates that the supply and demand effects are not solely driven by dislikes being private.

[39] This results in the exclusion of 34 channels. The estimated effects in this cross-sectional dataset are comparable to those estimated in the fixed-effects regressions. These are not reported in the paper.

of origin, verification and monetization status, and its assignment to one or more of 7 possible categories and 62 subcategories. I further estimate each content creator's gender and age based on their profile photo. I describe the methodology in Section 4.4.

I compare the observable characteristics of trending videos by male and female content creators in Table 1. While these videos are comparable in terms of video age, and number of likes and views, there are three important differences. First, videos by women on average receive 1,231 (43 percent) more dislikes and 3,338 (38 percent) more comments than those by men. Second, channels by female content creators whose videos are featured in trending videos tend to be about 10.3 months older than male channels. Male conntent creators produce 10.4 percentage points more content in entertainment, 10 percentage points more in sports, and 7.8 percentage points more in gaming. Female content creators produce 30 percentage points more videos 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 is 'trending' on YouTube is determined by the YouTube trending videos algorithm. Trending boosts the visibility of videos on the platform. The algorithm is non-personalized and relies on multiple factors, including view count, the rate of view accumulation ("temperature"), the sources of views (including those external to YouTube), the video's age, and its performance relative to other recent uploads from the same chan-

nel.[40] YouTube claims to prioritize what it calls 'safe content' in trending videos, filtering out severe profanity, mature content, violence, and other content deemed inappropriate, such as content that disparages others in the YouTube community.

Importantly, dislike counts do not feature in the trending videos algorithm. Consequently, variation in dislikes does not affect the selection of videos for the trending list. An independent study by Mozilla Foundation, based on data from over 20,000 users, confirms that dislikes are not part of the algorithm.[41] Additionally, there are no indications of changes to the trending videos algorithm around the time that the dislike counts were removed from YouTube on November 10, 2021.[42] Based on this evidence, I assume that trending videos before and after the treatment are comparable.

4.2 Productivity Data

Next, I construct a balanced panel of monthly video production for all channels that appear in the trending video sample at least once. I retrieve the full list of video uploads for each channel from the YouTube API.[43] Importantly, the data includes deleted videos along with their publication dates, which allows me to capture the full production history of each content creator.[44] It is important to include deleted videos as the probability that a creator deletes a video might be directly related to the number of dislikes it receives, and whether or not those dislikes are public.[45]

The resulting sample covers 1,583 channels, of which 1,277 channels belong to male content creators and 306 channels belong to female content creators. I restrict the sample to cover October 2020 to November 2022 and aggregate the content production statistics

[40] https://support.google.com/youtube/answer/7239739?hl=en

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

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

[43] The YouTube API is a publicly accessible application programming interface that offers structured access to YouTube data, including videos, playlists, and channels. It is primarily used by web developers seeking to integrate YouTube functionalities into their own applications. https://console.cloud.google.com/marketplace/product/google/youtube.googleapis.com YouTube offers researchers an extended quota for data retrieval to facilitate the use of its data in research activities. https://research.youtube/how-it-works/

[44] The date a video is added to the upload list serves as an approximate publication date, though creators may sometimes upload a video shortly before making it publicly available.

[45] While deleted videos are included, videos that were made private before I accessed the YouTube API are not included in this list. The likelihood of a video being made private (unlisted) may be influenced by the platform design change that hid dislike counts. Fortunately, my dataset on views per channel allows for testing any discontinuities in the rate at which videos are made private or deleted. This analysis does not identify a discontinuity in video unlisting patterns for either male or female creators at the time of the policy change (Figure A1).

on monthly level.[46] Prior to treatment, the average male content creator in the sample produced 13 videos per month, while the average female content creator produced 9 videos per month (Table 2). The productivity distribution, however, is skewed: The median male content creator in the sample produced 5 videos per month while the median female content creator produced 4 videos per month. The difference in productivity between male and female content creators is significant at both mean and median.

Table 2: 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: $n_{male} = 16,575$; $n_{female} = 3,978$.

4.3 Demand Data

I construct a balanced panel of monthly views for each channel. I use Socialblade API to obtain a daily snapshot of aggregate views and subscribers per channel over time.[47] As my main measure of demand, I calculate the daily change in aggregate views over time. The aggregate views over time exclude videos that are at any point deleted or made private, so that the difference between two consecutive dates can also be negative. To account for this, I aggregate data on monthly level and drop channels for which the difference between two consecutive months is negative.[48]

After these adjustments, I obtain measures of demand for 1,749 channels. 1,317 channels belong to male content creators and 432 channels belong to female content creators. I restrict the sample to cover October 2020 to February 2023.

There are no significant gender differences in average monthly views, however monthly

[46] While many creators appear to release videos on a weekly schedule, I aggregate the data monthly to maintain consistency with the content demand analysis, where a monthly level is used due to data limitations.

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

[48] I also test whether the pattern of these negative views (which capture how frequently videos are deleted or made private) changes following the treatment using a density test typically used for detection of self-selection sorting in regression discontinuity designs (Cattaneo et al. 2018). I find that there is no discontinuous change in the frequency of negative views after treatment (Figure A1).

views are highly skewed (Table 3). The average male content creator receives over 28.2 million views a month, while the average female content creator receives 33.7 million views a month (16 percent more, not statistically significant). At the median, male content creators receive 5.4 million views a month, while female content creators receive 3.1 million (56 percent fewer) views a month, a statistically significant difference (p=0.00). Additionally, female channels on average add 9 thousand (18 percent, p=0.06) fewer subscribers per month than male channels. The difference is greater at the median, where female content creators add 10 thousand (50 percent, p=0.00) fewer subscribers per month than male channels.

Table 3: 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: $n_{male} = 25,023$; $n_{female} = 8,208$.

4.4 Channel Classification

To examine the individual productivity effects of gendered feedback, I limit my sample to channels that are clearly identifiable as run by an individual, and where the gender of the individual is also distinctly recognizable based on their profile photo. While this approach might not perfectly isolate the channels that are directly exposed to consumer feedback,[49] it serves as a reasonable proxy and, on average, will select channels that are more responsive to feedback.

Content creators often use a personal photo as their profile image on YouTube. I use DeepFace, a Python package that wraps several state-of-the-art facial recognition models to analyze profile photos (Serengil and Ozpinar 2020).[50] This procedure first identifies the number of faces present in the profile image, and second, it determines the gender of each person based on those faces. The output includes a probability indicating whether what

[49] Some might still employ a sizeable team for their operations.

[50] https://github.com/serengil/deepface (explain in detail what this tool does and which of the facial recognition models I employ and why)

has been detected is indeed a face, as well as the probability that the face belongs to a male or female individual.[51]

In total, I identify 2,918 channels (37.4 percent) as belonging to individuals (a single face identified in the photo), 662 channels (8.5 percent) belonging to groups of individuals (at least two faces identified in the photo), and 4,219 channels (54.1 percent) with no faces identified (Table 4). As I want to study channels that are responsive to feedback, I focus on minimizing false positives (channels that are identified as belonging to an individual when they do not) rather than false negatives (channels that are not identified as belonging to an individual when they do). I therefore set a high threshold of 90 percent certainty for facial recognition, leaving me with 2,750 channels where an individual can be clearly identified in the profile photo.[52]

Table 4: Category: (1) individual, (2) group, (3) other

	Count	Pct
Individual	2,918	37.4
Group (2+ individuals)	662	8.5
Other (No face(s) found)	4,219	54.1
Total	7,799	100.0

Of these 2,750 channels where an individual face can be identified with 90 percent certainty, 668 faces (24.3 percent) are identified as belonging to a female and 2,082 (75.7 percent) are identified as belonging to a male (Table 5. To reduce the rate of miss-classification, I drop channels that are identified as belonging to a male respectively female individual with less than 70 percent certainty, leaving 2,635 channels.

Table 5: Gender

	Count	Pct
Female	668	24.3
Male	2,082	75.7
Total	2,750	100.0

To evaluate the performance of my facial-recognition approach for channel classification, I compare its outcomes to those of a manual classification of a subsample of channels. In this smaller sample of 339 channels, slightly more than 10 percent of the total, 72 channels

[51] Another measure that the package outputs is age, I use this to show that there are no differences in estimated content creator age in table 1.

[52] I.e. this results in dropping slightly less than 10 percent of the channels.

(22 percent) are identified as belonging to women, while 253 channels (78 percent) are identified as belonging to men. Treating this manual classification as the ground truth, the profile-photo classifier correctly identifies 248 (98 percent) of male-run channels and 63 (88 percent) of female-run channels. While the profile-photo classifier performs well overall, it tends to classify women (the treated group) as men (the control group) more frequently than vice versa. If anything, this tendency would bias my treatment estimates downwards.

4.5 Comments

To investigate whether reduced incentives to dislike a video lead to increased negative feedback through other public feedback channels, I analyze the text of video comments. I focus on comments that directly address the video, omitting replies to other comments to avoid potential bias from sentiments directed at other commentators.

Trending videos predominantly feature highly popular channels that receive a large number of comments on their content. I limit the data to 45 days pre-treatment and 45 days post-treatment in this analysis.[53] I also limit the analysis to comments posted within 2 weeks of the video being published, which captures 60 percent of all comments. To exclude partially treated videos from the analysis, I omit 2 weeks of videos just before the treatment from the data and take the 45 days before that as the pre-treatment group.

The resulting sample covers over 26 million comments across more than 50 thousand videos by 1256 channels. 940 of these channels are run my male creators and 316 by female creators. Prior to the platform design change, a video by a male content creator received on average 515 comments from 219 unique commentators, while a video by a female content creator received on average 607 comments from 262 unique commentators. These differences are not statistically significant (Table 6).

I employ two natural language processing algorithms to examine potential changes in the tone of comments after the platform design change.[54]

[53] I collect comments about two years after compiling the initial dataset of videos for each channel. During this time, some videos are deleted, limiting my ability to retrieve comments. To address this issue and reduce potential bias from selectively deleted videos, I limit the analysis to a shorter pre- and post-treatment period, focusing only on a subset of available videos (93 percent in this sample). (As explained in the data section, video deletions do not impact the productivity analysis, as deleted videos can still be observed, albeit with limited information.).

[54] Both tools have been used in previous academic literature (e.g. Aguiar (2024) use VADER and Beknazar-Yuzbashev et al. (2022) use Perspective API.) Perspective API is also well-established content-

First, I use a sentiment analysis tool specifically designed to capture sentiments expressed on social media: VADER (Valence Aware Dictionary and sEntiment Reasoner) (Hutto and Gilbert 2014). VADER relies on an extensive lexicon to process social-media specific features, such as emojis and internet slang.[55] Going beyond a simple bag-of-words approach, it also incorporates a list of rule-based heuristics to capture word-order sensitive relationships between terms. VADER scores sentiment on a scale from -1 to 1, with -1 representing very negative sentiment and 1 representing very positive sentiment.

Second, I use Perspective API, a tool specifically designed for content moderation to identify toxic comments.[56] It is trained on comments rated by 3 to 10 human raters, whose evaluations serve as input for a machine learning model that learns to provide similar scores for these and related comments.[57] Perspective API defines a toxic comment as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion." The scoring system outputs a probability that a comment will be perceived as toxic.[58]

The main strength of this measurement tool is its focus on directly assessing the toxicity of comments rather than overall sentiment. However, in contrast to VADER, it does not process social media-specific features such as emojis and internet slang, which are common in my dataset. Additionally, the quota limits of Perspective API allow me to process only a subset of comments.[59] Therefore, I present the results from the VADER sentiment analysis tool in the main Mechanisms section and include the findings from the Perspective API measure of toxicity in Appendix A.5.

Prior to the platform design change, videos by male content creators received comments with a less positive sentiment than videos by female content creators. Mean comment on a video by male content creator receives a score of 0.18 while a mean comment on a video by a female content creator receives a score of 0.33 on a scale from -1 to 1. This holds along the distribution, with 10th percentile of comments receiving a score of -0.31 for men compared to -0.15 for women, and 90th percentile of comments receiving a score of 0.70

moderation tool in the industry, used by, the New York Times, the Wall Street Journal, and Reddit, among others. Source: https://perspectiveapi.com/

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

[56] https://perspectiveapi.com/

[57] For more details, see https://developers.perspectiveapi.com/s/about-the-api-training-data

[58] For example, if 3 out of 10 human raters perceive a comment as toxic, the API will provide a score of 0.3 for that and similar comments.

[59] I select a random subset of 100 comments per video for which I retrieve the Perspective API toxicity score.

for men and 0.82 for women (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.

5 Empirical Approach

5.1 Effect on Dislikes

I estimate the average treatment effect of the platform design change in a two-way fixed effects difference-in-differences 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.[60] The coefficient β captures the average treatment effect for women's channels,[61] under the par-

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

[61] Throughout the paper, I interpret coefficient estimates as percentage effects when reporting results.

allel trends assumption: that in the absence of treatment, the proportional difference in growth rates of dislikes for channels run by men and women would have evolved in parallel (McConnell 2024).[62] To assess this assumption, I estimate an event study and plot the estimates to show that outcomes evolved in parallel before the platform design change (Figure 6.2).

To estimate the effect that the removal of public dislikes had on men, I exclude the time fixed-effects from the specification. This allows me to estimate a baseline coefficient for the post-period:

$$Y_{ict} = \alpha + \beta_1 Post_t + \beta_2 Post_t * Female_i + \gamma' \mathbf{X}_{ict} + \delta_{ic} + \epsilon_{ict} \tag{2}$$

where β_1, the estimated coefficient for $Post_t$, the indicator for the post-treatment period captures the differences in dislikes received by men after the design change.

To assess the size of the gender gap in feedback, I estimate a split-sample regression separately for the pre- and post-treatment samples. I exclude channel-country fixed effects, which allows me to estimate a coefficient for a female dummy variable:

$$Y_{ict} = \alpha + \beta Female_i + \gamma' \mathbf{X}_{ict} + \lambda_t + \epsilon_{ict} \tag{3}$$

where β, the estimated coefficient for $Female_i$ (the indicator for female content creators), captures the size of the gender gap in dislikes. I estimate the regression separately for the pre- and post-treatment samples to identify the size of the gender gap before and after the treatment. In addition to video characteristics, covariates include fixed effects for 62 channel categories to account for differences in content produced by channels, while omitting channel-country fixed effects.

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: $exp(\beta) - 1$.

[62] If the outcome variable were the absolute dislike count instead, the underlying parallel trends assumption would be that the outcomes change by the same *absolute* amount over time. Given the considerable variation in channel size, video popularity, and, subsequently, dislike counts, I therefore prefer to use the log transformation as my functional form.

5.2 Predicting Counterfactual Dislikes

Male and female content creators produce different types of content. To ensure that these differences in content production patterns do not bias my estimates, I account for them in two ways. First, I control for channel-country and video-topic fixed effects throughout my analyses. These fixed effects control for content categories, ensuring that gendered effects are estimated using only within-topic variation. Second, I predict the counterfactual dislikes that videos by women would have received if they had been produced by men. I include these predicted dislikes as an additional covariate in my main regression. This allows me to explicitly account for more nuanced differences in the types of content produced by men and women, which may not be fully captured by broader video-topic categories or by time-invariant channel-country fixed effects.

To generate these predictions, I train a machine-learning model on data describing videos produced by men prior to treatment.[63][64] To implement this approach, I estimate a Lasso (Least Absolute Shrinkage and Selection Operator, L1) regularized regression (James et al. 2023):

$$y_i = X_i\theta + \epsilon_i \tag{4}$$

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:

$$\hat{\theta} = \arg\min_{\theta} \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - X_i\theta)^2 + \alpha \sum_{j=1}^{p} |\theta_j| \right) \tag{5}$$

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

[63] I limit the training data to observations of videos prior to treatment since men also experienced some level of treatment.

[64] This approach is similar to that of Iaria et al. (2024), who document a gender gap in citations of scientific papers and face the challenge of differences between papers written by women and those written by men. They train a model to predict citations based on papers authored by men, demonstrating that the estimated gender gap in citations for women remains robust even when including these counterfactual citations in the regression.

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. Ridge regression, a common alternative, employs 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 regression includes all predictors in the final model. While this can improve prediction accuracy, it comes at a cost of reduced interpretability. Since the Lasso estimator achieves high accuracy in my setting, I choose it for its better interpretability.[65]

Predictors include the count of likes, views, comments, and their ratios, which measure popularity and engagement, reflecting audience interest. Video and channel categories account for variations in content type and target audience that might influence engagement. The country where the video is trending, along with its country of origin, captures geographic and demographic patterns. Timing variables, such as the time of publication, when the video began trending, and the elapsed time between these events, capture the effects of timing on viewer engagement. Additionally, indicators for whether the channel is monetized or verified assess the impact of channel status on engagement. Finally, interactions between these variables and their second-degree polynomials enable the model to capture potential nonlinear relationships.

I split the sample into a training and testing dataset (80/20 percent). I standardize the predictors by removing the mean and scaling them to unit variance. Using 10-fold cross-validation, I set the regularization parameter α=7.7. The Lasso selected coefficients explain 96 percent of the variation in the test sample.

I then roll out the model to predict the *counterfactual* dislikes for the entire sample. Finally, I estimate the following regression, which incorporates the predicted dislikes as an additional control:

$$Y_{ict} = \alpha + \beta Post_t * Female_i + \gamma' X_{ict} + \hat{Y}_{ict} + \delta_{ic} + \lambda_t + \epsilon_{ict} \tag{6}$$

[65] A regression with a Ridge (L2) loss function performs comparably well, explaining 96 percent of the variation in the test sample.

where $\hat{Y}_{ict}$ is the *predicted* logged daily mean count of dislikes per channel-country observation if the video had been produced by a man in the pre-treatment period.[66] Intuitively, controlling for these predicted dislikes allows me to condition the analysis on the 'objective' dislikability of content that would have been observed if the focal video were not *also* affected by gender-biased feedback.

5.3 Effect on Productivity and Demand

I estimate the effect of the platform design change on content supply and demand in a fuzzy difference-in-differences design. OLS estimates on the effect of the design change on dislikes demonstrate that dislikes decrease for both male and female content creators, but that the effect is much larger for women. Fuzzy difference-in-differences design allows me to leverage these relative effect magnitudes to identify the causal effect of the reduction in negative feedback on women's productivity and the audience's demand for their content. I can estimate the causal effect on women's productivity under the identifying assumption that the treatment effect is the same in the treatment and control group. That is, women and men *on average* react symmetrically to the removal of public dislike counts (De Chaisemartin and d'Haultfoeuille 2018). I estimate the average treatment effect on the productivity of the treated content creators using OLS:

$$Y_{it} = \alpha + \beta' Post_t * Female_i + \gamma' \mathbf{X}_{it} + \delta_i + \lambda_t + \epsilon_{it} \tag{7}$$

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 captures the average treatment effect for women's channels, under the assumption that channels run by men and women follow parallel trends in terms of the proportional difference in growth rates of produced videos respectively views per channel (McConnell 2024). To assess this parallel trends assumption, I estimate an event study version of the regression and plot the event study

[66] 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.

graph. The event study estimates also allow me to assess the persistence of the treatment effect over time.

5.4 Mechanisms

5.4.1 Heterogeneous Effects on Dislikes

To identify how the effect of the platform design change differs along the distribution of dislikes, I estimate a quantile regression. As my specifications include panel fixed effects, I cannot estimate a standard quantile regression. This is because traditional methods for controlling unobserved fixed effects, such as demeaning or differencing, rely on the linearity of expectations, which does not apply to quantiles. Instead, I implement a version of the 2-step quantile estimator of Canay (2011). First, I estimate the fixed effects regression specified in Equation 1, retrieve the estimated conditional fixed effects, and within-demean the outcome variable. Second, I estimate a quantile regression on the demeaned data:

$$Quant_{\tau}(Y_{ict}) = \alpha(\tau) + \beta' Post_t * Female_i(\tau) + \gamma' \mathbf{X}_{ict} + \epsilon_{ict} \tag{8}$$

where Y_{ict} is the demeaned logged daily mean count 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 are dependent on the quantile τ. The coefficient vector $\beta'(\tau)$ captures the treatment effect by quantile τ for women's channels, relative to men's channels in the same quantile.The main assumption underlying conditional quantile regression is linearity in the parameters at each quantile. That is, conditional quantiles of the dependent variable can be expressed as a linear combination of the independent variables (Cameron and Trivedi 2022). Following Canay (2011), I bootstrap the standard errors for correct inference.

To assess the presence of a gender gap in feedback across the entire distribution and examine how this gap changes after treatment, I split the data into a pre- and post-treatment samples and estimate a pooled quantile regression on each sample. In this specification, I exclude the channel-country fixed effects and include a female dummy

variable:

$$Quant_{\tau}(Y_{it}) = \alpha(\tau) + \beta' Female_i(\tau) + \gamma' \mathbf{X}_{ict} + \epsilon_{ict} \tag{9}$$

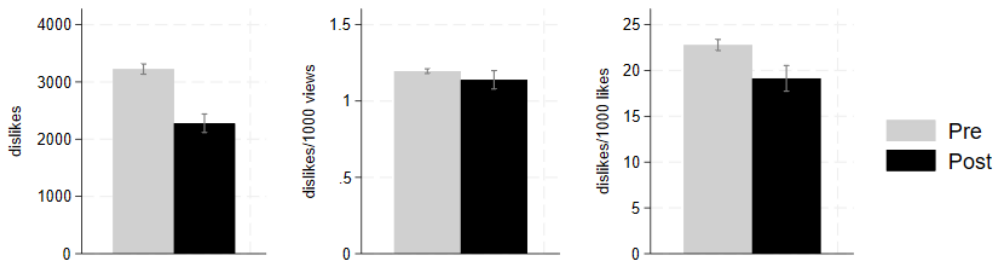
where the $\beta'(\tau)$ coefficient vector captures the gender gap by quantile of the dislike distribution before and after treatment.

6 Results

6.1 Dislikes Before and After the Platform Design Change.

To illustrate the effect of the platform design change on dislikes, I compare the average dislikes per video in the 454 days before and the 34 days after the design change. After the removal of public dislike counts, dislikes decrease. The average video received 937 (-30 percent) fewer dislikes in the post period (Figure 2). This effect, however, is partly explained by videos in the pre-period receiving more views (they have a longer 'air' time). Per 1,000 views, the reduction in dislikes is only marginally significant (p = 0.xx). The average video received 0.05 fewer dislikes per 1000 views in the post-period (-4 percent). Comparing dislikes per 1,000 likes – rather than views – the average video receives 3 fewer dislikes per 1000 likes after the platform design change (-14 percent).

Figure 2: Overall dislikes



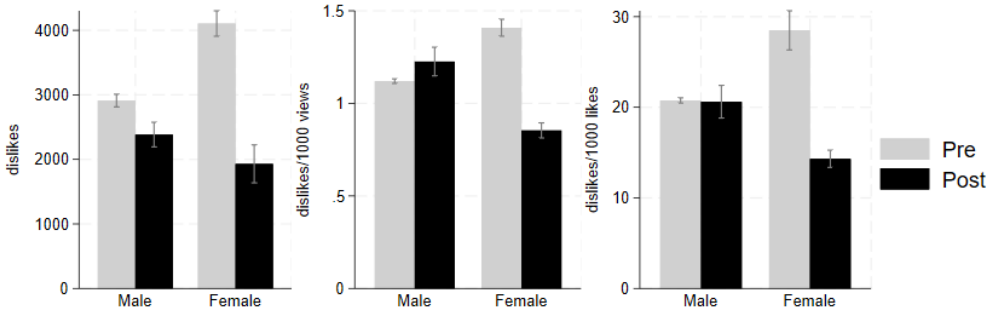
To illustrate the impact of the platform design change on dislikes by gender, I compare the average dislikes per video before and after the change, differentiating between channels run by men and women. This analysis reveals significant differences between the two groups. Notably, prior to the design change, women received substantially more dislikes than men, with an average of 4,067 dislikes compared to 2,836. This disparity is also evident when considering the dislikes relative to views (1.4 dislikes per 1,000 views for women versus 1.1 for men) and to likes (29 dislikes per 1,000 likes for women compared to 21 for men) (Figure 3).

After the design change, women experienced a significant reduction in dislikes compared to men. Women received an average of 1,574 dislikes overall, while men received 2,354. This trend continues when comparing dislikes relative to views (0.9 for women versus 1.2 for men) and relative to likes (15 for women compared to 21 for men).

In the post-treatment period, the average video by a woman received 2,313 fewer dislikes (-57 percent). When adjusted for views, the average video by a woman saw a decrease of 0.55 dislikes per 1,000 views (-39 percent). In terms of dislikes per 1,000 likes, women's videos experienced a reduction of 14 dislikes (-49 percent) after the design change.

Conversely, the average video by a man received 482 fewer dislikes (-17 percent) in the post period. Interestingly, per 1,000 views, the average video by a man received 0.1 *more* dislikes, although this difference not statistically significant at p=0.102. In terms of dislikes per 1,000 likes, the average for men remained unchanged at 21, indicating no significant shift compared to the pre-treatment period. Thus, while women experienced a substantial decrease in dislike counts after the platform design change, the effect on men appears less consistent.

Figure 3: Dislikes by gender



6.2 Causal Effect on Dislikes

I estimate the causal effect of the platform design change on dislikes in difference-in-differences regression (Table 7). The estimates show that women experience a 30 percent decrease in dislikes after the design change, relative to men (Column 1). This effect is robust to including additional controls. Conditional on the number of views per video and video age, women receive 21 percent fewer dislikes (Column 2) and controlling for the video's topic category, the coefficient remains unchanged (Column 3).

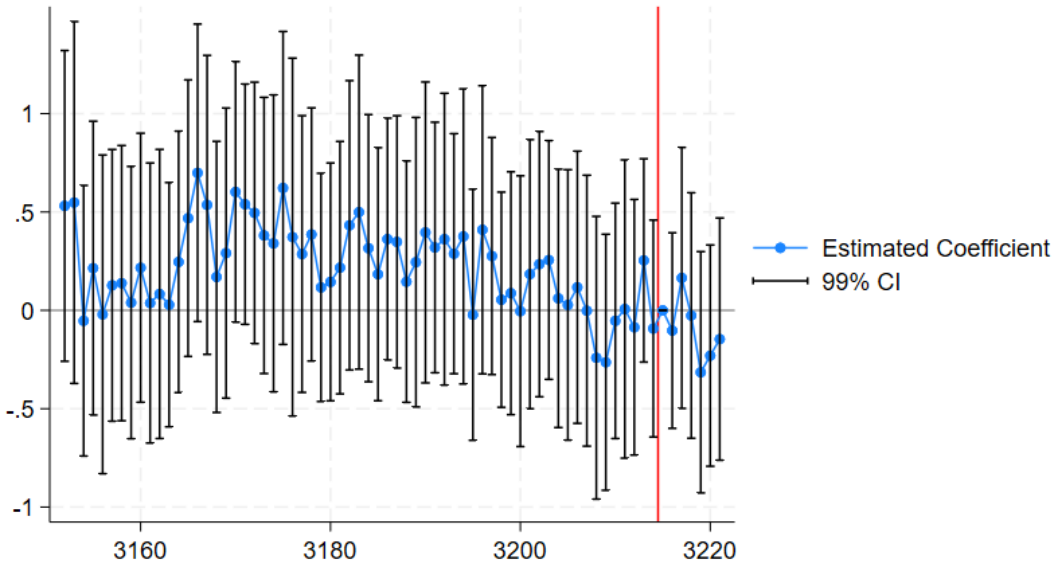
Table 7: Effect of platform design change on dislikes

VARIABLES	(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
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

Figure 6.2 plots the event study estimates based on data aggregated at weekly level for clarity, illustrating parallel trends in dislikes before the treatment between male and female creators. The coefficient estimates for the post-treatment period show that while there is a decrease in dislikes for women relative to men, it is moderate in magnitude at the mean. Instead, I find that there are large treatment effects in the tails of the dislike distribution.

Figure 4: Parallel trends: ln(Dislikes)



To account for the potential influence of content differences between female and male creators on the observed treatment effects, I condition my analysis on the counterfactual dislikes that videos by women would have received if they were created by men prior to the treatment. The estimated coefficient on predicted dislikes is positive and significant in all specifications (Table 8): A 1 percent increase in predicted dislikes is associated with 0.11 to 0.15 percent increase in actual dislikes across specifications. After accounting for predicted dislikes, women experience a 28 percent reduction in dislikes following the design change, relative to men (Column 1). When controlling for the number of views per video and video age, women receive 21 percent fewer dislikes (Column 2) and controlling for the video's topic category, the coefficient remains unchanged (Column 3). Conditioning on 'dislikability' of content *increases* the treatment effect, indicating that women select into producing content that make them less susceptible to receiving excessive dislikes.

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

VARIABLES	(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
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

The selection of videos into the trending sample may have changed with the treatment. Concerns regarding these changes are mitigated by the fact that dislikes do not directly influence YouTube's trending algorithm,[67] and no other platform changes were implemented at the time of the removal of public dislikes. Nonetheless, changes in dislike counts could have indirectly affected the selection of trending videos. To address the potential differential selection into trending videos after the platform design change, I restrict the sample to channels that had videos featured in the trending section both before and after this change (Table 11).

The estimates from this restricted sample align closely with the main results: women experience a 32 percent reduction in dislikes following the design change compared to men (Column 1). When controlling for the number of views per video and video age, women receive 20 percent fewer dislikes (Column 2), and when incorporating video controls and topic category controls, women see a 21 percent decrease in dislikes compared to men after the design change (Column 3). Additionally accounting for predicted dislikes increases the estimated coefficient to 24 percent (Column 4).

[67]

Table 9: Accounting for selection into trending videos

VARIABLES	(1) ln(Dislikes)	(2) ln(Dislikes)	(3) ln(Dislikes)	(4) ln(Dislikes)
Post × Female	-0.32***	-0.20**	-0.21**	-0.24***
	(0.12)	(0.09)	(0.09)	(0.09)
ln(Predicted dislikes)				0.13***
				(0.01)
Observations	18,263	18,263	18,263	18,263
R-squared	0.66	0.77	0.77	0.80
YT Channel-Country IDs	436	436	436	436
YT Channels	284	284	284	284
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

The platform design change significantly reduced dislikes on women's content. But what was the effect of the design change on male content creators? To estimate the effect that the removal of public dislikes had on men, I exclude the time fixed-effects from the specification. This allows me to estimate a baseline coefficient for the post-period, capturing the differences in dislikes received by men after the design change (Table 10). Controlling only for channel fixed effects (Column 1) men experience a 12 percent reduction in dislikes following the design change (p=0.029). However, when controlling for the number of views per video and video age (Column 2), when incorporating both video controls and topic category controls (Column 3) and adding the predicted dislikes (Column 4), the point estimate becomes smaller and statistically insignificant. This indicates that the platform design change has a limited impact on men.

Table 10: Treatment effects for men

VARIABLES	(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
Channel-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

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 to allow for identification of a female dummy (Table 11).

In the pre-treatment sample, the coefficient for the female dummy is positive at 13 percent and marginally statistically insignificant (p=0.106) without controlling for predicted dislikes (Column 1). When controlling for predicted dislikes, it is 14 percent and significant at p=0.041 (Column 2). In the post-treatment sample, the coefficient for the female dummy is negative and insignificant, at -8 percent without predicted dislikes (Column 3) and -4 percent with predicted dislikes (Column 4).[68] These results suggest that the platform design change effectively closed the gender gap in feedback and that women are no longer receiving more dislikes for similarly dislikable content than men.

[68] As men are also partially treated, I separately predict dislikes based on videos by men in the post-treatment sample, and control for these in the post-treatment sample for accurate comparison. I follow the same prediction procedure as for the predicted dislikes based on the pre-treatment sample and the model likewise performs exceptionally well. Using 10-fold cross-validation, I set the regularization parameter $\alpha = 10.5$. The Lasso selected coefficients explain 95 percent of the variation in the test sample.

Table 11: Gender gap in feedback before and after treatment

VARIABLES	(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
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

6.3 Effect on Productivity

I estimate the causal effect of the platform design change on female content creators productivity in a fuzzy difference-in-differences regression (Table 13). The estimates indicate that, relative to men, women increase their monthly supply of videos by 8.4 percent (p=0.01) after the design change. This result corresponds to an additional 0.4 videos uploaded per month for the median content creator in my sample.

Table 12: Effect on Productivity

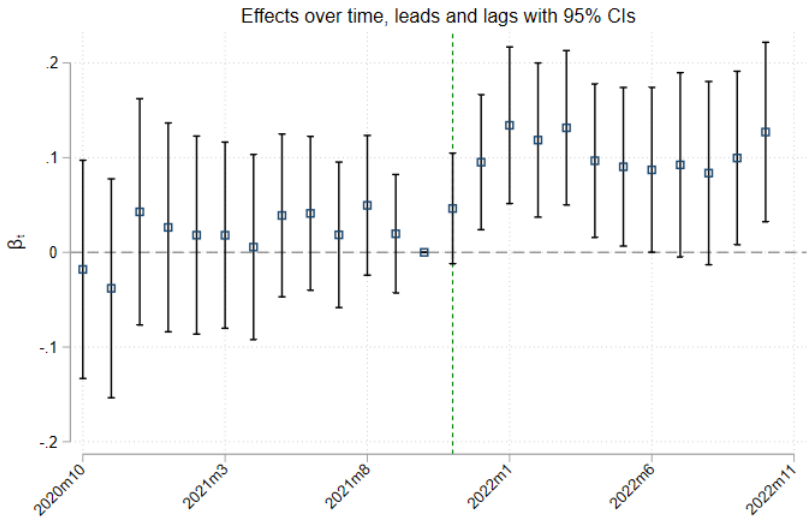
	(1) ln(Videos)
Post × Female	0.084
	[.011,.180]
Observations	37,992
Channels	1,583
R-squared	0.68
Median number of videos	5

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

Event study estimates show that the productivity of female and male content creators

followed a parallel trend before the platform design change (Figure 5). After the design change, the women's productivity increases relative to men's and remains at a higher level for the rest of my sample, suggesting a persistent treatment effect.

Figure 5: Parallel trends in monthly supply



6.4 Effect on Demand

I estimate the causal effect of the platform design change on demand for content by female content creators in a fuzzy difference-in-differences regression (Table 13). The estimates show that, relative to men, women experience a 15.5 percent (p=0.04) increase in monthly views after the design change. A back-of-the envelope calculation suggests that this translates into an increase of $193 to $3,100 in monthly revenues for the median channel in my sample.[69]

[69] Based on revenue estimates of $0.25 - $4.00 per 1000 views by Socialblade.com

Table 13: 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

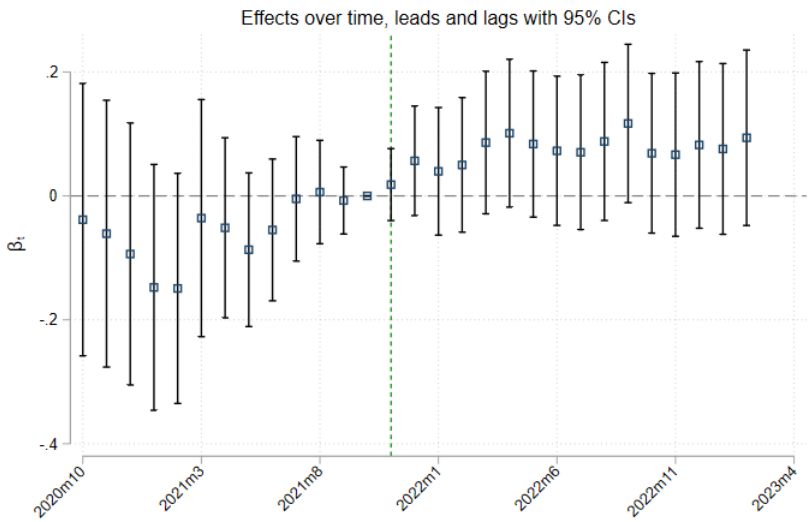
Note: Standard errors calculated using wild cluster bootstrap method with 200 replications. 95 percent confidence intervals reported.

Event study estimates show that the demand for female and male content creators' content evolved along a parallel trend before the platform design change (Figure 6).

Similar to my findings on content creator productivity, demand for women's content increases and remains at a higher level for the rest of my sample, suggesting a persistent treatment effect.

Plotting event study estimates shows that female and male channels followed a parallel trend in terms of monthly views per channel before the platform design change (Figure 6). After the change, the coefficients become steadily positive and remain so over the study period, suggesting a persistent treatment effect.

Figure 6: Parallel trends in monthly demand



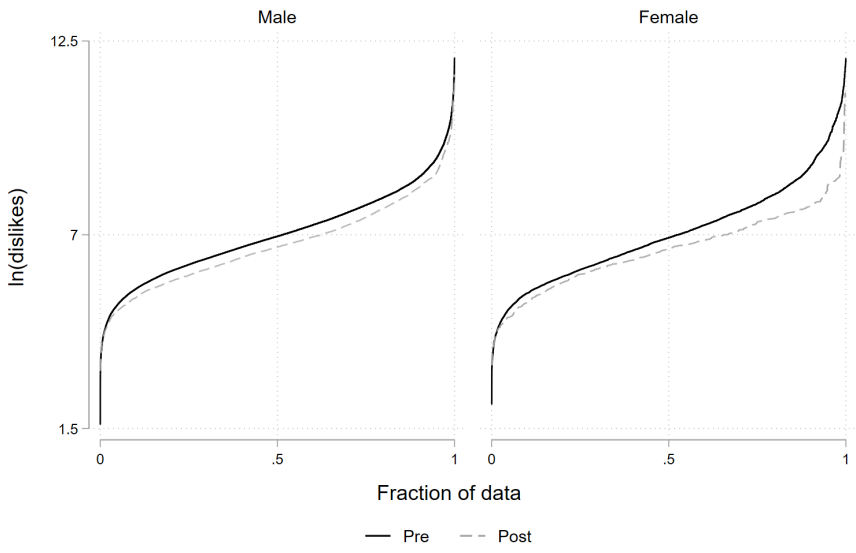
7 Mechanisms

My results document a gender gap in feedback on YouTube: Female content creators receive significantly more negative feedback but removal of public dislikes eliminates this gap. My 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 7). 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 7: Comparison of dislike distribution pre- and post-treatment by gender



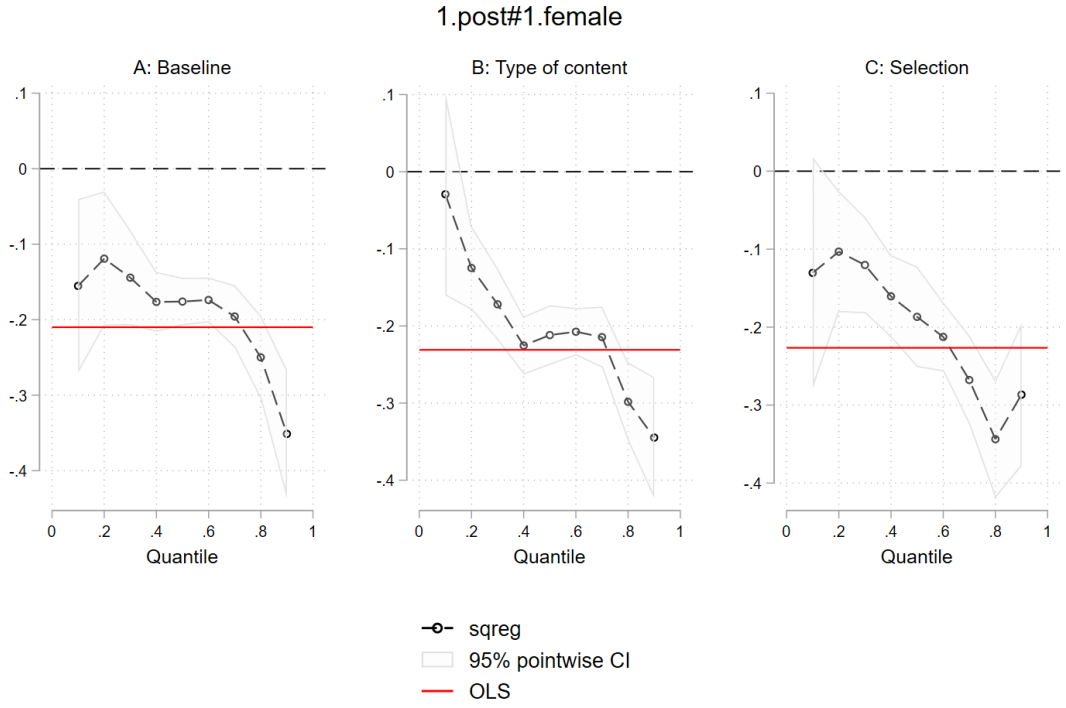
I estimate a quantile regression using the Canay (2011) two-step estimator to analyze the heterogeneous effects of making dislikes private on the number of dislikes (Panel A: Baseline, Figure 8). This specification is equivalent to the OLS difference-in-differences model that I use to obtain average treatment effects for the treated channels but allows me to capture changes in the treatment effect along the dislike distribution.[70] The treatment effect on dislikes towards women's content is negative throughout the distribution. The 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 8),[71] and in a smaller sample that only contains channels that appear in the trending video sample both before and after the treatment to account for changes to selection into trending videos (Panel C: Selection, Figure 8).[72]

[70] Specification in Panel A: Baseline in Figure 8 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.

[71] Specification in Panel B: Type of content in Figure 8 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.

[72] Specification in Panel C: Selection in Figure 8 corresponds to the OLS difference-in-differences regression in Column 2 of Table **??** and accounts for predicted dislikes, channel-country fixed effects, daily fixed effects, video age, and views per video. It is estimated on a restricted sample of channels that had videos featured in the trending section both before and after this change See Table A.3 in the Appendix for detailed results.

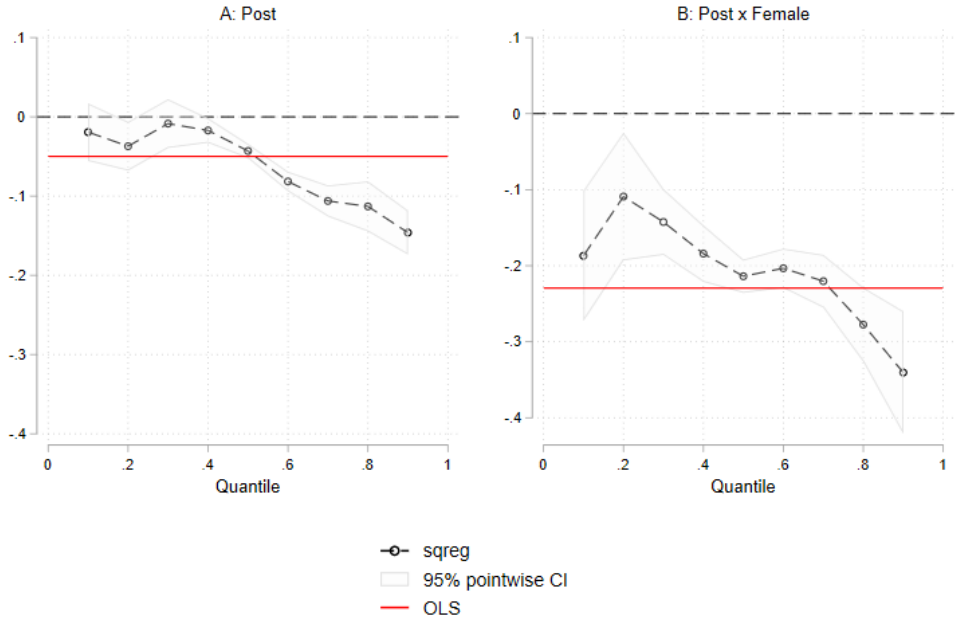
Figure 8: Quantile regression estimates of platform design change on dislikes



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 9).[73] The baseline estimate (Panel A: Post), Like the additional treatment effect female content creators (Panel B: Post x Female), the baseline treatment effect lead 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.

[73]Specification in Figure 9 corresponds to the OLS regression in Column 2 of Table 10 and accounts for channel-country fixed effects, video age, and views per video. See Table A.4 in the Appendix for detailed results.

Figure 9: Quantile regression estimates: Treatment effect for men and women



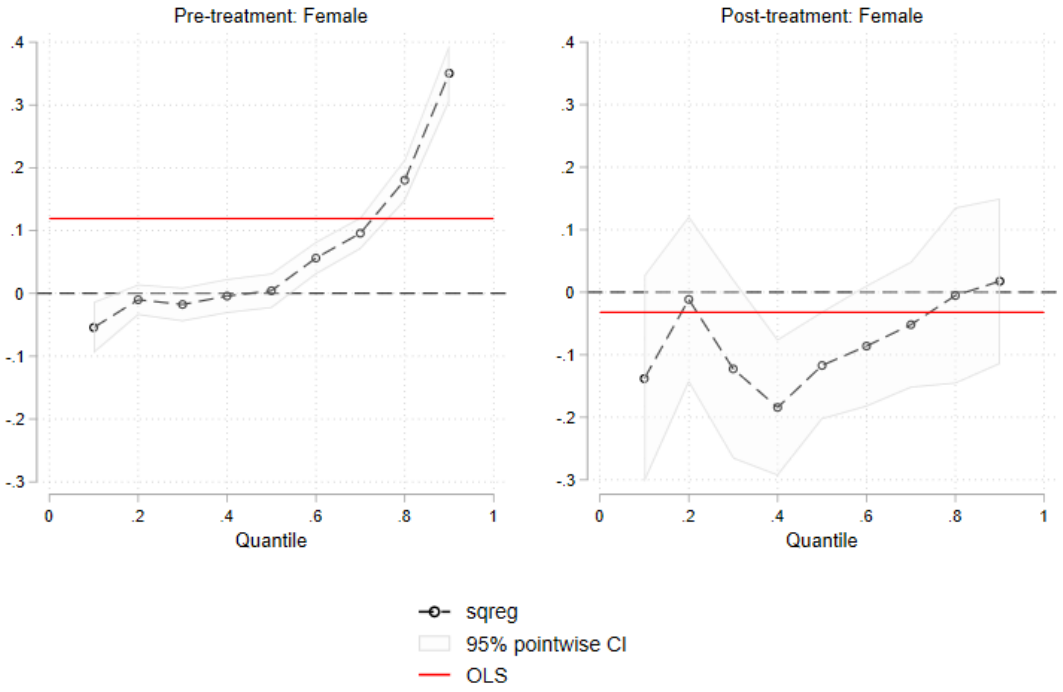
Next, I assess whether the platform design change closed the gender gap in feedback across the entire dislike distribution rather than 'just' on average (as shown in Table 11). To implement this analysis, I first divide the dataset into pre-treatment and post-treatment subsamples. I then estimate a pooled 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.[74]

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 any 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

[74]Specification in Figure 10 corresponds to the OLS regression in Columns 2 and 4 of Table 11 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.

dislike distribution.

Figure 10: 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 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 14). 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 14: 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).[75] Taken together, these results suggest that, if anything, the platform design change increased, rather than decreased, the average sentiment of comments.

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

Table 15: 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 16: 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 17). 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 17: 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 18). 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 18: 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 19). 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 19: 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 20). 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 20: 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. Working Paper, 2022.

Luis Aguiar. Bad apples on rotten tomatoes: Critics, crowds, and gender bias in product ratings. *Available at SSRN 4336108*, 2024.

Raphaela Andres, Michelangelo Rossi, and Mark J Tremblay. Youtube "adpocalypse": The YouTubers' journey from ad-based to patron-based revenues. *ZEW-Centre for European Economic Research Discussion Paper*, (59), 2023.

Guy Aridor, Rafael Jiménez Durán, Ro'ee Levy, and Lena Song. The economics of social media. *Available at SSRN 4708840*, 2024.

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.

George Beknazar-Yuzbashev, Rafael Jiménez Durán, Jesse McCrosky, and Mateusz Stalinski. Toxic content and user engagement on social media: Evidence from a field experiment. *Available at SSRN*, 2022.

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. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2024.

Alan Benson, Danielle Li, and Kelly Shue. Potential and the gender promotions gap. *Available at SSRN*, 2024.

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. 2024.

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

Ronit Bodner and Drazen Prelec. Self-signaling and diagnostic utility in everyday decision making. *The psychology of economic decisions*, 1(105):26, 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*, 2021.

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, 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.

Clément De Chaisemartin and Xavier d'Haultfoeuille. Fuzzy differences-in-differences. *The Review of Economic Studies*, 85(2):999–1028, 2018.

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

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.

Lena Abou El-Komboz, Anna Kerkhof, and Johannes Loh. Platform partnership programs and content supply: Evidence from the YouTube "Adpocalypse". 2023.

Olle Folke and Johanna Rickne. Sexual harassment and gender inequality in the labor market. *The Quarterly Journal of Economics*, 137(4):2163–2212, 05 2022. ISSN 0033-5533. doi: 10.1093/qje/qjac018.

Nickolas Gagnon, Kristof Bosmans, and Arno Riedl. The effect of gender discrimination on labor supply. *Journal of Political Economy*, 0(ja):null, Forthcoming. doi: 10.1086/733419. URL https://doi.org/10.1086/733419.

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)*, Ann Arbor, MI, June 2014.

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.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. Statistical learning. In *An introduction to statistical learning: With applications in Python*, pages 15–67. Springer, 2023.

Rafael Jiménez-Durán. The economics of content moderation: Theory and experimental evidence from hate speech on Twitter. *George J. Stigler Center for the Study of the Economy & the State Working Paper*, (324), 2023.

Garrett Johnson, Tesary Lin, James C. Cooper, and Liang Zhong. Coppacalypse? The youtube settlement's impact on kids content. *Available at SSRN*, 2023.

Tural Karimli. The costs of hidden workplace harassment. *Available at SSRN 4618312*, 2023.

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

Anna Kerkhof. Advertising and content differentiation: Evidence from youtube. *The Economic Journal*, page ueae043, 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. *Available at SSRN 3551103*, 2023.

Brendon McConnell. Can't see the forest for the logs: On the perils of using difference-in-differences with a log-dependent variable. February 2024. City, University of London.

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.

Sungsik Park, Woochoel Shin, and Jinhong Xie. The fateful first consumer review. *Marketing Science*, 40(3):481–507, 2021.

Martina Pocchiari, Davide Proserpio, and Yaniv Dover. Online reviews: A literature review and roadmap for future research. *International Journal of Research in Marketing*, 2024.

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. IEEE, 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. Technical report, Mimeo, 2023.

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.

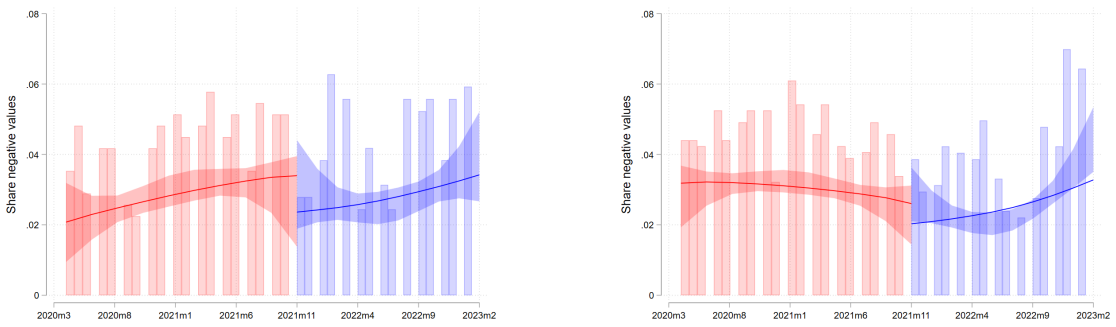
A Appendix

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).[76] 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

[76] 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

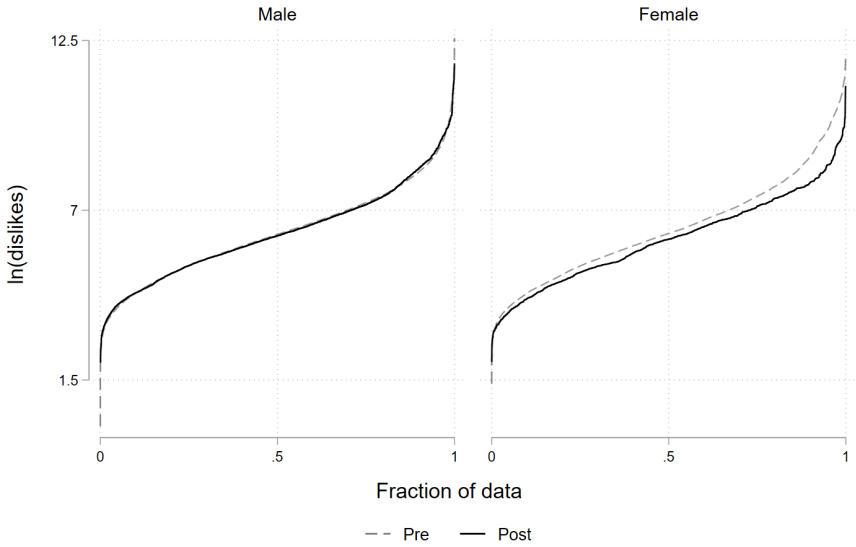


Figure A3: Quantile regression estimates of platform design change on dislikes: 5 country sample

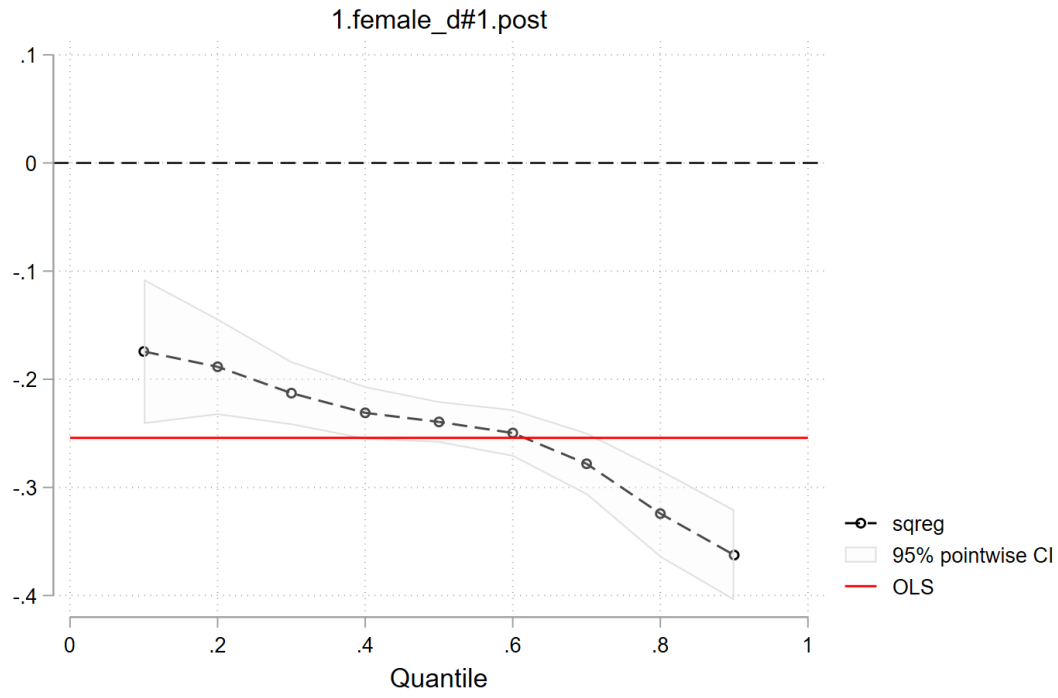


Figure A4: Quantile regression estimates: Treatment effect for men and women in 5 country sample

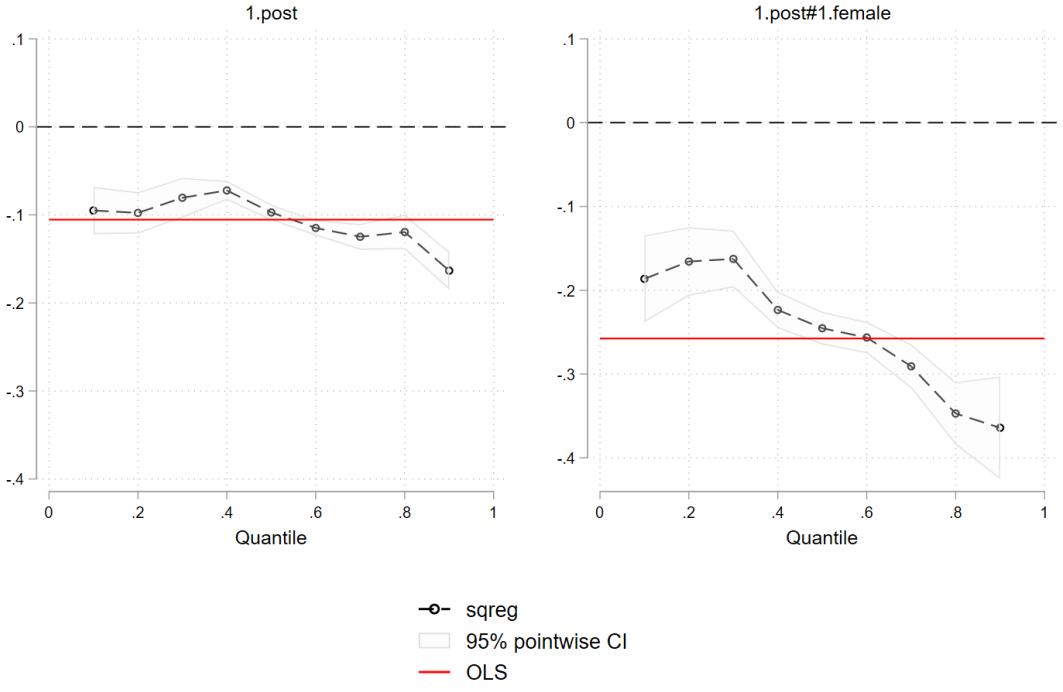
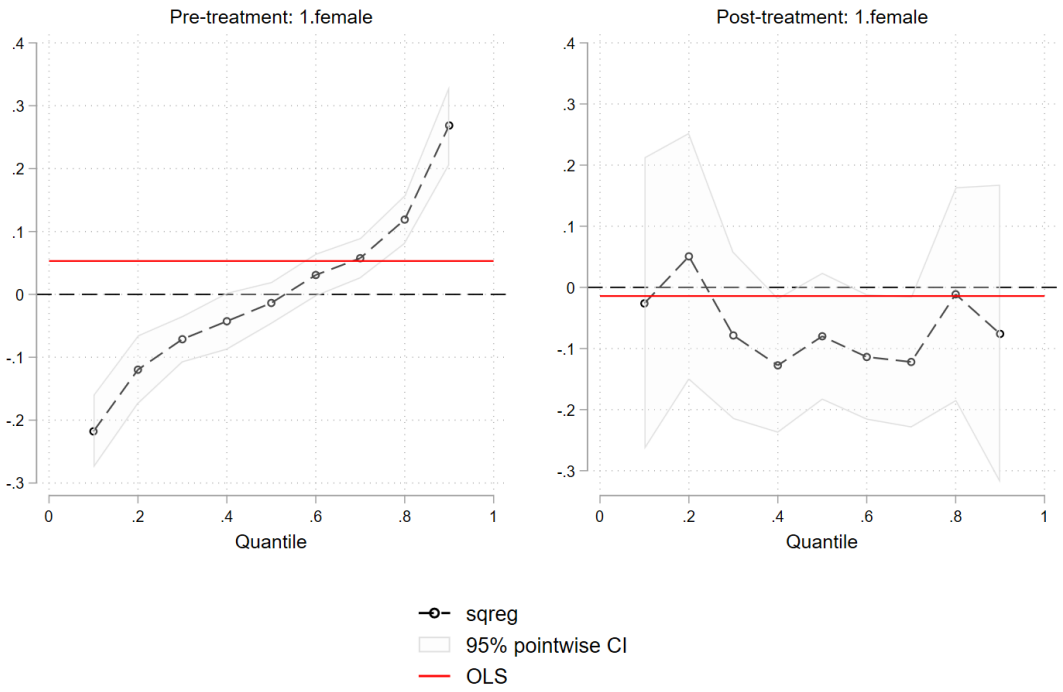


Figure A5: Quantile regression estimates: Gender gap in feedback before and after treatment in 5 country sample



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