Designing Inclusive Platforms: Experimental Evidence from Pinterest

Abstract

We present an *inclusive-by-design* approach for digital platforms, examining how strategic design choices can enhance representation while creating business value. In partnership with Pinterest, we study a redesign of the platform's content discovery interface to achieve a more balanced representation of skin tone in suggested content. Through a large-scale field experiment across six countries, we find that balanced content distribution significantly increases engagement with previously underrepresented content (+15%) without affecting overall platform activity. Users actively embrace content diversity, developing more balanced consumption patterns across different groups. Notably, users who historically sought diverse content show an 11% increase in overall engagement when structural barriers to discovering it are removed, addressing previously unmet user needs. These effects persist 9 months after implementation, indicating that inclusive platform design fundamentally enhances user experience rather than creating temporary novelty effects. Our findings provide the first large-scale causal evidence that thoughtfully designed inclusion initiatives can enhance representation without compromising business performance, offering practical guidance for integrating inclusion into core platform strategy.

Keywords: platform design, inclusion, responsible AI, fairness, field experiment, heterogeneity

Like every other 17 year old girl, I use social media and I am heavily influenced by it. I want to feel relevant. I want to feel beautiful. I want to feel up to speed on the new beauty standards. Lately, if I want to see people of color, I have to physically type the word "black". If I ever want to feel like I am a beautiful contribution, I cannot look at your app. Because when I look at it, I am told "people who look like me are not the norm."

~Anonymous User

1 Introduction

The poignant reflection above highlights a pervasive challenge in the digital age: online platforms, despite their global reach, often fail to represent the diversity of their users. Over 4 billion people across the world use social media today[1] to discover content, entertain themselves, and connect with friends and loved ones. Yet, these platforms frequently perpetuate biases, reinforce stereotypes, and marginalize underrepresented groups (Kay et al., 2015; Lam et al., 2018; Lambrecht and Tucker, 2019).

The design of these platforms, centered around algorithms that curate content and facilitate discovery, plays a pivotal role in determining who feels seen, valued, and included. Algorithmic interfaces, such as search and recommendations, are particularly influential in shaping users' digital experiences, determining what content users discover and engage with across platforms. For instance, personalized recommendations drive 80% of the hours streamed on Netflix (Gomez-Uribe and Hunt, 2016), 70% of consumption on YouTube (Kiros, 2022), and 35% of sales on Amazon (MacKenzie et al., 2013). The increasing prevalence of algorithmic curation on digital platforms has raised important questions about their impact on user experiences and social outcomes. These issues are particularly salient in visual content platforms, where the lack of diversity can marginalize certain user groups and shape societal perceptions of beauty, fashion, and representation.

Despite growing calls for more fair and diverse representation, evidence on how inclusivity impacts platform performance remains scarce. Platforms often hesitate to redesign their systems, fearing that prioritizing representation might compromise engagement or revenue. This hesitation stems from a critical gap in our understanding: Can platforms enhance inclusivity without sacrificing business objectives? Answering this question demands and rigorous causal evidence.

We address this challenge in partnership with Pinterest, one of the largest visual discovery platforms. Pinterest's content primarily consists of images and videos, with a significant portion featuring people. Historically, surfaced content heavily skewed toward items that featured people with "lighter" skin tones (Figure 2). Pinterest redesigned its content policy to strive for a more balanced representation across the skin tone spectrum (Figure 5b) and promote a more inclusive online experience for its users. Our study is among the first to provide causal evidence on how such *inclusive-by-design* initiatives affect user engagement and platform performance.

[1] https://www.forbes.com/advisor/business/social-media-statistics/ (May 2023)

Using a large-scale field experiment with ∼600,000 users across six countries, we find that balanced content distribution significantly increases engagement with previously underrepresented content—without compromising overall platform activity. Notably, users who historically sought diverse content show an 11% increase in overall engagement when structural barriers to discovery are removed, revealing unmet demand.

We make three key contributions. First, our findings challenge conventional assumptions about the trade-offs between inclusivity and engagement. They show that thoughtfully designed inclusion initiatives can enhance representation while simultaneously driving business value. This insight transforms how we think about platform design, shifting the focus from reactive diversity patchwork to proactive, *inclusive-by-design* strategies.

Second, we identify the mechanisms through which inclusive design creates value. We find meaningful heterogeneity in user responses: individuals historically underserved by the platform's existing content policy–due to limited exposure to content matching their preferences–exhibit substantial increases in engagement. This suggests that traditional platform designs create artificial constraints for certain user segments, and that addressing these constraints can unlock new engagement opportunities.

Third, we establish the sustainability and scalability of inclusive design as a platform strategy. The persistence of positive effects nine months after implementation shows that balanced representation fundamentally enhances user experience rather than creating temporary novelty effects. Additionally, our successful extension to body type diversity demonstrates that platforms can simultaneously address multiple representation dimensions without compromising performance. Together, these findings offer a practical framework for implementing and measuring the impact of inclusive design across various platform contexts.

Beyond its immediate findings, our study has broader implications for the future of digital platforms. As online spaces increasingly mediate societal discourse and personal identities, the design of these systems carries profound ethical and economic consequences. By embracing inclusivity as a core design principle, platforms can create more equitable and engaging online environments that promote diversity and mitigate algorithmic pigeonholing. Our research offers practical insights and guidelines for managers aiming to achieve inclusive design at scale.

Related Literature

Our work contributes to multiple research streams examining how platform design choices can create both business value and social impact. Substantively, we contribute to a nascent but growing body of work on the design of inclusive products and on firm initiatives on diversity, equity, and inclusion. Aneja et al. (2023) demonstrate that highlighting minority business ownership on a restaurant platform increases both customer engagement and firm performance. Hartmann et al. (2023) document an increase in demographic diversity in online display ads and show that this increase correlates with higher user engagement. Shulman and Gu (2023) provide a theoretical framework showing how organizational factors influence investments in inclusive product design. We extend this literature by providing large-scale causal evidence on how intentional platform design for inclusion impacts core business metrics.

Conceptually, research examining how platform design choices, both algorithmic and informational, shape user behavior provides critical context for our work. Recent studies show how changes in core platform features affect user engagement and platform outcomes. For instance, Sun et al. (2023) show that removing content personalization on Alibaba reduces user engagement and decreases sales. Relatedly, Lei et al. (2023), remove the use of external data provided by a leading search engine to a smaller engine. They find that removing the data access significantly lowers the click-through rate for the smaller search engine. Chen et al. (2023) find that "active" users increase their content discovery when platforms enhance recommendation diversity, although there is no change in overall engagement. On the informational side, Bairathi et al. (2025) provide experimental evidence on how explicit platform endorsements such as "Best Seller" or "Platform's Choice" impact user search and purchase. Building on this work, we show how algorithms explicitly designed to be inclusive can remove structural barriers to content discovery, revealing untapped engagement opportunities for previously underserved user segments.

On the technical front, our work builds upon Silva et al. (2023), which details the engineering implementation of diversification across different Pinterest surfaces. While they describe algorithmic approaches, we provide causal evidence of how inclusive design affects user experience and platform performance through field experiments. Key distinctions include: (1) extensive analysis of heterogeneous treatment effects across user segments, showing how inclusive content policies address previously unmet user needs; (2) nine-month longitudinal analysis demonstrating sustained positive effects beyond initial implementation; (3) discussion of specific mechanisms through which inclusive design creates value; (4) extension to multi-dimensional diversification with body type representation; and (5) focus on Pinterest's highest-traffic interface–the Related Pins surface–which accounts for the majority of user engagement on the platform but was not studied in Silva et al. (2023). These contributions extend the understanding of inclusive design from a technical implementation to a core platform strategy with significant implications for user experience and engagement.

Lastly, our work connects to broader discussions about platform responsibility in the digital age. As online platforms increasingly influence public dialogue and individual self-perceptions, questions of representation and inclusion have gained prominence (Barocas and Selbst, 2016; Corbett-Davies et al., 2023). Studies examine how platforms can promote fairness through intentional design choices. For example, Geyik et al. (2019) show how LinkedIn enhanced gender representation in talent search results, while Singh and Joachims (2018) develop frameworks for ensuring balanced content exposure across user groups. This growing focus on responsible platform design reflects an increasing recognition that platforms must balance business objectives with broader social impact. Our work contributes to this conversation by showing how thoughtful platform design can enhance both user experience and representation. We provide causal evidence that inclusion need not come at the cost of engagement, informing how platforms approach the relationship between social responsibility and business growth.

To summarize, we show how inclusive platform design affects business performance, reveal untapped engagement opportunities from removing structural barriers to content discovery, and offer practical guidance for implementing inclusion initiatives at scale. Our findings establish that platforms can leverage design choices to simultaneously enhance user experience and promote inclusion, transforming how businesses think about the relationship between social impact and growth. To the best of our knowledge, no prior work causally demonstrates the effectiveness of an *inclusive-by-design* product strategy in a real-world field setting.

The remainder of the paper is structured as follows. Section 2 provides an overview of the institutional setting. Section 3 describes the experiment and data. Section 4 presents the main results. Section 5 discusses the mechanism with treatment effect heterogeneity and quantifying supply-side changes. Section 6 discusses the managerial implications. Section 7 enumerates robustness checks, and Section 8 concludes.

2 Platform Design and Representation Challenges

Digital platforms increasingly face design trade-offs between algorithmic relevance and societal values like fairness (Hardt et al., 2016; Corbett-Davies et al., 2017; Menon and Williamson, 2018). Prior work has shown how design changes Geyik et al. (2019) can improve representation. Building on this platform-design perspective, we investigate the impact of inclusive platform design at Pinterest, a global online platform used for content discovery, visual inspiration, and shopping. Pinterest is a visual discovery platform and each piece of content on it is called a Pin. Pins are bookmarks that people use to save content they like and they can be either images, videos, or products. Figure A1 in the Online Appendix shows examples of Pins served to a sample user. Pinterest serves over 518 million monthly active users[2] and has a corpus of over 12.5 billion Pins. To keep the discussion general, we refer to Pins as *items* or *content* in the paper.

Users discover content through multiple surfaces such as a personalized home feed, search, related Pins, and related products. All the surfaces are powered using in-house content curation algorithms that serve personalized (to various degrees) content to the users. Our study focuses on the platform's most heavily used interface – the Related Pins surface, which shows users visually similar content after they click on a focal item. Parallels from other platforms include the "products related to this item" carousel on Amazon product pages, and the playlist sidebar shown to a user while watching a video on YouTube. Figure 1 illustrates this interface, showing content presented to a user after selecting an initial item (shown at the top).

On the content side, each item on Pinterest that includes people has an underlying skin tone signal. This signal classifies the item into one of four buckets based on the skin tone range of the people depicted in the image (Fawaz et al., 2020). For the purposes of this work, we will

[2]Pinterest Announces First Quarter 2024 Results, Reports 23% Revenue Growth and More Than Half A Billion Monthly Active Users. https://investor.pinterestinc.com/financial-results/quarterly-results/default.aspx

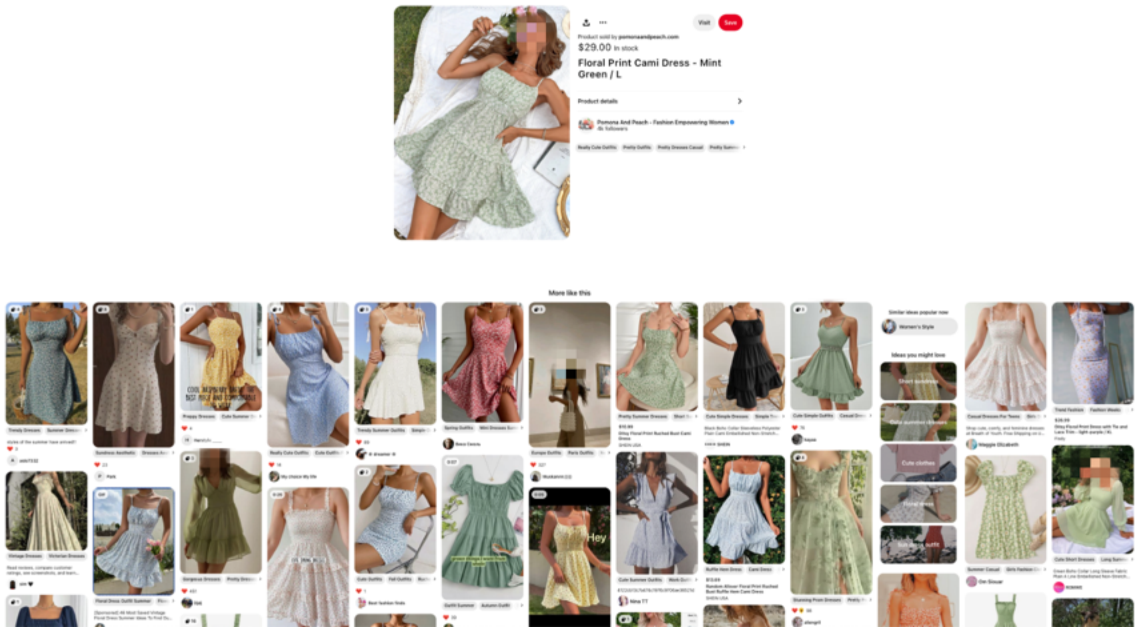


Figure 1: Suggested Content on the Related Pins Surface After Clicking on a Focal Item

call the skin tone ranges "Lightest", "Second Lightest", "Second Darkest", and "Darkest". If an item does not have the image of a person, then the skin tone signal has no bucket assignment. Throughout the paper, we refer to items that are classified as either having the "Second Darkest" or the "Darkest" skin tone bucket as "Deeper" skin tone content. Figure A2 shows examples of items from each skin tone bucket.

2.1 The Need for Inclusive Platform Design

We focus on Pinterest's Related Pins surface, a discovery interface that promotes related content. When a user selects a focal item of interest, the platform shows visually similar content. While the similarity between the focal and suggested content items is a basic criterion for showing relevant content, often, the definition of similarity can be myopic and may overlook opportunities for relevant diversification.

Visual similarity algorithms tend to prioritize features such as color, pattern, and composition – including the skin tone of people featured in the content. This creates a systematic pattern where interactions with content featuring people of one skin tone predominantly lead to recommendations of similar skin tones. Figure 2 illustrates this effect, showing how content featuring lighter skin tones dominates the surface. This distribution represents the pre-intervention exposure of content across skin tone categories, revealing a significant imbalance that *may not* necessarily reflect the underlying preferences of the user base.

These representation patterns create structural barriers to content discovery. Pinterest's visual nature makes addressing these challenges particularly complex but also presents unique opportunities. Consider fashion content like the summer dresses shown in Figure 1. Users naturally

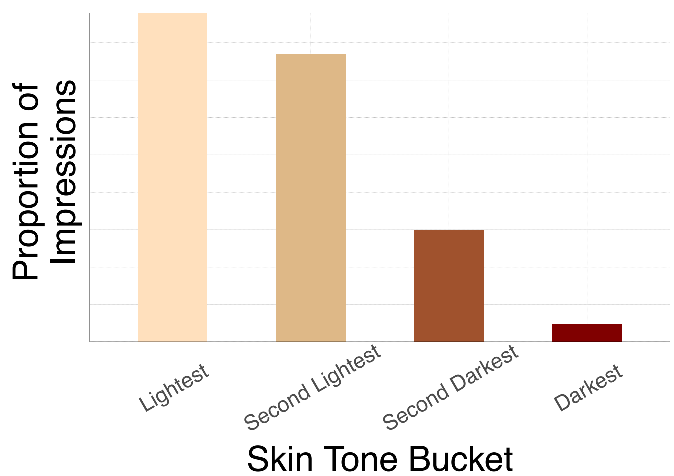


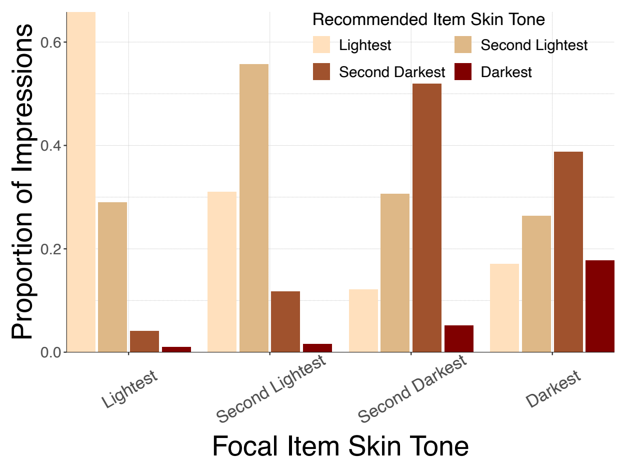
Figure 2: Pre-treatment Distribution of Impressions by Skin Tone
Note: Y-axis labels have been masked to preserve confidential information

expect to see items similar in style, texture, and design, but conventional approaches prioritize visual similarity in all dimensions, including the skin tone of models wearing the clothing. Figure 3a translates this observation to a data-backed argument. The figure shows the proportion of impressions of suggested items by skin tone conditional on the skin tone of the focal item. For example, the left-most stack shows the proportion of suggested items in the Lightest (L), Second Lightest (SL), Second Darkest (SD), and Darkest (D) skin tones condition on the focal item belonging to the Lightest (L) skin tone. We see the extent of visual similarity here since the majority of the items in the recommended set belong to the Lightest skin tone. This pattern generally holds for other skin tones as well, except for focal items with the Darkest skin tone bucket, where the recommended set is more balanced. These patterns highlight systematic skews in content presentation that may limit users' ability to discover diverse content that matches their interests.

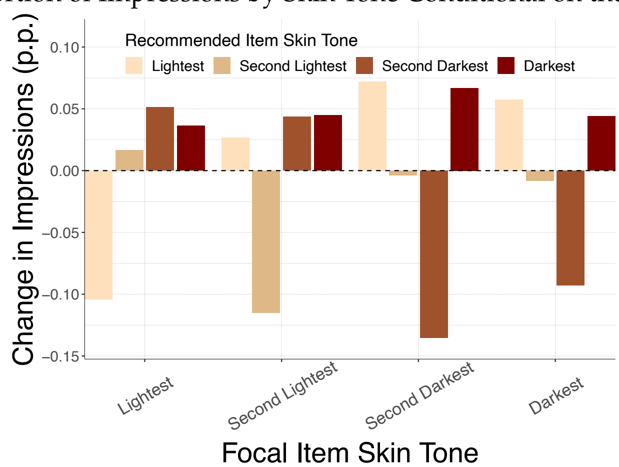
These self-reinforcing patterns have significant business and user experience implications. In categories such as fashion and beauty, where the platform has a strong presence, diverse representation directly affects content relevance and user satisfaction. The platform risks marginalizing users who cannot find content that resonates with their identities or interests, potentially limiting both user engagement and market reach. As highlighted by the user quote in the introduction, and discussed in Strapagiel (2020) and Asare (2021), this experience gap can fundamentally shape how users perceive the platform's value and relevance to their needs.

Beyond individual user experience, limited representation has broader implications for platform strategy and growth. Platforms that optimize purely for short-term engagement metrics risk creating *filter bubbles* (Pariser, 2011) and *echo chambers* (Sunstein, 2001) that confine users to narrow silos of content. This not only constrains users' ability to discover new interests and perspectives but also hinders the platform's capacity to learn and adapt to evolving user preferences (Chaney et al., 2018; Jiang et al., 2019). Over time, narrowly focused content can impoverish the entire platform and potentially limit its market reach and appeal (Mansoury et al., 2020).

The business case for addressing these representation challenges extends beyond immediate social impact. First, inclusive design can reveal untapped engagement potential by connecting users with content they value but struggle to discover. Second, platforms with diverse content ecosystems are better positioned to serve global audiences with varied preferences and expectations. Third, as user expectations around representation evolve, platforms that proactively address these concerns may build stronger user trust and loyalty.



(a) Proportion of Impressions by Skin Tone Conditional on the Focal Item



(b) Change in Impressions (p.p.) by Skin Tone after Diversification

Figure 3: Distribution of Skin Tones in Suggested Content Conditional on the Focal Item Skin Tone

These perspectives offer strong motivation to fundamentally rethink its approach to diversifying content distribution. Research in academia, as well as industry, has recognized this need and has made substantial progress over the past decade in diversifying recommendations and search engine results with different constraints (Carbonell and Goldstein, 1998; Zhu et al., 2007; Agrawal et al., 2009; Wilhelm et al., 2018; Chen et al., 2017; Wang et al., 2020; Chen et al., 2023). However, much of the previous work has defined diversity in terms or content categories or subject matter. A salient but overlooked dimension is the representation of people shown in the content itself.

Our work extends this literature by introducing skin tone as a feature over which visual content can be effectively diversified to create more inclusive platform experiences.

This idea provides a foundation for our experiment, which tests whether balanced content distribution can enhance representation while maintaining or improving user engagement. The next section details our experimental design and measures for evaluating both inclusion and business outcomes.

3 Experiment Design

We test the impact of exposure to inclusive content using a field experiment on Pinterest's Related Pins surface[3] for users in USA, Canada, Great Britain, Ireland, Australia, and New Zealand. During the experiment, users interacting with fashion-related focal items are triggered into either the control or treatment conditions. In the control condition, users are shown content based on the current system, i.e., the status quo. The status quo content tends to be more concentrated in terms of skin tone ranges displayed. In the treatment condition, users are shown more diverse content in terms of the visual skin tone of the underlying items.

Specifically, the new content policy strives for a more balanced exposure across all four skin tone categories. The policy is operationalized using a two-stage deep neural network that is common in large-scale industrial applications (Covington et al., 2016; Wang et al., 2018; Meta, 2019; Huang et al., 2020; Zhang et al., 2020). Skin tone diversity is promoted using a Determinantal Point Process, a probabilistic model of repulsion that can be used to diversify a set of items (Kulesza and Taskar, 2012; Wilhelm et al., 2018). For technical implementation details on other Pinterest surfaces, please refer to Silva et al. (2023).

Figure 4 presents an example of a focal item and the corresponding items shown in the control and treatment conditions. Skin tone diversity among the items served in the treatment condition is apparent. See Figure A3 for another example.

3.1 Data

Our working data contains a randomly drawn proportion of all users in the experiment. The sample contains 319,235 users in control and 319,448 users in treatment from March 28, 2023, to April 28, 2023. These users represent an undisclosed fraction of users who were randomized into the experiment. We provide randomization checks using pre-treatment historical data in Online Appendix C.

For each user, we observe their experimental assignment and their activity on the platform. Importantly, for each query request generated by the user, we observe the focal item for which the content suggestions are generated, the items suggested by position, and whether the user engaged with them. We also observe the time stamp of each query which we use for ruling out novelty effects.

[3] The experiment was conducted in April 2023.

(a) Focal Item

(b) Control Suggestions

(c) Treatment Suggestions

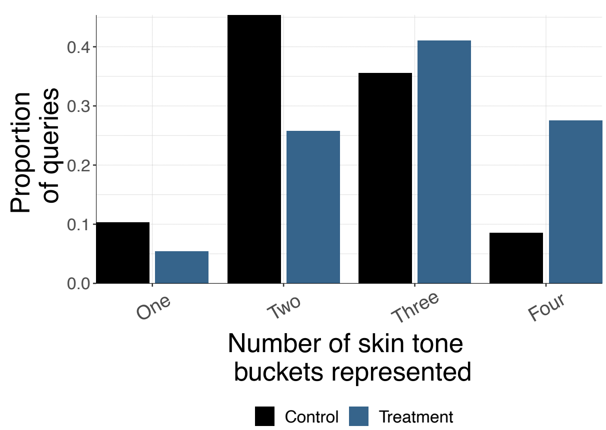
Figure 4: Suggested Content: Treatment vs. Control

3.2 Inclusion Metrics

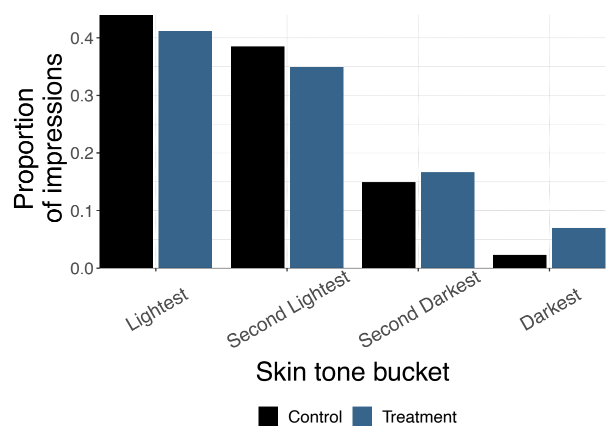
The new content policy diversifies the suggested items for each fashion-related query[4] made by users in treatment. This resulted in users' content exposure becoming visually more inclusive and having a more even distribution across skin tone ranges. We quantify the distributional impact of diversification on suggested items using the following metric

$$\text{Div@k(P)} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \prod_{d_i \in \mathcal{D}} \mathcal{I}[d_i \in P_k(m)] \tag{1}$$

where m is a query belonging to $\mathcal{M}$, the entire set of queries. P is the content policy that serves k ordered items. $\mathcal{D}$ is the diversity dimension and d_i is a single element of $\mathcal{D}$. In our case, d_i takes on 4 values, one for each skin tone bucket. Intuitively, the measure captures the proportion of queries where all four skin tone buckets are represented among the top-k items. During our experiment, the proportion of queries in which all four skin tone buckets were represented among the top-20 items increased 3-fold in the treatment group relative to the control group, as shown by the last bar in Figure 5a. Furthermore, in attempting to balance the representation across skin tones, the new policy also increases exposure to content with deeper skin tones. In Figure 5b, we plot the proportion of impressions by each skin tone among the top-20 suggested items. We see that treatment increased the proportion of impressions for deeper skin tone content by ∼33%. We investigate the impact of this provision of more visually inclusive content on user engagement. In Appendix B, we show the shift in distribution between treatment and control using other measures of diversity.



(a) Proportion of queries with different numbers of skin tone buckets represented in the top-20 results



(b) Proportion of impressions by each skin tone bucket

Figure 5: Distribution of recommended content in treatment vs. control.

[4]A query is defined as a click to a focal item that results in content suggestions being presented to the user.

3.3 Outcomes of interest

Our goal is to investigate how inclusive design policies influence content consumption, as measured by both overall engagement and the diversity of content engaged. We concretize these notions of engagement and consumption diversity using four metrics, which collectively form our outcome variables in subsequent analysis.

3.3.1 Overall Engagement

We measure overall engagement using the number of total items *Repinned* from the suggested content set by each user. A Repin is equivalent to saving an item. For generality, we will use the term "saving" or "saved" henceforth. Like most other measures of online user activity, items saved has a long-tail (Brynjolfsson et al., 2006). Consequently, we consider the top-20 suggested items for a given query, which collectively represent a significant majority of the total engagement. Once the items are shown the user can choose to engage with any number of them (or not engage at all). We count the number of items saved by a user for a given query and then aggregate that over all the queries made by the user to get the total number of items saved. We mainly focus on the number of items saved since it is a key engagement metric used by the platform to make product decisions. In Online Appendix D, we check the robustness of our results with other measures of engagement.

3.3.2 Consumption Diversity

Recall that the primary motivation behind the platform redesign is to enhance the diversity of content that users discover and engage with, particularly across underrepresented visual characteristics. In this spirit, we measure consumption diversity using three metrics that capture different aspects of diversity – 1) the number of deeper skin tone items saved, 2) the Shannon entropy of items saved computed over the four skin tone buckets, and 3) the proportion of users engaging with content from all four skin tone buckets.

Focusing on consumption diversity is important for two reasons. First, it gives insight into how the engagement patterns are likely to change once the new system is launched platform-wide. This is valuable for managers as it helps inform subsequent content strategy for platform growth and monetization. Second, extant research has shown that diverse consumption is associated with better long-term outcomes, such as retention (Anderson et al., 2020; Wang et al., 2022). If our treatment causes consumption diversity to increase, then this could potentially have favorable long-term implications.[5]

The first metric for consumption diversity is similar to the metric for overall engagement, except that it only focuses on items with deeper skin tone content, which are items classified as having either the 'Second Darkest' or the 'Darkest' skin tone bucket. The experiment effectively

[5]Our experiment ran only for four weeks and hence cannot draw a causal link between short-term consumption and long-term outcomes.

increased the proportion of impressions (relative to the control group) for deeper skin tone items (Figure 5b), and hence the increase in engagement with this content gives a direct measure of the impact of treatment on consumption diversity.

We complement this with Shannon entropy, which accounts for the evenness in the distribution of engagement over the four skin tone buckets (Holtz et al., 2020; Chen et al., 2023), rather than concentrating on a smaller subset. We compute the Shannon Entropy for each user over the 4 skin tone bucket types as follows:

$$S^i = -\sum_{b \in \mathcal{B}} s_{bi} \ln(s_{bi}) \tag{2}$$

where S^i is the Shannon entropy for user i, s_{bi} is the share of engagement for user i belonging to skin tone bucket b, and $\mathcal{B}$ is the full set of skin tone buckets, which has cardinality 4 in our case. The value of this metric ranges from 0 to 2, where $S^i = 0$ corresponds to no diversity (only one skin tone bucket in the items engaged), while $S^i = 2$ corresponds to an even distribution across the four buckets.

Finally, our last metric captures whether users are engaging with content from all 4 skin tone buckets. This provides an intuitive and easily interpretable measure for the proportion of the user base engaging with diverse content.[6] This is a binary variable for each user, which takes the value 1 if the user engaged with content from all four skin tone buckets and 0 otherwise.

4 Results

Consider a user i with treatment status given by T_i, where T_i is a binary variable that takes the value 1 when user i is in treatment and 0 otherwise. As the user explores content on the platform, the user clicks on focal items, generating a request for content suggestions. The platform then produces an ordered set of items that are relevant to the user's query. Once the suggested items are surfaced, the user can choose to engage with any number of them (or not engage at all). Our primary measure of engagement is *saving* an item. We collect activity data for all the users during the course of the experiment, aggregate it to the user level, and do our main analysis with the user-level data. We estimate treatment effects using the following regression

$$Y_i = \alpha + \beta T_i + \epsilon_i \tag{3}$$

where Y_i is a measure of engagement that varies with the outcome of interest. For overall engagement, it is the total number of items saved by the user. β captures the effect of getting exposed to more diverse skin tone content on user engagement.

[6] A limitation of just using Shannon entropy is that we cannot easily distinguish whether an increase in Shannon entropy is coming from consuming a more even proportion of items across the buckets conditional on the number of buckets a user engages with or just from engaging with more number of buckets.

Arguably, one can estimate the treatment effect at different levels of data aggregation, e.g., at a query level. However, once exposed to the treatment, any subsequent queries a user makes could be influenced by the treatment itself. In the extreme case, there could be strong path dependence among the queries where the user clicks on the results of the previous query to generate new content suggestions and so on. Hence, we focus our main analysis on aggregate user-level results that can answer a simple but pivotal question, "How do the top-line metrics change once users are exposed to diverse and inclusive content suggestions"? Nevertheless, we show the robustness of our results by re-estimating Model 3 at the query level, at the user-query-item level, and by using only the first query of each user. All robustness tables are present in the Online Appendix.

4.1 Average Treatment Effect

We estimate the treatment effect using Model 3 on different measures of engagement described in Section 3.1. The results are shown in Table 1. In all columns, we suppress the intercept to protect sensitive information. Heteroskedasticity-robust standard errors are shown in parentheses.

Columns 1 and 2 show the impact of the treatment on overall engagement (total number of items saved by the user) and engagement with deeper skin tone content (total number of second darkest and darkest skin tone items saved) respectively. Since both these measures are long-tail outcomes with a huge mass at zero, we estimate the treatment effect on a log scale using a Quasi-Poisson specification. This also has the added benefit of directly reading the treatment effect as a percentage change.[7]

Next, we measure the impact on Shannon Entropy (Column 3) which accounts for the evenness in the distribution of engagement over the four skin tone buckets (Holtz et al., 2020; Chen et al., 2023). For ease of interpretation, we normalize the raw outcome in the control group to 100 so that the coefficient on the treatment indicator can also be read as a percentage change. The last measure captures whether users are engaging with content from all four skin tone buckets. This provides a simple and easily interpretable measure for the proportion of the user base that is engaging with diverse content. This is a binary variable and we estimate it using a logistic regression.

We find that overall engagement rates are stable in treatment and control (Column 1). Although the magnitude shows a small increase of ∼0.5% increase in engagement, this change is indistinguishable from zero. The result is consistent if we define engagement using other common measures such as the number of successful queries (queries with at least 1 save) or the daily logins (Online Appendix, Table D3).

While overall engagement rates remain stable, consumption diversity goes up substantially, as shown in Columns 2, 3, and 4. We find that both engagement with deeper skin tone content and the Shannon entropy of consumption increase by ∼15%, and the proportion of users engaging with content from all four skin tone buckets increases by 70% ($\exp(0.5363) - 1 = .70$). These

[7] Arguably, one can also estimate this model by $\log(\cdot + 1)$ transforming the outcome. However, recent research (Chen and Roth, 2023) has shown that ATEs for log-like transformations cannot always be interpreted as approximating a percentage change since they depend on the units of the outcome. The authors recommend a Poisson/Quasi-Poisson for expressing ATE as a percentage.

results are strong, significant, and important from a managerial perspective. Taken together, they indicate that platforms can meaningfully increase consumption diversity without hurting top-line engagement rates.

Table 1: Average Treatment Effect on Engagement and Consumption Diversity

	Overall Engagement	Consumption Diversity		
Dependent Variables:	Total Saves	Saves of Deeper ST Content	Shannon Entropy	Saved All 4 ST (Binary)
Model:	(1)	(2)	(3)	(4)
	Quasi-Poisson	Quasi-Poisson	OLS	Logit
Variables				
Treatment	0.0046	0.1492**	0.1463**	0.5365**
	(0.0197)	(0.0316)	(0.0083)	(0.0258)
Fit statistics				
Observations	638,683	638,683	638,683	638,683
Log-Likelihood	-170,730,442.9	-249,880,080.0	-4,610,119.8	-35,994.5

This table shows the average treatment effect of inclusive content on user engagement. The first column shows estimates from a Quasi-Poisson regression of the total number of items saved on the treatment indicator (Results with OLS are in the Online Appendix). Coefficients in all columns can be interpreted as percentage change. Columns 2, 3, and 4 regress measures of consumption diversity on the treatment indicator. Heteroskedasticity-robust standard errors are shown in parentheses. Signif. Codes: **: 0.05, *: 0.1

4.2 User Preferences vs. Mechanical Effects

Admittedly, one may worry that the consumption diversity effects shown in Table 1 are largely mechanical. For example, content that receives increased visibility naturally receives more engagement. Although this concern is plausible, it is important to contextualize it within our broader motivation for inclusive platform design. From an existentialist perspective, Sartre's "existence precedes essence" principle applies here (Sartre, 2007): the value of platform design choices is ultimately defined by how users interact with them, not by predetermined algorithmic intentions. Content discovery interfaces should fundamentally connect users with content they value, and a direct test of their effectiveness is whether users meaningfully engage with the presented content. In this section, we provide analytical evidence distinguishing between user preference satisfaction and mechanical visibility effects.

To begin, it is worth emphasizing that exposure does not automatically translate to engagement. User agency is critical in separating passive consumption from active engagement. To this end, it is important to select appropriate metrics that capture meaningful interaction rather than incidental attention. Surface-level or "up-the-funnel" metrics such as clicks require minimal user investment and are more susceptible to mechanical effects. Instead, we focus on content-saving behavior, which represents a deliberate user action signaling genuine interest and intention for future reference. This metric represents a more *meaningful* form of engagement and serves as a key performance indicator for the platform's overall effectiveness.

We further provide four pieces of data-based evidence to rule out mechanical effects. First, if users don't prefer the suggested content, they can either explore deeper in the content stream or disengage entirely, both of which would result in lower overall engagement. If this were the case, we would have seen a lower overall engagement in the treatment group. However, as shown in

Column 1 of Table 1, the overall engagement rate is stable, suggesting that content relevance is maintained despite the redistribution.

Second, we estimate the impact on engagement with deeper skin tone content normalized by the exposure to such content, i.e., we estimate the *save rate* of deeper content. Using query-level data, we find that the *save rate* of deeper skin tone content increases significantly in the treatment group (Table E8), indicating higher user receptivity to this content beyond mere visibility effects.

Third, we address the concern that the treatment effect is essentially driven by position bias, where deeper skin tone content receives more attention due to its higher placement in the interface. To test this, we use the user-query-item level data to estimate treatment effects for each position separately. Figure E5 in the Online Appendix echoes the result from Table 1. Across the top-20 positions, overall engagement remains stable, and engagement with deeper skin tone content goes up significantly. This position-controlled analysis provides strong evidence that users are actively selecting more diverse content rather than passively responding to placement.

Finally, we examine explicit user *dissatisfaction* signals to evaluate preference alignment. The platform interface allows users to *hide* content items with a single click, providing a direct negative feedback mechanism. We estimate the impact of inclusive design on the number of items hidden. This allows us to gauge whether users are being exposed to irrelevant content in the pursuit of diversity. If users were being shown content that does not satisfy their preferences, then we would see an increase in the *hide rate*. However, this is not the case. Table F9 shows the results. We find that the number of hides is slightly reduced in the treatment group, although the effect is not statistically significant. This evidence, at least in part, assures us that the new algorithm is catering to user preferences while actively promoting content skin tone diversity.

Beyond these immediate indicators, we assess longer-term persistence effects using data collected nine months after platform-wide implementation. This longevity analysis (see Section 6.1) provides additional perspective on whether the observed effects represent genuine preference satisfaction rather than novelty-driven temporary changes in behavior.

4.3 Treatment Effects by User Characteristics

We check for heterogeneity in treatment effects across key demographics – user's self-reported gender, geography, and the user's tenure with the platform. Investigating the differential treatment impact of new product features is important from a product manager's perspective. Ideally, platforms would like to make universally appealing product changes. Seldom, however, is the case that a proposed change benefits the entire spectrum of users. Hence, it is common to establish guardrails and make feature launch decisions after ruling out potential harms to any pre-defined user sub-groups.

About 77% of the users report their gender to be female, and ∼79% of the users are from the US. We do not find any meaningful differences across user demographics. Table G10 in the Online Appendix shows the estimated coefficients. We find consistent results of stability in overall engagement and an increase in engagement with deeper skin tone content.

Similarly, we estimate treatment effects for new vs. older platform users, where new users are those who joined the platform during the experiment period. New users make up ∼4% of our sample. We do not find any substantial difference in the impact of the treatment on engagement. The results are shown in Table G11.

4.4 Treatment Effects by Content Category

Currently, the inclusive content policy only applies to fashion-related queries, which is one of the largest interest categories on Pinterest. We explore whether there is heterogeneity in treatment effects across the main sub-categories of interest. These sub-categories include content from women's styles, dresses, occasion outfits, bottoms, jewelry and accessories, hair, nails, etc. Ex-ante, we may expect that for some of these interest sub-categories, skin tone could either be more salient or more relevant to the query, and hence, the treatment effect could be more pronounced.

To estimate treatment effect heterogeneity by sub-category, we use data at the query level and classify the focal item into one of the top 20 sub-categories by content supply. We then estimate heterogeneous treatment effects for these sub-categories separately.

The results are shown in Figure H6a for overall engagement and Figure H6b for engagement with deeper skin tone content. For overall engagement, the treatment effect is fairly stable across all interest sub-categories. However, for deeper skin tone content, we see substantial heterogeneity with women's style, dresses, and occasion outfits seeing more than a 25% increase in engagement.

5 Who Benefits from Inclusive Design?

The aggregate results show that inclusive design can enhance representation without compromising engagement. However, these averages could potentially mask critical heterogeneity in how users respond to more diverse content distribution. Understanding this variation is essential to unpacking the mechanisms driving the results and to informing platform strategies that balance inclusivity with user satisfaction. In this section, we analyze two dimensions of heterogeneity: (1) how users' historical engagement shapes their response to more inclusive content distribution and (2) how strategic content redistribution affects exposure across skin tone categories. Together, these analyses reveal which users benefit most from inclusive design and why.

5.1 Heterogeneity in User Preferences

An aggregate *null effect* could either be driven by a uniform near zero effect across the spectrum, or it could be masking offsetting effects across different user segments. To identify these patterns, we examine how users historically discovered and engaged with diverse content before our experiment. Specifically, we posit that the new inclusive content policy improves the match quality between suggested items and preferences for some users. This is possible if users have heterogeneous preferences over content representing different skin tones. Since deeper skin tone content was initially underrepresented (as compared to lighter skin tone content), users with a preference for this type of content may have been underserved. The new policy, which strives for a more even representation of content, could potentially lower the costs of accessing deeper skin tone content for users with a strong preference for it.

To test for this, we consider users' prior exposure to and engagement with deeper skin tone content. We segment users in our experiment on two dimensions – diversity of content exposed to in the 3-month window prior to the experiment and engagement (saves) with deeper skin tone content in the 3-months prior to the experiment.[8] We include both exposure and engagement in this segmentation because it helps us identify users who actively sought diverse content despite potential barriers to discovering it.

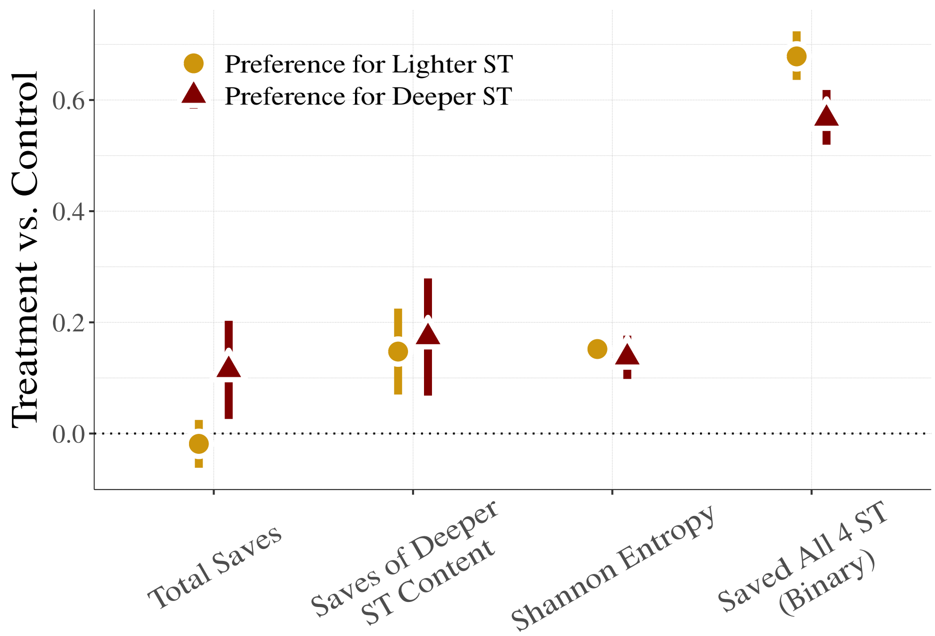


Figure 6: Heterogeneous Treatment Effects on Engagement and Consumption Diversity

We measure content exposure diversity using Shannon Entropy (Holtz et al., 2020; Chen et al., 2023). Historical engagement with deeper skin tone content is calculated as the number of deeper skin tone content items saved divided by the total number of impressions. This gives us a direct measure of how much the user prefers deeper skin tone content, normalized by all the items they have been exposed to. We then use the 80-20 rule to create a 2x2 matrix based on whether a user was in the top quintile of these dimensions or not. We employ the 80-20 rule since this approach aligns with the Pareto principle commonly observed in user behavior patterns, where a smaller subset of users often exhibits distinct behavioral characteristics. Our segmentation effectively isolates users who demonstrated the strongest preference signals while maintaining sufficient sample size for statistical analysis. Summary statistics and robustness checks for this segmentation

[8]For this analysis, we do not consider new users since they do not have any pre-treatment historical data.

are shown in Online Appendix I. It is important to note that the platform does not assume or attempt to infer a user's skin tone based on the content they are engaging with. Signals from content interactions are only used to infer preferences and not identities.

After segmenting the users, we classify users who are in the top quintile of exposure to diverse content and the top quintile of engagement with deeper skin tone content as users with *preference for deeper skin tone content*. The remaining users are classified as users with *preference for lighter skin tone content*. Under this classification scheme, we re-estimate the treatment effects for both user groups separately. The results are shown in Figure 6.

We find that the treatment effect on overall engagement for some users is significantly greater than zero. Users with *preference for deeper skin tone content* see about an 11% increase in overall engagement. The treatment effect on overall engagement for users with *preference for lighter skin tone content* is negative but statistically indistinguishable from zero. When aggregated and weighted by the corresponding number of users in each group, this gives a total effect that is close to zero. We provide robustness checks for this specification, including the user segmentation, in Online Appendix I.

These findings suggest that historical platform patterns may have created artificial barriers for users seeking diverse content, constraining their engagement. When the barriers are removed, these users engage more actively with the platform. Importantly, this enhanced engagement occurs without disrupting the experience of other users, explaining the stable overall engagement we observe at the platform level.

5.2 Supply-Side Content Redistribution

The heterogeneous effects in Figure 6 suggest that user preferences play a key role in explaining stable overall engagement. However, this raises a follow-up question: How did the platform's content redistribution strategy affect exposure patterns, and did this redistribution change relevance for different user segments? We address this next by analyzing shifts in the distribution of suggested content.

The design intervention rebalanced the distribution of skin tone categories in suggested content while preserving core relevance signals (e.g., visual similarity, user preferences). Specifically, the inclusive design approach increased the proportion of queries where all four skin tone categories were represented in the top-20 suggested items by 3-fold (Figure 5a). Content featuring deeper skin tones (Second Darkest and Darkest) saw a 33% increase in impressions relative to the control group (Figure 5b). Despite these shifts, lighter skin tone content remained prominent, accounting for more than 70% of impressions in the treatment group. This calibrated balance ensured that underrepresented content gained visibility without compromising the overall content relevance.

The specific changes in content distribution reveal why different user segments respond distinctly to the intervention. Consider users who historically engaged primarily with lighter skin tone content, i.e., users with a *strong preference for lighter skin tone content*. In the control condition, these users typically see 16 out of 20 items featuring lighter skin tones per query. Under treatment, this number decreases to 14 items – a modest 12.5% reduction that maintains their ability to find preferred content. See the top two lines in Figure 7.

The impact is more substantial for users who historically sought content featuring deeper skin tones, i.e., users with a *strong preference for deeper skin tone content*. These users see their preferred content increase from an average of 4 items to 6 items per view – a 50% increase in relevant content availability (bottom two lines in Figure 7). While the shift from 4 to 6 deeper skin tone items per query may appear modest in isolation, it represents a significant structural change when considered at scale. The cumulative effect across a user's session—which typically includes multiple queries—leads to substantially greater exposure diversity over time. At the platform level, this translates to millions of additional impressions for previously underrepresented content daily, fundamentally reshaping the content ecosystem. This substantial improvement in content accessibility explains the ∼11% increase in overall engagement shown in Figure 6.

A key concern is whether promoting representation balance reduces the average quality of recommended content. To test this, we compare the historical engagement rates (e.g., saves per impression) of items shown in treatment and control. If the new approach sacrificed relevance for diversity, we would expect lower-quality items to surface in the treatment group. However, we find no significant difference in the quality of recommended content between treatment and control groups (Table 2).

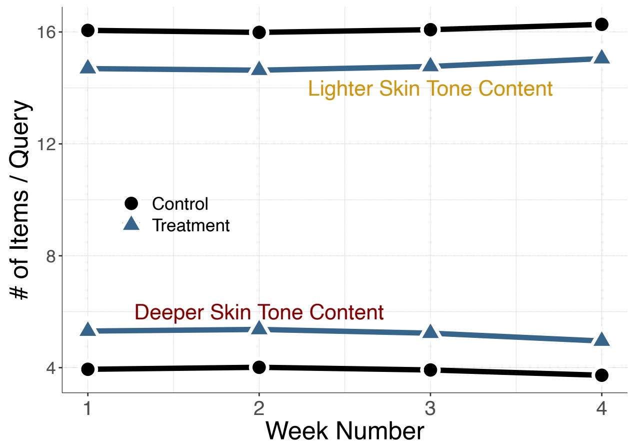


Figure 7: Change in the Supply of Lighter and Deeper Skin Tone Content (Treatment vs. Control)

Dependent Variable:	Historical Item Quality	
	Overall Quality	Position Adjusted Quality
Model:	(1)	(2)
Variables		
Treatment	0.0055	0.0054
	(0.0302)	(0.0301)
Fixed-effects		
Position		✓
Fit statistics		
Observations	26,822,658	26,822,658
Log-Likelihood	-118,590,259.0	-118,589,539.8

*Clustered (User) standard errors in parentheses. Signif. Codes: **: 0.05, *: 0.1*

Table 2: Quality of Items Suggested in Treatment and Control

These findings challenge conventional assumptions about inevitable trade-offs between representation and engagement. By strategically reallocating visibility to underrepresented content—rather than enforcing rigid quotas—the inclusive design approach achieves a more balanced representation while maintaining the platform's core discovery value and user experience.

6 Scaling Inclusive Design

We examine the managerial implications of our research from two critical angles. First, we analyze the longer-term impact on user engagement and consumption patterns several months after the platform-wide implementation of the inclusive design policy. Second, we evaluate the generalizability of this design approach to other platforms and diversity dimensions, demonstrating its potential to drive inclusivity across a wide range of applications. By considering both the long-term consequences and the broader applicability of our findings, we aim to provide actionable insights for managers and policymakers seeking to foster more equitable and representative digital environments.

6.1 Longer-term Persistence

A crucial question for any platform innovation is whether its effects persist beyond initial implementation. We examine this question using data from $\sim$800,000 users collected over nine months following Pinterest's platform-wide launch of more inclusive content suggestions in May 2023. This analysis reveals how changes in platform design durably affect user behavior and engagement.

We preface this discussion by noting that our analysis here is more exploratory and is meant to serve as a data-based guide while discussing the long-term platform-wide implications of the product launch. Although we cannot interpret the results in the section as clean long-term causal effects since there is no active control group after the product launch, we believe that this longer-term analysis strongly complements our shorter-term causal findings.

For this analysis, we compute our four outcome variables – 1) overall engagement, 2) engagement with deeper skin tone content, 3) Shannon entropy, and 4) engagement with content from all 4 skin tone buckets. We plot the four outcome variables in separate panels in Figure 8. For each variable, we show mean values for control and treatment during the experiment, as well as the means nine months after platform-wide implementation. In each case, we also plot the 95% confidence intervals. The intervals for the post-launch period are much smaller since we have substantially more data, both cross-sectionally and over time. Note that the post-launch sample is a randomly selected subset of users *not* included in the experiment.[9]

The results paint a compelling picture of sustainable positive impact. Overall engagement remains stable in the months following implementation, while the diversity of content consumption continues to significantly exceed pre-launch levels (Figure 8). Although the magnitude of increased consumption diversity moderates somewhat from the initial experimental period, it maintains substantially higher levels than the pre-launch baseline. This pattern suggests that in-

[9]We use a new random subset from the entire user base rather than tracking the same users in the experiment to ameliorate *survivorship bias*. The set of users who are active nine months after the launch are likely to be more engaged, and hence, their outcome measures will not be representative of the entire user base. Therefore, we pull a fresh sample of users from the post-launch period and compare them with the pre-launch baseline. This provides a more representative view of the platform's long-term outcomes.

clusive content distribution creates lasting improvements in user experience rather than merely generating temporary novelty effects.

These sustained elevated levels of consumption diversity are encouraging and have important implications for platform strategy. First, prior research has shown that users with diverse consumption patterns tend to exhibit favorable long-term outcomes (Anderson et al., 2020; Wang et al., 2022). We validate this relationship in our context using historical observational data. In Online Appendix K, we examine how short-term consumption diversity relates to two key long-term metrics: login frequency and overall engagement. Controlling for historical user behavior and demographics, Tables K16 and K17 show that users with higher short-term consumption diversity exhibit better long-term outcomes across all specifications. While we cannot make tight causal claims, these findings suggest potential long-term benefits from sustained increases in content diversity.

Second, and more importantly, the sustained increase in diverse content consumption without any decline in overall engagement reveals previously unmet user needs. The persistence of these effects indicates that historical platform patterns may have created structural barriers to content discovery for certain user segments. This shift in consumption toward more diverse content helps inform the platform's subsequent content strategy while achieving broader social goals of fostering inclusive digital environments.

6.2 Multidimensional Inclusivity with Body Type Diversity

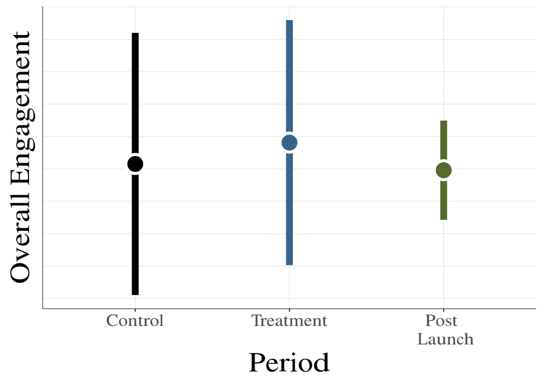
The success of any platform innovation ultimately depends on its scalability and adaptability across different contexts. We demonstrate these qualities by extending the *inclusive-by-design* approach to another critical dimension of visual representation: body type diversity. This extension provides a rigorous test of whether platforms can simultaneously promote multiple dimensions of inclusion while maintaining user engagement.

Similar to skin tone representation, Pinterest classifies content featuring people into four body type categories: "Lower Range", "Lower Middle Range", "Upper Middle Range", and "Upper Range". This classification enables us to examine whether the positive effects of balanced content distribution extend beyond our initial implementation.

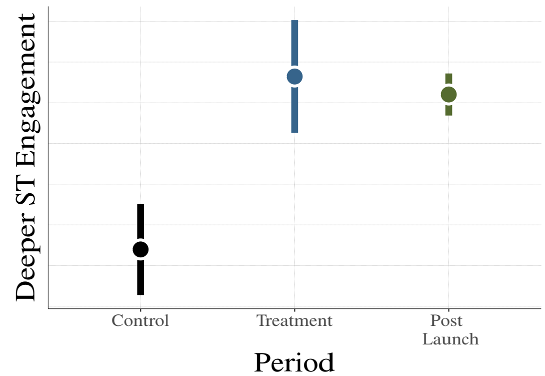
We test this extension through a second field experiment with 839,000 users, building on our initial findings. Half of these users receive content balanced across both skin tone and body type dimensions, while the control group was exposed to content with the skin-tone-balanced distribution already implemented across the platform. It is important to note that this experiment design isolates the incremental impact of adding body type diversity to existing inclusion efforts.

The results strongly validate our approach's scalability. The intervention successfully increases representation – queries showing all four body types increase three-fold, with particular gains in visibility for previously underrepresented Upper Middle Range and Upper Range content (200% increase in impressions).

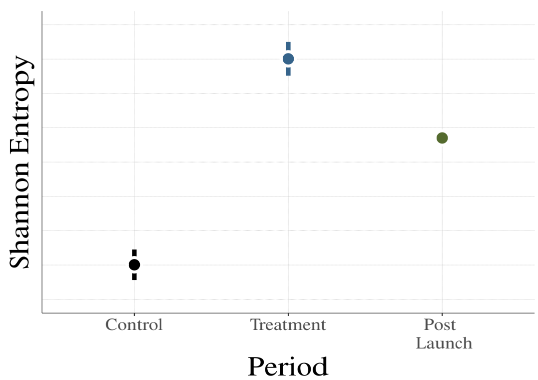
Table 3 shows the impact on overall user engagement and consumption diversity. In this case,



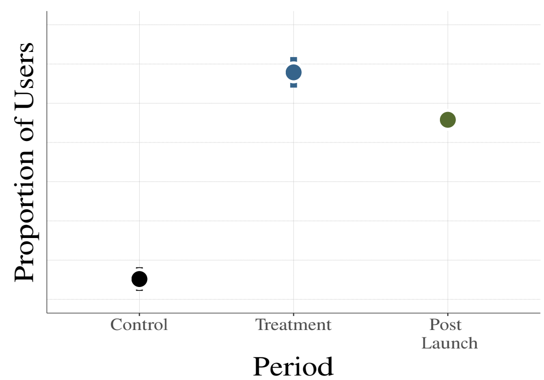
(a) Overall Engagement



(b) Engagement with Deeper ST Content



(c) Shannon Entropy Across ST Buckets



(d) Engagement with Content from all 4 ST Buckets

Figure 8: Overall Engagement and Consumption Diversity 9 Months After the Platform-Wide Launch
Note: Y-axis labels have been masked to preserve confidential information

consumption diversity is defined only over the set of body type buckets. The results are qualitatively similar to the case of skin tone diversification – overall engagement remains stable, and engagement with all four body type buckets increases significantly.

These findings have broad implications for platform strategy. First, they show that inclusive design creates compounding rather than competing benefits when implemented across multiple dimensions. Adding body type diversity to existing skin tone representation further improved diversity metrics, suggesting platforms can address multiple representation dimensions simultaneously. Second, these results challenge the conventional view of defining user preferences solely based to their historical consumption. The positive response to multidimensional representation suggests users value content diversity across various attributes, even beyond those they might explicitly search for, pointing to a broader conception of content relevance than traditionally operationalized in platform design.

The success across both skin tone and body type dimensions suggests that strategic platform design choices can create more inclusive digital spaces while maintaining strong business perfor-

Table 3: Inclusivity by Body Type: Average Treatment Effect on Engagement and Consumption Diversity

Dependent Variables:	Total Saves	Saves of Diverse BT Content	Shannon Entropy	Saved All 4 BT (Binary)
	Overall Engagement		Consumption Diversity	
Model:	(1)	(2)	(3)	(4)
	Quasi-Poisson	Quasi-Poisson	Gaussian	Logit
Variables				
Treatment	0.0010	0.5200**	0.3705**	0.7741**
	(0.0076)	(0.0473)	(0.0195)	(0.1351)
Fit statistics				
Observations	838,304	838,304	838,304	838,304
Log-Likelihood	-183,596,446.9	-29,857.3	-3,023,578.4	-2,294.1

This table shows the average treatment effect of body type and skin tone content diversification on user engagement. The control group receives content diversified based on skin tone only. The first column shows estimates from a Quasi-Poisson regression of the total number of items saved on the treatment indicator. Coefficients in all columns can be interpreted as a percentage change. Columns 2, 3, and 4 regress measures of consumption diversity on the treatment indicator. Heteroskedasticity-robust standard errors are shown in parentheses. Signif. Codes: **: 0.05, *: 0.1

mance. This insight opens new possibilities for platforms to systematically enhance representation across multiple dimensions of diversity, creating digital environments that are representative of the users they serve.

7 Robustness Checks

We perform several robustness checks for our results using different model specifications, alternative definitions of engagement, and testing for novelty effects. All results are shown in the Online Appendix.

1. **Alternate functional forms.** We estimate the average treatment effects using OLS overall engagement and the three consumption diversity variables. The results are consistent and are shown in Table D2.

2. **Alternate outcome variables.** We show the robustness of our results with respect to two other measures of engagement – a) the number of successful queries, where a query is considered successful if the user saved at least one item among the top-20, and b) the number of unique dates a user logged onto the platform and made a query during the experiment after being triggered. The results are shown in Table D3 and are qualitatively similar to our main results.

3. **Additional controls.** We estimate treatment effects after controlling for user demographics, cohort, and user activity index. The user activity index is Pinterest's internal classification of users based on their activity history. The results are unchanged and are shown in Table D4.

4. **Multiple data aggregations.** We estimate treatment effects at the user-query level. This allows us to control for time-varying factors such as the calendar date and the number of days the user has spent in the experiment. The estimates are similar to the ones obtained

using user-level data and are shown in Table E5. Furthermore, we estimate the average treatment effect on overall engagement using user-query-item level data. This is the most granular level of data in which each row represents the item a user was shown, the position of the item in the user's feed, and whether the user saved that item or not (engagement). Recall the new content policy re-organized the user's feed by moving items with different skin tone signals up or down in order to increase the representation of all skin tone buckets. While our results show that the overall engagement levels are stable, it could still be that some positions gain at the expense of other positions. We test for this directly by estimating the treatment effect for each position separately. The results are shown in the Online Appendix in Figure E5a. Again, we find consistent results – overall engagement results for each position are stable and are *not* distinguishable from zero.

5. **Robustness for heterogeneous effects.** Our analysis in Section 5 first segments users into quintiles based on historical exposure diversity and historical engagement with deeper skin tone content. In the Online Appendix, we show the full table of results where the model is estimated jointly using interactions. Further, since users with "*preference for lighter skin tone content*" have a negative but statistically insignificant treatment effect, we show estimates for users across all the quintiles to check whether any segment of users had strong negative results. Again, our robustness results are consistent with the effects presented in Section 5.

6. **Novelty effects.** Novelty effects are commonly observed in experiments relating to product changes (Kohavi et al., 2020). We test the robustness of our results against potential novelty effects in two ways.

   (a) We estimate the treatment effects separately for each week of the experiment at a user-query level. If we expect the results to be primarily driven by novelty effects, we are likely to see the treatment effect dampen as the experiment time progresses. However, contrary to this, we find that the effects are stable and consistent during the experiment period. The results for overall engagement and for engagement with deeper skin tone content are shown in Figure J7 in the Online Appendix.

   (b) A drawback of the above approach is that it does not account for the amount of time a user has spent in the experiment. For example, although the overall treatment effect is stable across the four weeks, this could entirely be driven by new users being triggered into the experiment every week and each cohort of new users exhibits novelty effects that are not captured in Figure J7. We circumvent this limitation by explicitly accounting for the number of days a user has spent in the experiment and then re-estimating the treatment effect for each of these "days-in" experiment time periods. The results are shown in Figure J8. We do not see any systematic pattern of treatment effects increasing or decreasing over time.

8 Discussion

We investigate the impact of inclusive platform design on user engagement at Pinterest, one of the world's largest visual content discovery platforms. Pinterest implemented a strategic redesign of its content discovery interface to achieve a more balanced representation across skin tone ranges. While traditional visual recommendation systems have primarily optimized for narrow content similarity, our approach reconceptualizes platform design to incorporate representation as a core dimension of the user experience.

Specifically, items on Pinterest consist of images or videos and items that include people who have an underlying skin tone signal. This signal classifies the item into one of four buckets, which we label as Lightest, Second Lightest, Second Darkest, and Darkest. The inclusive design approach diversifies content suggestions, resulting in users' exposure to a more balanced distribution across skin tone ranges. This strategic intervention increased the proportion of queries where all four skin tones were represented in suggested content by more than 200%. By addressing the historical under-representation of deeper skin tone content, the platform increased visibility for this content by 33% while maintaining overall visual relevance.

Our field experiment reveals that users respond positively to this platform redesign. Overall engagement remains stable while consumption diversity increases significantly. Specifically, engagement with deeper skin tone content and the evenness of engagement across skin tone categories both increased by 15%, while the proportion of users engaging with content from all four skin tone buckets rose by 53%. We provide evidence that these results reflect the satisfaction of user preferences rather than just mechanical effects.

The heterogeneity in user response to more inclusive content reveals important insights about digital platform design. Users who previously sought but struggled to find diverse content show substantial increases in engagement when given better access to it. This suggests that historical platform patterns may have created artificial barriers for certain user segments, limiting both their experience and their potential platform engagement. By removing these barriers, platforms can unlock new sources of user engagement while creating more inclusive digital spaces. Regarding this segmentation, we note that we do not assume or attempt to infer a user's skin tone based on their engagement; historical signals are only used to learn preferences and not identities.

Finally, we discuss managerial implications from two critical perspectives – long-term persistence diversification effects, and generalizability to other diversity dimensions. For the former, we study user activity and engagement nine months after the platform-wide launch of the new algorithm. Tracking persistence in the effects witnessed during the experiment period allows us to better understand the longer-term implications for the platform's content strategy. We continue to observe elevated consumption diversity with stable engagement levels. Not only does this complement the experimental findings, but it also indicates the platform's potential role in fostering a more inclusive digital culture. A consistently diverse content consumption pattern means that users are being exposed to varied perspectives, cultures, and ideas, promoting inclusivity and improving representation.

For the latter, we show evidence from another large-scale field experiment that exposes users to diverse items based on content body type and skin tone. We find encouraging results here as well – overall engagement remained stable, and engagement with content across all body types increased significantly.

These findings inform how platforms might approach the relationship between inclusion and business growth. Rather than viewing representation as a constraint that potentially limits relevance or performance, our work demonstrates that thoughtfully designed inclusion can be value-creating. By proactively addressing representation in core platform interfaces, digital platforms can enhance user experience, broaden market reach, and potentially strengthen user retention and loyalty.

Through this lens, inclusive design emerges not merely as a social responsibility initiative but as a strategic approach to platform design that recognizes and addresses the diverse preferences of users. We hope this study promotes further research exploring how platforms can leverage design choices to simultaneously enhance user experience and promote representation across various dimensions.

References

Agrawal, Rakesh, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong, "Diversifying Search Results," in "Proceedings of the Second ACM International Conference on Web Search and Data Mining" WSDM '09 Association for Computing Machinery New York, NY, USA 2009, p. 5–14.

Anderson, Ashton, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas, "Algorithmic Effects on the Diversity of Consumption on Spotify," in "Proceedings of The Web Conference 2020" WWW '20 Association for Computing Machinery New York, NY, USA 2020, p. 2155–2165.

Aneja, Abhay, Michael Luca, and Oren Reshef, "The Benefits of Revealing Race: Evidence from Minority-owned Local Businesses," Working Paper 30932, National Bureau of Economic Research February 2023.

Asare, Janice Gassam, "Does Tiktok have a race problem?," Dec 2021.

Bairathi, Mimansa, Xu Zhang, and Anja Lambrecht, "The Value of Platform Endorsement," *Marketing Science*, 2025, *44* (1), 84–101.

Barocas, Solon and Andrew D. Selbst, "Big Data's Disparate Impact," *California Law Review*, 2016, *104* (3), 671–732.

Brynjolfsson, Erik, Yu Jeffrey Hu, and Michael D. Smith, "From Niches to Riches: The Anatomy of the Long Tail," *Sloan Management Review*, March 2006, *47*, 67–71.

Carbonell, Jaime and Jade Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries," in "Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval" SIGIR '98 Association for Computing Machinery New York, NY, USA 1998, p. 335–336.

Chaney, Allison J. B., Brandon M. Stewart, and Barbara E. Engelhardt, "How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility," in "Proceedings of the 12th ACM Conference on Recommender Systems" RecSys '18 Association for Computing Machinery New York, NY, USA 2018, p. 224–232.

Chen, Guangying, Tat Chan, Dennis Zhang, Senmao Liu, and Yuxiang Wu, "The Effects of Diversity in Algorithmic Recommendations on Digital Content Consumption: A Field Experiment," *Available at SSRN 4365121*, 2023.

Chen, Jiafeng and Jonathan Roth, "Logs with Zeros? Some Problems and Solutions*," *The Quarterly Journal of Economics*, 12 2023, *139* (2), 891–936.

Chen, Laming, Guoxin Zhang, and Hanning Zhou, "Improving the diversity of top-N recommendation via determinantal point process," in "Large Scale Recommendation Systems Workshop" 2017.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq, "Algorithmic Decision Making and the Cost of Fairness," in "Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining" KDD '17 Association for Computing Machinery New York, NY, USA 2017, p. 797–806.

__ , Johann D Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel, "The measure and mismeasure of fairness," *The Journal of Machine Learning Research*, 2023, *24* (1), 14730–14846.

Covington, Paul, Jay Adams, and Emre Sargin, "Deep Neural Networks for YouTube Recommendations," in "Proceedings of the 10th ACM Conference on Recommender Systems" RecSys '16 Association for Computing Machinery New York, NY, USA 2016, p. 191–198.

Fawaz, Nadia, Bhawna Juneja, and David Xue, "Powering inclusive search & recommendations with our new visual skin tone model," Pinterest Engineering Blog 2020. Accessed: 2023-06-26.

Geyik, Sahin Cem, Stuart Ambler, and Krishnaram Kenthapadi, "Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search," in "Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining" KDD '19 Association for Computing Machinery New York, NY, USA 2019, p. 2221–2231.

Gomez-Uribe, Carlos A. and Neil Hunt, "The Netflix Recommender System: Algorithms, Business Value, and Innovation," dec 2016, *6* (4).

Hardt, Moritz, Eric Price, Eric Price, and Nati Srebro, "Equality of Opportunity in Supervised Learning," in D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, Vol. 29 Curran Associates, Inc. 2016.

Hartmann, Jochen, Oded Netzer, and Rachel Zalta, "Diversity in Advertising in Times of Racial Unrest," *Working Paper*, 2023.

Holtz, David, Benjamin Carterette, Praveen Chandar, Zahra Nazari, Henriette Cramer, and Sinan Aral, "The Engagement-Diversity Connection: Evidence from a Field Experiment on Spotify," 2020.

Huang, Jui-Ting, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang, "Embedding-based Retrieval in Facebook Search," *CoRR*, 2020, *abs/2006.11632.*

Jiang, Ray, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli, "Degenerate Feedback Loops in Recommender Systems," in "Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society" AIES '19 Association for Computing Machinery New York, NY, USA 2019, p. 383–390.

Kay, Matthew, Cynthia Matuszek, and Sean A. Munson, "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations," in "Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems" CHI '15 Association for Computing Machinery New York, NY, USA 2015, p. 3819–3828.

Kiros, Hana, "Hated that video? YouTube's algorithm might push you another just like it," Technical Report, MIT Technology Review 2022.

Kohavi, Ron, Diane Tang, and Ya Xu, *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*, Cambridge University Press, 2020.

Kulesza, Alex and Ben Taskar, *Determinantal Point Processes for Machine Learning*, Hanover, MA, USA: Now Publishers Inc., 2012.

Lam, Onyi, Brian Broderick, Stefan Wojcik, and Adam Hughes, "Gender and Jobs in Online Image Searches," https://www.pewresearch.org/social-trends/2018/12/17/gender-and-jobs-in-online-image-searches/ 2018. Accessed: 2023-06-26.

Lambrecht, Anja and Catherine Tucker, "Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads," *Management Science*, 2019, *65* (7), 2966–2981.

Lei, Xiaoxia, Yixing Chen, and Ananya Sen, "The value of external data for digital platforms: Evidence from a field experiment on search suggestions," *SSRN*, 2023.

MacKenzie, Ian, Chris Meyer, and Steve Noble, "How retailers can keep up with consumers," Technical Report, McKinsey & Company 2013.

Mansoury, Masoud, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke, "Feedback Loop and Bias Amplification in Recommender Systems," in "Proceedings of the 29th ACM International Conference on Information & Knowledge Management" CIKM '20 Association for Computing Machinery New York, NY, USA 2020, p. 2145–2148.

Menon, Aditya Krishna and Robert C. Williamson, "The cost of fairness in binary classification," in Sorelle A. Friedler and Christo Wilson, eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Vol. 81 of *Proceedings of Machine Learning Research* PMLR 23–24 Feb 2018, pp. 107–118.

Meta, 2019.

Pariser, Eli, *The filter bubble: How the new personalized web is changing what we read and how we think*, Penguin, 2011.

Sartre, Jean-Paul, *Existentialism Is a Humanism*, Yale University Press, 2007.

Shulman, Jeffrey D. and Zheyin (Jane) Gu, "Making Inclusive Product Design a Reality: How Company Culture and Research Bias Impact Investment," *Forthcoming, Marketing Science*, 2023.

Silva, Pedro, Bhawna Juneja, Shloka Desai, Ashudeep Singh, and Nadia Fawaz, "Representation Online Matters: Practical End-to-End Diversification in Search and Recommender Systems," in "Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency" FAccT '23 Association for Computing Machinery New York, NY, USA 2023, p. 1735–1746.

Singh, Ashudeep and Thorsten Joachims, "Fairness of Exposure in Rankings," in "Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining" KDD '18 Association for Computing Machinery New York, NY, USA 2018, p. 2219–2228.

Strapagiel, Lauren, "Influencers say Instagram is biased against plus-size bodies, and they may be right," May 2020.

Sun, Tianshu, Zhe Yuan, Chunxiao Li, Kaifu Zhang, and Jun Xu, "The Value of Personal Data in Internet Commerce: A High-Stakes Field Experiment on Data Regulation Policy," *Management Science*, 2023.

Sunstein, Cass R., *Republic.com*, Princeton University Press, 2001.

Wang, Jizhe, Pipei Huang, Huan Zhao, Zhibo Zhang, Binqiang Zhao, and Dik Lun Lee, "Billion-scale Commodity Embedding for E-commerce Recommendation in Alibaba," *CoRR*, 2018, *abs/1803.02349.*

Wang, Yichao, Xiangyu Zhang, Zhirong Liu, Zhenhua Dong, Xinhua Feng, Ruiming Tang, and Xiuqiang He, "Personalized Re-ranking for Improving Diversity in Live Recommender Systems," 2020.

Wang, Yuyan, Mohit Sharma, Can Xu, Sriraj Badam, Qian Sun, Lee Richardson, Lisa Chung, Ed H. Chi, and Minmin Chen, "Surrogate for Long-Term User Experience in Recommender Systems," in "Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining" KDD '22 Association for Computing Machinery New York, NY, USA 2022, p. 4100–4109.

Wilhelm, Mark, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, and Jennifer Gillenwater, "Practical Diversified Recommendations on YouTube with Determinantal Point Processes," in "Proceedings of the 27th ACM International Conference on Information and Knowledge Management" CIKM '18 Association for Computing Machinery New York, NY, USA 2018, p. 2165–2173.

Zhang, Han, Songlin Wang, Kang Zhang, Zhiling Tang, Yunjiang Jiang, Yun Xiao, Weipeng Yan, and Wenyun Yang, "Towards Personalized and Semantic Retrieval: An End-to-End Solution for E-commerce Search via Embedding Learning," *CoRR*, 2020, *abs/2006.02282.*

Zhu, Xiaojin, Andrew Goldberg, Jurgen Van Gael, and David Andrzejewski, "Improving Diversity in Ranking using Absorbing Random Walks," in "Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference" Association for Computational Linguistics Rochester, New York April 2007, pp. 97–104.