Designing Inclusive Platforms: Experimental Evidence From Pinterest

Madhav Kumar*[1], Pedro Silva[2], and Ashudeep Singh[3]

[1]Harvard University
[2]Pinterest
[3]Microsoft

March 26, 2026

Abstract

Can digital platforms broaden representation in algorithmically curated content without sacrificing user engagement? We address this question through a large-scale field experiment on Pinterest with over 630,000 users. Exposing users to more balanced skin tone representation substantially increases engagement with previously underrepresented content while maintaining overall platform activity. Users who historically sought diverse content but faced higher search frictions to find it increase overall engagement by 11% when these frictions are reduced, suggesting that the baseline algorithm underserved this user segment. Effects persist nine months after platform-wide implementation and extend to body type diversity, indicating that inclusive design is a viable and scalable platform strategy.

JEL Codes: C93, D83, L86, J15, D22, L14

Keywords: platform design, algorithmic curation, content discovery, field experiment, representation, search frictions, digital markets

*Kumar (corresponding author): Harvard Business School, Morgan Hall, 167, Soldiers Field, Boston, MA 02163, (email: mk@hbs.edu). The author received no financial assistance for this project. The author declares that he has no relevant or material financial interests that relate to the research described in this paper.; Silva: The author is a current employee of Pinterest, Inc. During the writing of this paper, he held stock that may constitute a material financial position; Singh: work done while at Pinterest. The author was an employee of Pinterest, Inc. during the writing of the paper. He has held stock that may constitute a material financial position. The views expressed here are solely those of the authors and do not represent Pinterest, Inc. All errors are our own.

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 experience described above reflects a broader pattern on digital platforms: algorithmic systems that curate content for billions of users can produce distributions of representation that are sharply skewed along dimensions such as race and gender (Luca, 2017; Lambrecht and Tucker, 2019; Evsyukova et al., 2024). These imbalances constrain content discovery for some users, limit the platform's ability to serve the full breadth of its user base, and, if left unaddressed, risk leaving measurable engagement on the table. Yet platforms have been reluctant to intervene, in part because rigorous evidence on whether inclusive design is compatible with strong platform performance has been largely absent. This paper provides such evidence. Using a large-scale field experiment, we show that rebalancing skin tone representation in algorithmically curated content substantially increases the diversity of content consumed, maintains overall engagement, and recovers demand from user segments that the baseline algorithm underserved.

On most major platforms, algorithmic recommendations determine the vast majority of content users encounter, driving 80% of 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 its impact on consumer welfare and social outcomes (Eckles, 2022). These questions are particularly salient on visual content platforms, where algorithms that prioritize content similarity can inadvertently concentrate recommendations along dimensions they do not explicitly target, including the visual characteristics of people featured in the content. When this concentration is systematic, users whose preferences are not well-served by the dominant distribution face higher search frictions, potentially limiting both their engagement and the platform's growth. Beyond these economic consequences, persistent concentration in visual representation can shape societal perceptions of beauty, fashion, and belonging, amplifying the marginalization of underrepresented groups (Kay et al., 2015; Lam et al., 2018).

We address this challenge in partnership with Pinterest, one of the largest online platforms. Pinterest's content primarily consists of images and videos, with a significant portion featuring people. Historically, surfaced content has been heavily skewed toward items featuring people with "lighter" skin tones (Figure IV). Pinterest redesigned its core content curation algorithm to strive for a more balanced representation across the skin tone spectrum (Figure VI) reducing the concentration of recommended content along skin tone dimensions. Our study is among the first

to provide causal evidence on how such design changes affect user experience and platform performance.

Using a large-scale field experiment with over 630,000 users across six countries, we find that a more balanced content distribution significantly increases engagement with previously underrepresented content while maintaining overall platform activity. Notably, users who historically sought diverse content show an 11% increase in overall engagement when search frictions for content discovery are reduced, revealing unmet demand. Multiple features of the data, including stable overall engagement, increased save rates beyond what exposure alone would predict, and no rise in user dissatisfaction signals, indicate that these patterns reflect active content selection rather than a mechanical consequence of changed exposure.

These findings advance the literature in three ways. First, we uncover meaningful heterogeneity in consumer responses that reveals how algorithmic design shapes content access, and where it falls short. Users historically underserved by the platform's existing curation system, who sought diverse content but faced higher search frictions to find it, exhibit an 11% increase in overall engagement when these frictions are reduced. This finding suggests that standard platform design, by concentrating content along dimensions it does not explicitly target, can create unnecessary search frictions for identifiable user segments, leaving measurable engagement on the table. For platform managers, this points to a concrete diagnostic: auditing recommendation outputs for systematic concentration may reveal underserved demand that simple design changes can recover. Second, we show that the perceived tradeoff between representation and platform performance is not inevitable (Berk et al., 2017; Lepri et al., 2018; Williams et al., 2018; Rhue and Clark, 2020; Kasy and Abebe, 2021). Thoughtfully designed inclusion initiatives can substantially enhance representation while maintaining aggregate engagement, and critically, without reducing engagement for any identifiable user segment. Third, we provide evidence on the durability and scalability of inclusive design as a platform strategy. Effects persist nine months after platform-wide implementation, though the absence of a concurrent control group limits causal interpretation of the post-launch data, and extend successfully to body type diversity, indicating that platforms can address multiple representation dimensions simultaneously. Together, these findings offer a practical framework for implementing and measuring the impact of inclusive design in the context of online platforms.

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 online environments that promote diversity and mitigate algorithmic pigeonholing. Our research offers actionable insights and guidelines for managers aiming to achieve inclusive design at scale.

Related Work

Our work contributes to multiple research streams examining how platform design choices can create both economic value and social impact. Digital platforms create value by facilitating interactions between users and content providers, with matching algorithms determining which content reaches which users and thereby directly shaping market outcomes (Evans and Schmalensee, 2016; Banerjee et al., 2026). A growing body of work in information systems and marketing examines how specific design features mediate this matching process, including recommendation engines (Fleder and Hosanagar, 2009; Hosanagar et al., 2014), personalization policies (Holtz et al., 2020; Sun et al., 2023), search algorithms (Ghose et al., 2014; Lei et al., 2023), reputation systems (Luca, 2017; Fradkin and Holtz, 2023), and information disclosure mechanisms (Sahni and Nair, 2019; Fong et al., 2024; Bairathi et al., 2025). While this literature has advanced our understanding of how platform design affects efficiency, it has devoted less attention to how the same design choices shape the distribution of content across user-relevant dimensions such as representation. Our paper extends this work by showing that a platform's core content suggestion algorithm, redesigned to balance content representation, can reduce search frictions for underserved users without compromising overall engagement.

Conceptually, our work contributes to the expanding literature documenting disparate outcomes across user segments on digital platforms. Edelman et al. (2017) and Luca et al. (2026) document racial discrimination on Airbnb, and Evsyukova et al. (2024) show similar biases against African American users on LinkedIn. Lambrecht and Tucker (2019) find that STEM career ads are shown disproportionately less to women due to higher costs of reaching female users. Bohren et al. (2019) provide evidence of gender-based discrimination on a popular Q&A forum. Relatedly, research has shown that algorithmic systems can systematically underrepresent certain groups in search results and image recommendations (Kay et al., 2015; Lam et al., 2018), shaping users' perceptions of who belongs in particular professional and aesthetic categories and entrenching these imbalances over time. A smaller set of recent papers has begun to shift from diagnosis to intervention: Babar et al. (2023) and Aneja et al. (2025) show that highlighting minority business ownership on a restaurant platform increases customer engagement and firm performance, and Athey et al. (2022) utilize AI-generated images to identify *stylistic* features that boost micro-lending transactions while reducing disparities. Our paper extends this interventionist line of work to one of the largest content platforms in the world, providing causal evidence that redesigning a core content curation algorithm to balance representation can ameliorate these imbalances without reducing engagement.

Understanding how inclusive design shapes engagement requires connecting to research on how algorithmic curation influences consumption patterns and how consumers value variety. When platforms optimize content selection for similarity, they can create self-reinforcing cycles that narrow the range of content users encounter (Lee and Hosanagar, 2019; Chaney et al., 2018), confining users to progressively narrower content silos and potentially suppressing the variety that users would otherwise prefer. Our finding that more balanced content distribution leads to

more balanced consumption without reducing overall engagement speaks directly to the literature on variety-seeking behavior (Kahn and Louie, 1990; Simonson, 1990; Kahn, 1995) and to research showing that expanded product variety can generate substantial consumer surplus in digital markets (Brynjolfsson et al., 2003, 2006; Aguiar and Waldfogel, 2018). Recent platform-level evidence reinforces this connection: diverse consumption patterns are associated with improved long-term engagement on music streaming platforms (Anderson et al., 2020), and enhancing content diversity increases discovery among active users without reducing overall engagement (Chen et al., 2023). By revealing substantial engagement increases among users who historically sought diverse content despite structural barriers, our results suggest that narrowly concentrated content curation may create unnecessary search frictions for certain market segments, constraining content discovery and engagement for these users.

Taken together, our findings contribute to research on platform design and consumer behavior, algorithmic curation and content diversity, and the economics of digital marketplaces. They suggest that inclusive design need not come at the cost of platform performance but may instead unlock engagement that narrowly concentrated content curation leaves unrealized.

2 Empirical Context and Representation Challenges

With an annual revenue of over $4 Billion[1], Pinterest is a global online platform used for content discovery, visual inspiration, and product exploration. It is a visual platform where users search for and save images and videos – called Pins – related to their interests. To keep the discussion general, we refer to Pins as *items* or *content* in the paper. Unlike social networking platforms focused on interpersonal connections, Pinterest primarily functions as a tool for visual inspiration and product discovery. Users create thematic collections ("boards") of saved content for future reference, spanning categories including fashion, beauty, travel, home decor, recipes, and many more. Our study focuses on fashion-related discovery, one of Pinterest's largest content categories. Figure I (see also Figure A1 in the Online Appendix) shows examples of fashion items served to a sample user. Pinterest's business model relies primarily on advertising revenue, with the platform connecting users with content that matches their interests. Downstream commercial activity, such as clicks to retailer websites and product purchases, drives advertiser demand and represents a key component of the platform's value proposition.

The platform serves over 600 million monthly active users globally[2] who discover content on the platform through multiple surfaces, such as a personalized home feed, search, and related items. 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

[1] https://s204.q4cdn.com/369458543/files/doc_earnings/2025/q4/earnings-result/Q425-PressRelease.pdf
[2] https://investor.pinterestinc.com/financial-results/quarterly-results/default.aspx

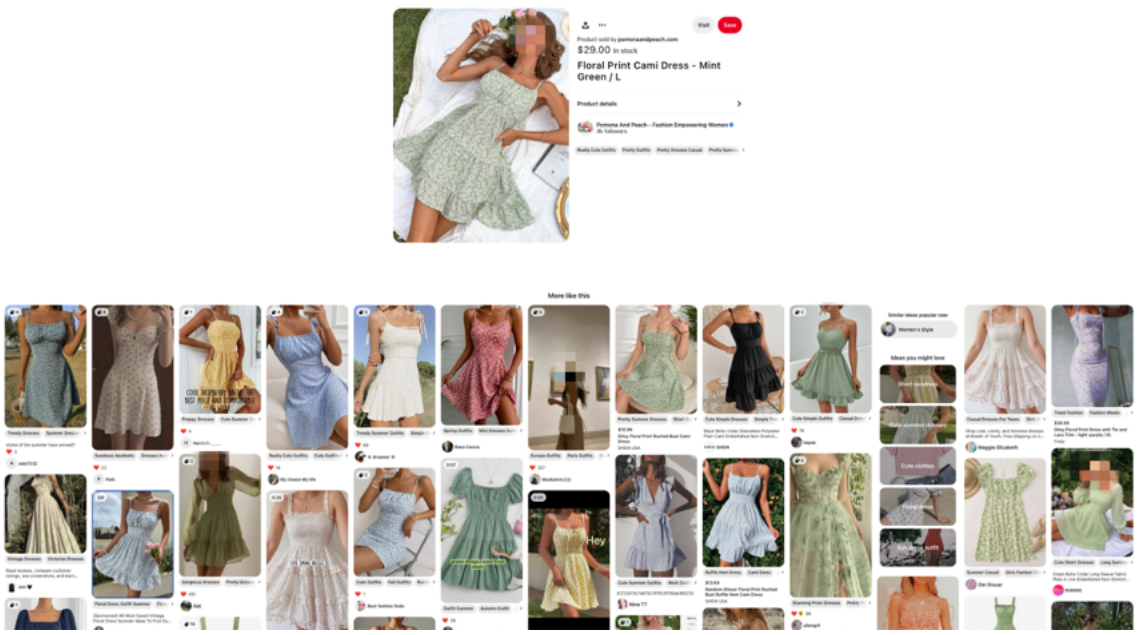


Figure I: Recommended Content on the Related Pins Surface After Clicking on a Focal Item

on Amazon product pages, and the playlist sidebar shown to a user while watching a video on YouTube. Figure I illustrates this interface, showing content presented to a user after selecting an initial item (shown at the top). The interface displays recommended items in a scrollable grid, with users typically viewing a subset of the full recommendation set before navigating away. Our analysis focuses on the top-20 recommended items, which collectively account for a substantially large majority of engagement on this surface.

On the content side, Pinterest has over 12.5 billion items in its corpus. Each of these items that includes people has an underlying skin tone signal. This signal is generated by a proprietary deep learning model developed internally at Pinterest. It is designed to handle diverse conditions, including varied lighting, poses, and occlusions, and has been iteratively refined using feedback from multiple internal evaluators (Fawaz et al., 2020). The signal classifies an item into one of four buckets based on the skin tone range of the people depicted in the image, spanning the spectrum from lighter to darker skin tones. The four categories, which we label 'Lightest', 'Second Lightest', 'Second Darkest', and 'Darkest', are derived from the model's continuous skin tone estimates and represent ordered ranges along this spectrum rather than discrete demographic categories. 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 II shows examples of items from each skin tone bucket.



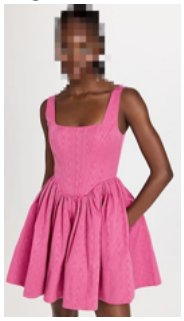
(a) Lightest Skin Tone Bucket



(b) Second Lightest Skin Tone Bucket



(c) Second Darkest Skin Tone Bucket



(d) Darkest Skin Tone Bucket

Figure II: Example Pins from Each Skin Tone Bucket

2.1 Engagement on the Platform

Most content-based digital platforms use engagement as the primary metric to evaluate design changes and optimize their algorithms (Covington et al., 2016; Kohavi et al., 2020; Holtz et al., 2020; Chen et al., 2023). It captures how users interact with the content surfaced to them and is closely tied to the commercial outcomes, such as retention and monetization, that platforms ultimately care about. On Pinterest, the core engagement measure is saving an item, which represents a deliberate user action indicating genuine interest and intent for future reference. Unlike passive engagement metrics such as impressions[3] or clicks, saving requires users to actively select content for their personal collections. This action serves as Pinterest's core engagement metric and predicts downstream behaviors, including return visits and commercial activity. Importantly, saving is considered a positive endorsement signal; users who dislike content can hide it (which we examine separately) rather than saving it.

2.2 Need for Inclusive Platform Design

Visual discovery platforms face unique representation challenges. When users search for fashion inspiration, the images they encounter shape their perception of which styles, aesthetics, and products are "for them." In categories like fashion and beauty, seeing models of similar skin tone or body type helps users visualize how items might look on themselves, which is particularly rel-

[3] An impression occurs when an item is displayed to a user.

evant for purchase decisions. However, when entire aesthetic categories (e.g., minimalist fashion, bohemian style) are predominantly represented by models of one "type" (e.g., skin tone or body type), users from other groups may feel excluded from those communities of interest (as noted by the user in the opening quote of this paper).

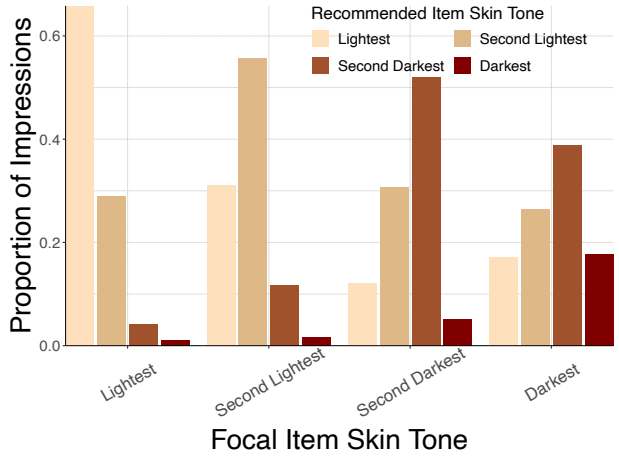
To make this idea more concrete, consider Pinterest's Related Pins surface, a discovery interface that promotes related content, and the one that we focus on in this paper. When a user selects a focal item of interest, the platform shows visually similar content (Figure I). While the similarity between the focal and recommended content items is a basic criterion for showing relevant content, the definition of similarity can often be myopic and may overlook opportunities for relevant diversification. For instance, Pinterest's baseline recommendation algorithm incorporates visual similarity and engagement signals to rank candidate items.[4] As with most industrial recommendation systems, the precise weights and features are proprietary. Importantly, skin tone is not an explicit feature in the baseline algorithm. The concentration of recommendations by skin tone is an emergent property of optimizing for visual similarity, i.e., while prioritizing features such as color, pattern, and composition, the system inadvertently clusters content by the skin tone of people featured in it. This creates a systematic pattern where interactions with content featuring people of one skin tone predominantly lead to recommendations of similar skin tones.

Figure IIIa translates this observation to a data-backed argument. The figure shows the proportion of impressions of recommended items by skin tone, conditional on the skin tone of the person in the focal item. For example, the left-most stack shows the proportion of recommended 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 reveal systematic concentration, creating frictions that hinder users from finding diverse content aligned with their interests.

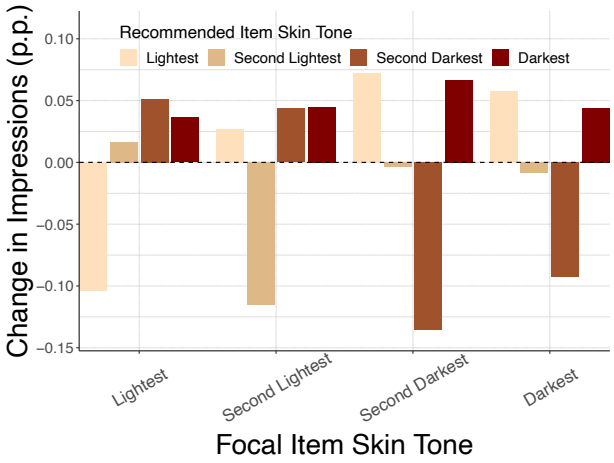
The resulting aggregate distribution across the platform is shown in Figure IV. Content featuring lighter skin tones heavily dominates the surface: the Lightest category alone accounts for a larger share of impressions than the Second Darkest and Darkest categories combined. While we mask the y-axis to preserve confidential platform data, the qualitative pattern is stark. The pre-intervention distribution is substantially skewed toward lighter skin tones, an imbalance that may not reflect the underlying preferences of the user base.

User search behavior provides direct evidence that these algorithmic patterns create meaningful frictions. Before Pinterest introduced optional skin tone filters, the majority of top search queries for skin-related content explicitly referenced skin tone, suggesting users were actively working around algorithmic limitations to find representative content (Melendez, 2020). Those

[4]The Related Pins surface focuses on displaying items that are similar to the focal item selected, and hence, user personalization is limited. More generally, the level of personalization is different across different surfaces on the platform and depends on the purpose the surface is designed to serve.



(a) Proportion of Impressions by Skin Tone Conditional on the Focal Item (Pre-Diversification Distribution)



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

Figure III: Distribution of Skin Tones in Recommended Content Conditional on the Focal Item Skin Tone

filters are user-initiated refinement tools for text-based search queries, distinct from the algorithmic intervention we study, but the behavior they were designed to address speaks directly to the demand for more representative content discovery. If this demand goes unmet, users may not be able to find content that resonates with their interests, increasing the risk of disengagement and reducing the platform's ability to retain and grow its user base. As discussed in Strapagiel (2020) and Asare (2021), and echoed by the user quote in the introduction, this experience gap can fundamentally shape how users perceive the platform's value and relevance to their needs.

When aggregated across millions of users on the platform, these representation gaps also have broader implications for market efficiency and platform 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 limits the platform's ability to learn from and adapt to the full range of consumer preferences (Chaney et al., 2018; Jiang et al., 2019). In the aggregate, narrow representation limits both consumer activity and market reach.

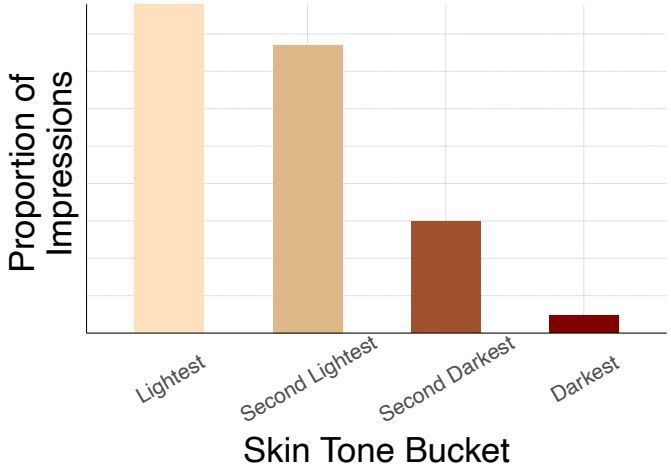


Figure IV: Pre-treatment Distribution of Impressions by Skin Tone
Note: Y-axis labels have been masked to preserve confidential information

While personalization might seem like a natural solution to these representation challenges, it is insufficient on its own for several reasons. First, personalizing based on skin tone would require inferring users' racial identity from their engagement history, an approach that is both imprecise and ethically fraught. Second, even setting aside identity inference, engagement-based personalization can only optimize over preferences that users have had the opportunity to reveal. If the baseline algorithm underexposes users to diverse content, their engagement history reflects a constrained choice set rather than their unconstrained preferences. Personalizing on that history would perpetuate the very frictions it aims to resolve, a feedback loop that prior work has documented in other recommendation settings (Acemoglu et al., 2023; Mansoury et al., 2020; Chaney et al., 2018). Third, personalization provides no solution for new users who lack engagement history, leaving them subject to the same skewed default distribution. Diversification offers a different approach: rather than tailoring content to inferred identities or constrained histories, it ensures baseline representation across categories for all users, expanding the range of content users can discover while avoiding the feedback loops that narrow content curation systems over time.

Research in academia and industry has recognized this challenge and has made substantial progress over the past decade in diversifying recommendation algorithms 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 of 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 in two ways. First, we introduce skin tone, the visual representation of

people depicted in content, as a dimension over which platform content can be effectively diversified, a salient attribute in visual discovery contexts that existing diversification approaches have overlooked. Second, whereas prior work, including the technical implementation underlying our intervention (Silva et al., 2023), has focused on how such diversification can be achieved, our focus is on the economic and behavioral consequences of deploying it at scale: how inclusive design shapes consumer engagement, content discovery, and the distribution of platform outcomes across user segments. Whether platforms can diversify along this dimension without sacrificing engagement is an open empirical question, and one with direct economic stakes, since the case for inclusive design weakens if it reduces content relevance or user satisfaction. Our experiment provides causal evidence on this tradeoff.

3 Experiment Design & Data

We examine the impact of inclusive platform design on consumer engagement and platform outcomes using a field experiment on Pinterest's Related Pins surface. We focus on this surface for two reasons. First, Related Pins is the primary content discovery interface on Pinterest, generating more impressions and engagement than any other surface. Second, the Related Pins surface is where algorithmic curation most directly shapes user experience: unlike search, where users exert substantial control over what they see through query formulation, the content suggested here is predominantly algorithm-driven, making it both more susceptible to representation imbalances and more amenable to algorithmic correction. We note that the initial search surface is unaffected by our intervention; to the extent that search results also exhibit skin tone concentration, our estimates may understate the potential gains from platform-wide inclusive design. To be precise, users in the treatment group experienced the diversified algorithm only on the Related Pins surface. Their home feed, search results, and all other platform surfaces remained identical to those of control users throughout the experiment.

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 items based on the current content system, i.e., the status quo. The status quo content tends to be more concentrated in terms of skin tone ranges displayed (Figure IIIa). In the treatment condition, users are shown items from the new recommendation system that surfaces more diverse content in terms of the visual skin tone of the underlying items, leading to a more balanced representation across skin tones (Figure IIIb).

This raises the question of what "more balanced representation" means in this context. There is no established theoretical benchmark for the "right" level of skin tone representation on a platform. Reasonable targets could be based on platform user demographics, national population shares, content supply distributions, or equal representation across categories, each with different normative justifications and each yielding a different target distribution. Rather than optimizing for a specific distributional target, the platform adopted a minimum visibility criterion: at least

one item from each of the four skin tone categories (see Figure II) among every ten recommended items. This ensures that all skin tone categories appear in a user's recommendations without imposing strict proportional parity across them. Figure V presents an example of a focal item and the corresponding items shown in the control and treatment conditions. Given the same focal item, the control condition surfaces recommended items that are visually similar but concentrated in skin tone. The treatment condition draws from the same candidate pool but reorders the items, producing a set of recommendations that retains visual relevance to the focal item while representing a wider range of skin tones. The lighter skin tone content continues to comprise the majority of the treatment items; the intervention narrows the imbalance, but does not eliminate it.

We emphasize that this is a deliberately modest target, akin to a representational floor rather than a distributional optimum. This modesty is a feature of our research design rather than a limitation. Our goal is not to identify the welfare-maximizing level of skin tone representation, which would require both a normative framework and richer preference data than we possess. Instead, we test whether a conservative increase in representation is compatible with maintaining engagement. If even this modest shift produces meaningful diversity gains without engagement costs, it establishes a lower bound on the scope for inclusive design: platforms can move at least this far toward balanced representation without adverse consequences. The question of how much further they can or should go remains open and is an important direction for future work.

3.1 Treatment Implementation

Pinterest's content curation system is operationalized using a two-stage recommendation architecture (Figure A2 in Online Appendix A.2 provides a schematic) common in large-scale industrial applications (Covington et al., 2016; Wang et al., 2018; Meta, 2019; Huang et al., 2020; Zhang et al., 2020). In the first stage, a candidate generation model retrieves a set of items relevant to the user's focal item based on visual similarity, content features, and engagement signals. This retrieval stage narrows the full item corpus to a set of several hundred candidate items for each focal item. This candidate set is identical for treatment and control users. The two conditions diverge only in how items from this common set are ranked and displayed in the second stage.

In the second stage, a ranking model scores and orders these candidates for display. Our intervention modifies only the second stage:[5] holding the candidate pool fixed, the platform re-ranks candidates to achieve more balanced skin tone representation. The re-ranking operates probabilistically rather than imposing a hard quota, increasing the likelihood that all four skin tone categories appear in the set of suggested items while preserving the relevance-based ranking signal. This distinguishes the approach from a simple slot-based rule, which would enforce

[5]One might ask why the intervention targets the ranking stage rather than the retrieval stage. The retrieval stage operates over billions of items using fast approximate nearest-neighbor search in high-dimensional embedding spaces, where imposing distributional constraints on skin tone composition is computationally infeasible at the latencies required for real-time serving. The ranking stage, which operates over the much smaller candidate set of several hundred items, is better suited to incorporating diversity objectives without compromising response times. This staging choice is standard in large-scale industrial recommendation systems (Covington et al., 2016).

representation mechanically at the potential cost of surfacing less relevant content. The minimum representation threshold described above characterizes the approximate outcome of this reranking rather than a binding constraint. Because the re-ranking operates over the available candidate pool, the degree of adjustment varies across queries depending on the skin tone composition of the candidate set for a given focal item. Crucially, the treatment does not introduce new items or alter the relevance criteria; it only adjusts the display order of items from the existing candidate pool. Silva et al. (2023) detail the technical implementation.

3.2 Data

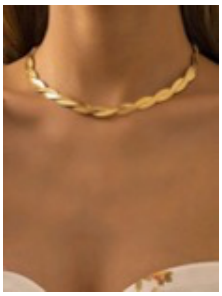
Our experimental sample includes randomly selected users from six English-speaking countries (United States, Canada, Great Britain, Ireland, Australia, and New Zealand) who made at least one fashion-related query on the Related Pins surface during the experiment period. Here, a query is defined as a user clicking on a focal item that generates a request for related content. The platform's algorithm responds with an ordered list of suggested items. Randomization is done at the user level, ensuring that each user experiences consistent recommendations throughout the experiment period. The final sample contains 319,235 users in control and 319,448 users in treatment for a period of four weeks. These numbers represent an undisclosed fraction of randomly selected users from the six countries mentioned above. About 77% of the users report their gender to be female, and ∼79% of the users are from the US.

Users are enrolled upon their first triggering query and randomly assigned to treatment or control with equal probability. Once enrolled, users remained in their assigned condition for the entire duration of the experiment. Our main analyses include all enrolled users regardless of subsequent activity; results are robust to restricting the sample to users' first query only (Table F8). Table C in the Online Appendix presents randomization checks on pre-treatment characteristics, including historical engagement, consumption diversity, and demographics. We find no evidence of imbalance across treatment arms.

For each user, we observe their experimental assignment and their activity on the platform. Further, for each query request generated by the user, we observe the focal item for which the content suggestions are generated, the items recommended 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.

3.3 Inclusivity Metrics

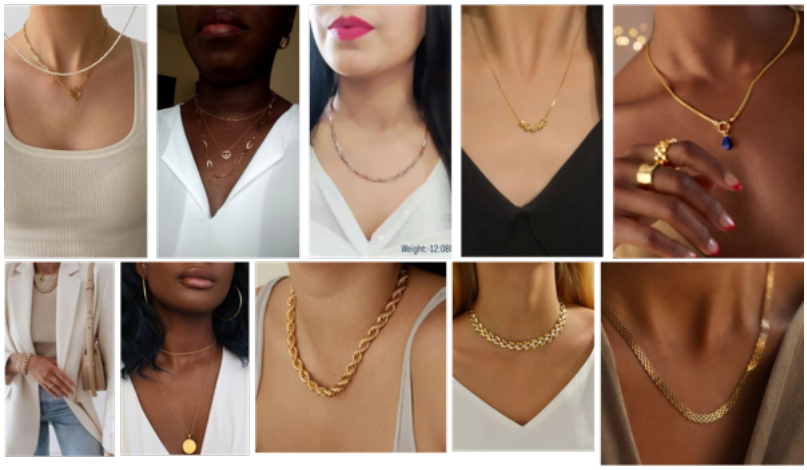
The new content curation algorithm diversifies the suggested items for each fashion-related query, reducing frictions for users seeking diverse content by more efficiently allocating visibility across the spectrum of available content options. This results in users' content exposure becoming visually more inclusive and having a more even distribution across skin tone ranges. We quantify the



(a) Focal Item



(b) Control



(c) Treatment

Figure V: Recommended Items: Treatment vs. Control

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 (recommendation system) 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 VI. We investigate the impact of this provision of more visually inclusive content on consumer engagement. In Appendix B, we show the shift in distribution between treatment and control using other measures of diversity.

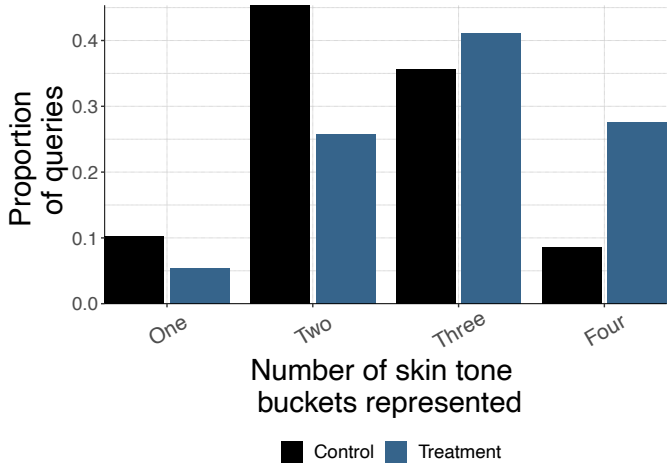


Figure VI: Proportion of queries with different numbers of skin tone buckets represented in the top-20 results

3.4 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.

3.4.1 Overall Engagement

We measure overall engagement using the number of items saved from the recommended content set by each user. On Pinterest, saving (called "Repinning" on the platform) is the action of adding an item to one's personal collection (analogous to bookmarking a product on an e-commerce site

or adding a song to a playlist) for future reference. This requires deliberate user action: clicking a save button and selecting a destination board, making it a strong signal of genuine interest rather than passive consumption. Unlike lighter engagement metrics such as impressions or clicks, which can reflect incidental attention or algorithmic placement, saving represents an intentional decision to retain content for future use. Saves are also the platform's primary engagement metric for evaluating product changes, suggesting the measure captures behavior that the platform's own analyses have found to be economically meaningful. We use saving and engagement interchangeably throughout the paper.

Like most measures of online user activity, items saved has a long-tail (Brynjolfsson et al., 2006). Consequently, we consider the top-20 recommended 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. In Online Appendix E, we check the robustness of our results with other measures of engagement.

3.4.2 Consumption Diversity

Since the intervention changes the skin tone composition of content users are exposed to, it is important to assess whether it also shifts the composition of content they choose to engage with. To this end, 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.

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 increased the exposure (relative to the control group) of deeper skin tone items, 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:

$$S^i = -\sum_{b \in \mathcal{B}} s_{bi} \log_2(s_{bi}) \tag{2}$$

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.

Interpreted jointly with overall engagement, these outcomes speak to a central platform-design question: can the platform expand representation in what users *choose* without degrading engagement? A shift toward more balanced consumption, accompanied by stable (or higher) total engagement, is consistent with the intervention relaxing a relevance constraint, i.e., surfacing content users value but were not being shown, rather than pushing users into unwanted content. The heterogeneity analysis in Section 6 provides a further test of this interpretation by examining whether users with stronger revealed preferences for diverse content respond differently to the intervention.

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 recommended 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.

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,

[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.

which allow us to directly assess whether inclusive content distribution affects overall engagement and the diversity of content consumed. 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 presented in Online Appendix D-K, with specific sections noted in the text.

4.1 Average Treatment Effect

We estimate the treatment effect using Model 3 on different measures of engagement described in Section 3.2. The results are shown in Table I. 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 an approximate 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. For ease of interpretation, we normalize the raw outcome in the control group to 1 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 $\sim$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 daily logins (Online Appendix, Table E4). The results are also consistent if we estimate Model 3 using OLS, as shown in Table E3.

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 $\sim$15%. The proportion of users engaging with content from all four skin tone buckets also increases substantially, with a 71% ($\exp(0.5363) - 1 = .71$) increase in the odds. Given the low base rate of this outcome in the control group, this closely approximates a 70% increase in the proportion itself, as confirmed by the linear probability model in Table E3.

[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 paper recommends a Poisson/Quasi-Poisson for expressing ATE as a percentage.

Taken together, these results indicate that platforms can meaningfully increase consumption diversity along a dimension users demonstrably care about, i.e., skin tone representation, without reducing overall engagement. The increase in diverse consumption is consistent with users engaging with content that was previously harder to discover due to algorithmic concentration.

Table I: 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 columns 1-3 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

4.2 Treatment Effects by User Characteristics

We examine heterogeneity in treatment effects across key user characteristics: self-reported gender, geography, and tenure with the platform. It allows us to assess whether the aggregate null effect on overall engagement masks offsetting effects across user subgroups, e.g., gains for some users at the expense of others. An intervention that maintains aggregate engagement while making certain groups worse off would raise distributional concerns distinct from the average effects we document. Additionally, examining heterogeneity by observable demographics helps characterize the generality of our findings and identify whether particular user segments respond differently to increased representation.

Table G10 in the Online Appendix shows estimated coefficients for treatment-by-demographic interactions. For overall engagement and engagement with deeper skin tone content, we find no significant heterogeneity by gender or country (Columns 1-4). For Shannon entropy and engagement with all four skin tone buckets, we observe significant positive interactions with female gender (Columns 5 and 7), indicating that women show larger increases in consumption diversity under treatment. This may reflect women's greater representation in fashion categories where skin tone diversity is particularly salient, or baseline gender differences in variety-seeking behavior. Importantly, treatment effects on overall engagement remain stable across gender, indicating the intervention does not differentially impact any demographic group.

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 $\sim$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.

5 Consumer Preferences vs. Mechanical Effects

Admittedly, one may worry that the consumption diversity effects shown in Table I are largely mechanical. For example, content that receives increased visibility naturally receives more engagement, regardless of whether users value it. Before addressing this concern empirically, it is useful to clarify what the mechanical framing implies. Any recommendation system operates through exposure: exposure is a precondition for engagement, and users can only engage with content they are shown. This is the fundamental mechanism through which platforms create value. Consequently, when a system surfaces content that matches user preferences, and users engage with it, we consider it effective content matching and not a mechanical effect. In this sense, if the intervention exposes users to diverse content and they choose to save it while maintaining their overall level of engagement, this is consistent with the system surfacing content that users find relevant but previously had difficulty discovering.

A distinct concern, however, is that users engage with the newly surfaced content not because they value it, but simply because it occupies positions in the recommendation list that would otherwise have shown content they preferred more. Under this conjecture, we would expect overall engagement to decline, dissatisfaction signals to increase, or engagement with diverse content to merely track its increased exposure share without exceeding it. We address this concern along two dimensions. First, we examine whether users actively value the diverse content they are shown, drawing on multiple indicators of preference alignment beyond raw engagement counts. Second, we consider whether engagement effects operate through skin tone representation specifically or through correlated item characteristics that may vary with skin tone. We address each in turn.

5.1 Do Users Value Diverse Content?

It is worth emphasizing that exposure does not automatically translate to engagement. User agency is critical in separating passive consumption from active engagement. Hence, 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 I, the overall engagement rate is stable, suggesting that match quality and preference alignment are 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 F9), indicating higher user receptivity to this content beyond mere visibility effects.

Third, we address the concern that the treatment effect might be 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 F4 in the Online Appendix echoes the result from Table I. Across the top-20 positions, overall engagement remains stable, and engagement with deeper skin tone content goes up significantly. Since we are comparing engagement within the same position, this analysis mitigates concerns that the results are driven by positional advantages.

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 II 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, shows that the new algorithm is catering to consumer preferences while actively promoting content skin tone diversity.

Table II: Average Treatment Effect on User Dissatisfaction Measured by Items Hidden

Dependent Variable:	Number of Items Hidden
Model:	(1)
	Quasi-Poisson
Variables	
Treatment	−0.0316
	(0.0243)
Fit statistics	
Observations	638,683
Log-Likelihood	-275,543,399.3

*This table compares the number of items hidden by users in treatment vs. control. The platform interface allows users to hide content items with a single click, i.e., give a direct signal back to the system that the user would not like to see content like this. This measure allows us to gauge whether users are being exposed to irrelevant content in the pursuit of diversity. Heteroskedasticity-robust standard errors in parentheses. Signif. Codes: **: 0.05, *: 0.1*

In addition to these immediate indicators, we also assess longer-term persistence effects using data collected nine months after platform-wide implementation of the new diversified system. While the post-launch period lacks a concurrent control group, the sustained elevation of consumption diversity (see Section 7.1) is inconsistent with the notion that the effects were driven by short-lived novelty responses.

5.2 Skin Tone Intervention vs. Correlated Item Characteristics

A related concern is whether the effects we document reflect preferences for skin tone representation specifically or preferences for the item characteristics that may correlate with skin tone. If content featuring darker-skinned models systematically differs in style, color palette, or fashion category, then increased engagement with diverse skin tones could partly reflect increased exposure to varied items rather than varied representation per se.

While we cannot fully decompose these channels with our data, several features of our design suggest that skin tone representation plays a meaningful independent role. To begin, it is helpful to note how the experimental design and algorithmic structure inherently constrain item variation. Users reach the Related Pins surface by clicking on a focal item of their choice, so all suggested content, in both treatment and control, must be relevant to that user-selected item. Within this already-constrained candidate pool, the diversification algorithm operates purely by reranking: both treatment and control users draw recommendations from the same set of candidate items, with the treatment adjusting only the order in which the items are surfaced from this pool. The algorithm does not introduce new items that would not have been candidates under the baseline system. Thus, the scope for systematic differences in item characteristics between treatment and control is bounded twice over, first by relevance to the focal item, and second by the common candidate pool. However, this does not eliminate the possibility that items promoted by the reranking differ from demoted items on characteristics correlated with skin tone; we provide two additional tests to assess this possibility.

First, we examine the consistency of treatment effects across content sub-categories. If item variety rather than skin tone representation were the primary driver, the treatment effect on deeper skin tone engagement to be inconsistent across categories: positive in categories where darker-skinned models happen to be paired with more desirable items, attenuated or negative in categories where no such pairing exists. Figure H5b shows treatment effects on engagement with deeper skin tone content across the twenty largest sub-categories. The effect is consistently positive across most of the sub-categories, including those with very different item compositions, e.g., women's style, dresses, occasion outfits, jewelry, and hair. This pattern is more consistent with a role for skin tone representation than with category-specific confounds driving the results.

Second, we test this more directly by exploiting the structure of our experimental design. For any given focal item, the candidate pool of suggested items is identical across treatment and control, and only the reranking differs. We estimate treatment effects on engagement using query-level data, restricting the sample to focal items observed in both treatment and control conditions and conditioning on the focal item. This within-focal-item comparison holds constant the candidate pool and all query-level characteristics, substantially narrowing the scope for confounding item variation. Table III reports the results. Consistent with our main findings, treated users save significantly more deeper skin tone content even when the candidate pool is held constant, providing further evidence that the results are not driven by differences in the types of queries or candidate pools across conditions, and are consistent with users responding to the skin tone

composition of displayed content.

Table III: Treatment Effect Conditional on the Focal Item

Dependent Variables:	Total Saves	Saves of Darker ST Content
Model:	(1)	(2)
	Quasi-Poisson	Quasi-Poisson
Variables		
Treatment	0.0206	0.1188**
	(0.0126)	(0.0229)
Fixed-effects		
Focal Item	✓	✓
Fit statistics		
Observations	1,702,972	524,838
Log-Likelihood	-978,640.7	-204,090.0

*This table estimates treatment effects at the query level, restricting the sample to focal items observed in both treatment and control conditions. Column 1 regresses overall engagement (total saves) on the treatment indicator with focal-item fixed effects. Column 2 repeats the analysis for engagement with deeper skin tone content. Both columns are estimated using Quasi-Poisson, and coefficients can be interpreted as approximate percentage changes. The focal-item fixed effects hold constant the candidate pool and all query-level characteristics. Clustered (User) standard errors in parentheses. Signif. Codes: **: 0.05, *: 0.1*

Finally, and independent of the tests above, even if item variety contributes to these effects, this does not diminish the practical value of the findings. Skin tone and item characteristics are not independent in the real world: fashion items are photographed with specific models, and these pairings are not randomly assigned by content creators. Our intervention works within this structure, surfacing existing content more evenly rather than artificially separating skin tone from items. The managerially relevant question is whether diversifying on the observable dimension of skin tone harms engagement, and our results indicate it does not. The extent to which the effects are driven by skin tone representation versus correlated item attributes is an important question for future research, but it does not alter our central conclusion that platforms can increase representation without sacrificing engagement.

In summary, the evidence presented in this section is collectively inconsistent with a purely mechanical interpretation of our main results: overall engagement does not decline, save rates for deeper content increase, the effects are not driven by position, dissatisfaction signals do not increase, and the patterns hold across sub-categories and within focal-item fixed effects. However, we cannot fully rule out the possibility that correlated item characteristics contribute to the observed engagement shifts, nor can we definitively separate active preference from passive exposure-driven behavior. What the evidence does establish is that the intervention increased representation without degrading engagement, and that the increase in diverse consumption was accompanied by patterns consistent with user preference rather than pure mechanical displacement.

6 Who Benefits from Inclusive Design?

The aggregate results show that inclusive design can enhance distributional outcomes without sacrificing platform performance. 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. To this end, we analyze how users' historical engagement shapes their response to more inclusive content distribution.

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 platform design 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 design, 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.

We test for this using users' prior exposure to and engagement with deeper skin tone content. We segment users in our experiment on two dimensions – 1) diversity of content exposed to in the 3-month window prior to the experiment and 2) 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 seek diverse content despite potential barriers to discovering it.

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 consumer 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 use the top quintile as the threshold in order to isolate users with the strongest revealed preference signals while maintaining sufficient sample size for precise estimation. In Online Appendix I, we show that results are robust to examining treatment effects across all quintiles of each dimension separately (Table I13) and to replacing the discrete segmentation with continuous pre-treatment measures (Table I14), ensuring that this particular cutoff is not driving our findings.[9]

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 a *preference for deeper skin tone content*.

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

[9] 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.

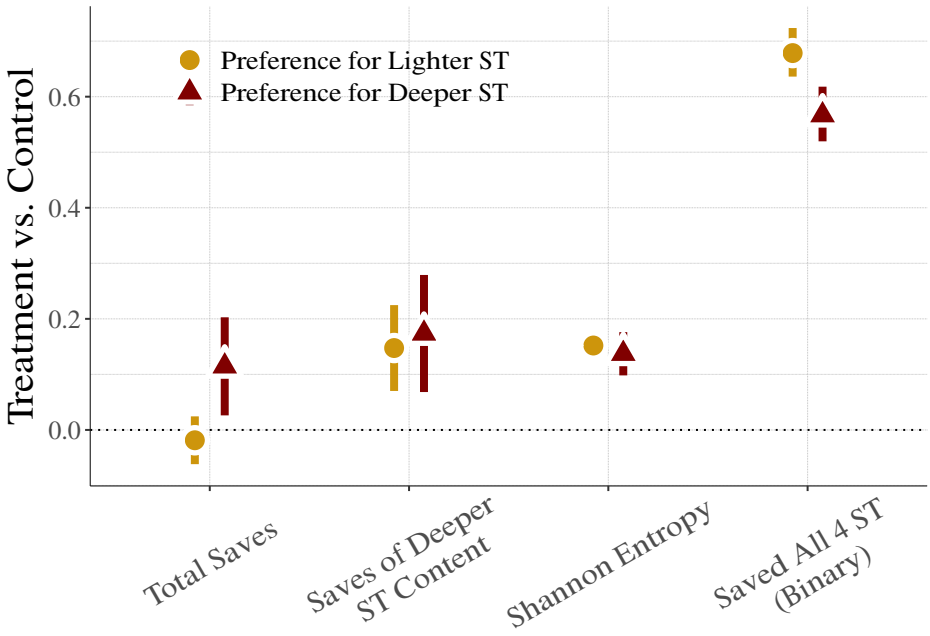


Figure VII: Heterogeneous Treatment Effects on Engagement and Consumption Diversity

The remaining users are classified as users with a *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 VII.

Decomposing the Aggregate Effect. We find that users classified as having a *preference for deeper skin tone content* see an 11% increase in overall engagement. The treatment effect on overall engagement for users with *preference for lighter skin tone content* is -0.08% 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 (as shown in Table I).

Table IV decomposes these overall effects by the skin tone of the content engaged with. For users with a preference for deeper skin tone content, the 11% overall engagement gain is primarily driven by a 17% increase in engagement with deeper skin tone content. Engagement with lighter skin tone content increases by 7% but is not statistically significant. This pattern is consistent with the unmet demand interpretation: these users increase engagement with the type of content the baseline algorithm made difficult to discover. For users with a preference for lighter skin tone content, the near-zero overall effect reflects a reallocation of engagement. A significant 15% increase in deeper skin tone engagement offset by a significant 5% decrease in lighter skin tone engagement, leaving total volume unchanged.

A notable pattern in Figure VII is that even users *without* a strong prior preference for deeper skin tone content increase their engagement with such content. A most straightforward explanation is compositional. The treatment increases exposure to diverse content for all users, and some of this increased exposure translates to engagement as users select from a changed set. Under

Table IV: Decomposition of Treatment Effects on Engagement by Skin Tone Content

	Lighter ST Engagement	Deeper ST Engagement
Users with Lighter ST Preferences	−0.049**	0.147**
	(0.023)	(0.039)
Users with Darker ST Preferences	0.071	0.173**
	(0.048)	(0.053)

*This table decomposes the overall engagement effects from Figure VII into engagement with lighter and deeper skin tone content separately for each user segment. Lighter ST Engagement captures saves of items classified in the Lightest or Second Lightest skin tone buckets. Deeper ST Engagement captures saves of items classified in the Second Darkest or Darkest skin tone buckets. Coefficients are estimated using Quasi-Poisson and can be interpreted as approximate percentage changes. Standard errors are shown in parentheses. Signif. Codes: **: 0.05, *: 0.1*

this account, the reallocation from lighter to deeper content reflects the mechanical relationship between exposure and choice, not an active preference for diversity. However, several aggregate patterns from Section 5.1, including elevated save rates beyond exposure shares (Table F9), no increase in content hiding (Table II), and stable effects across positions (Figure F4b), weigh against a purely mechanical interpretation. Importantly, the paper's core findings hold under either interpretation: the intervention increases representation without reducing overall engagement for any user segment, expanding access for previously underserved users without disrupting the experience of others. These patterns suggest that at least some of the shift reflects active content selection. A complementary possibility, consistent with the variety-seeking literature (Kahn and Louie, 1990; Simonson, 1990; Kahn, 1995), is that the baseline algorithm produced recommendations more concentrated in skin tone than even these users would prefer, and the intervention moves them closer to a preferred level of variety.

We cannot cleanly separate these channels with our data. Importantly, however, the paper's core findings hold under either interpretation. Under both accounts, the intervention increases representation without reducing overall engagement for any user segment. Users with the strongest revealed preferences for diverse content see meaningful gains in total engagement, while other users shift the composition of their engagement toward more balanced content without any reduction in volume. The intervention expands access for previously underserved users without disrupting the experience of others, a result that is informative for platform design.

7 Scaling Inclusive Design

We extend our main findings in two directions. 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 multiple diversity dimensions, demonstrating its potential to drive inclusivity across a wide range of applications. These analyses allow us to assess the durability and generalizability of our main findings, key considerations relevant both to platforms evaluating inclusive design as a long-term

strategy and to policymakers assessing whether interventions promoting representation are likely to achieve lasting effects.

7.1 Longer-term Persistence

A crucial question for our experimental findings is whether they reflect persistent changes in user behavior or transient responses to novelty. We examine this using data from ∼800,000 users collected over nine months following Pinterest's platform-wide launch of more inclusive content suggestions to all users 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 implications of inclusive design policies. We cannot interpret the results in the section as clean long-term causal effects since there is no active control group after the platform-wide launch. Nevertheless, 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 VIII. 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.[10]

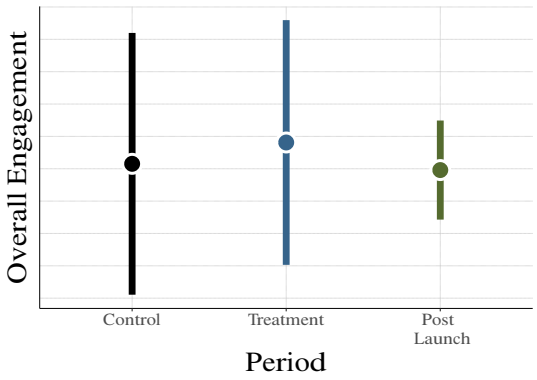
The results are suggestive of sustainable impact, though they should be interpreted with the caveat that no concurrent control group exists in the post-launch period. Overall engagement remains stable in the months following implementation, while the diversity of content consumption continues to significantly exceed pre-launch levels (Figure VIII). 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 inclusive 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-

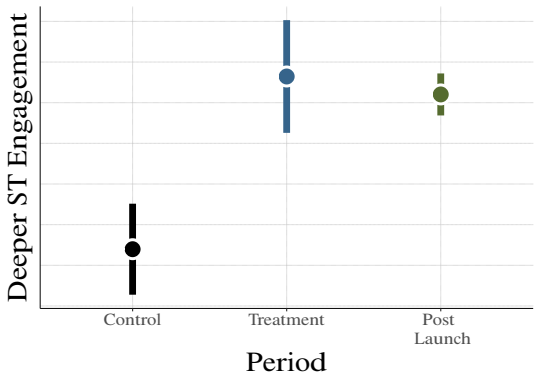
[10]We use a new random subset from the entire user base rather than tracking the same users in the experiment to ameliorate *survivorship bias*. The set of users who are active nine months after the launch is likely to be more engaged, and hence, their outcome measures will not be representative of 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.

term metrics: login frequency and overall engagement. Controlling for rich measures of historical user behavior and demographics, Tables K15 and K16 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.

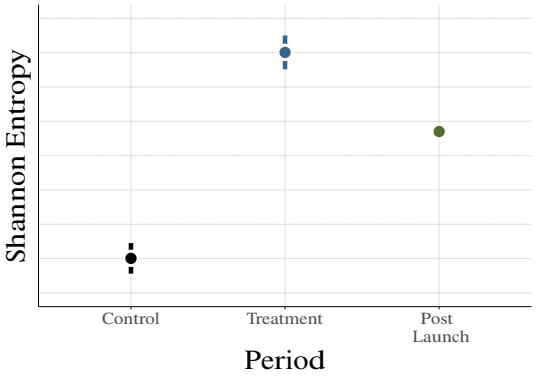
Second, and more importantly, the sustained increase in diverse content consumption without any decline in overall engagement reveals previously unmet user needs. The persistence of these effects indicates that historical platform patterns may have created artificial market frictions to content discovery 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.



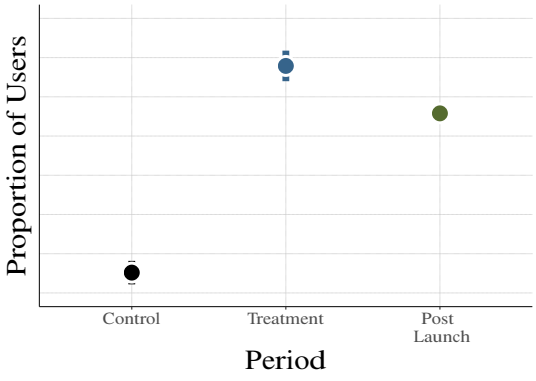
(a) Overall Engagement



(b) Engagement with Deeper ST Content



(c) Shannon Entropy Across ST Buckets



(d) Engagement with Content from all 4 ST Buckets

Figure VIII: Overall Engagement and Consumption Diversity 9 Months After the Platform-Wide Launch

Note: Y-axis labels have been masked to preserve confidential information

7.2 Multidimensional Inclusivity with Body Type Diversity

Our main experiment focuses on skin tone, but representation challenges on visual platforms extend to other dimensions such as body type, age, and physical ability. Here, we explicitly test whether inclusive design can address multiple dimensions simultaneously, or whether diversifying along one attribute comes at the cost of others. We assess this by extending inclusive platform design to another critical dimension of visual representation: body type diversity. This extension provides a rigorous generalization test of whether online 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 threefold, with particular gains in visibility for previously underrepresented Upper Middle Range and Upper Range content (200% increase in impressions).

Table V shows the impact on overall user engagement and consumption diversity. In this case, consumption diversity is defined only over the set of body type buckets. Like skin tone, body type diversity is a dimension with clear user relevance in fashion contexts: users benefit from seeing how clothing fits on bodies similar to their own. The consistency of results across both dimensions, each representing a meaningful aspect of visual self-relevance rather than an arbitrary content attribute, strengthens confidence that the findings reflect genuine user preferences rather than novelty effects or mechanical exposure increases.

These findings have broad implications for inclusive platform design. 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 consumer preferences solely based on their historical consumption. The positive response to multidimensional representation suggests that users' notion of relevance is broader than what is inferred from historical consumption alone: users respond to representation along multiple dimensions, 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

Table V: 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 columns 1-3 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

design choices can create more inclusive digital spaces while maintaining business performance. 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.

8 Robustness Checks and Limitations

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 for overall engagement and the three consumption diversity variables. The results are consistent and are shown in Table E3.

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 number of days a user logs into the platform is another key metric that most platforms carefully track. The results are shown in Table E4 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 E5.

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 F6. 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 that the new content policy reorganized 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 F4a. Again, we find consistent results – overall engagement results for each position are stable and are *not* distinguishable from zero.

5. **Robustness for heterogeneous effects.** The analysis in Section 6 segments users into quintiles based on historical exposure diversity and historical engagement with deeper skin tone content. In Online Appendix I, we show the results across all quintiles of historical engagement and exposure estimated separately. We also estimate the heterogeneous effects using the raw continuous measures instead of the discretized segmentation. The results are consistent.

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. 1) We estimate treatment effects separately for each week of the experiment and find stable effects throughout (Figure J6). 2) Since stable weekly effects could be driven by new user cohorts entering each week, with each exhibiting short-lived novelty, we re-estimate treatment effects by the number of days each user has spent in the experiment. We find no systematic attenuation over time (Figure J7).

8.1 Limitations

Three features of our study design merit discussion when interpreting the results.

Consumption vs. Welfare. Our outcomes capture engagement behavior and not user welfare directly. The intervention shifts the composition of content consumed toward greater skin tone balance without reducing total engagement, but we cannot fully disentangle how much this compositional shift reflects genuine preference satisfaction vs. accommodation of a changed choice set. The stability of overall engagement, the absence of increased content hiding (Table II), and the elevated save rate of deeper skin tone content beyond its exposure share (Table F9) are consistent with users finding the redistributed content at least as relevant as the content it displaced. But we cannot rule out more nuanced welfare effects, particularly for users whose consumption shifted without a corresponding increase in total engagement. Similarly, we do not observe downstream economic outcomes such as purchases or user satisfaction.

Scope of Intervention. Our treatment operates within a specific design: holding the candidate pool of relevant items fixed, the diversification algorithm reranks candidates to achieve more balanced skin tone representation among the set of suggested items. This design means that both treatment and control users draw from the same pool of items for any given focal item, and the treatment adjusts display order rather than introducing new content. While this constrains the scope for systematic differences in item characteristics across conditions, it does not eliminate the possibility that items featuring models of different skin tones also differ in other attributes such as style, color palette, or aesthetic category. Our analysis in Section 5 shows that item quality (measured by historical save rates) does not differ between treatment and control, and the heterogeneity results in Section 6, where segmenting users by their skin tone engagement history predicts differential responses, are more consistent with skin tone relevance than with item variety alone. Nonetheless, we cannot fully decompose how much of the observed engagement shift reflects preferences for skin tone representation versus preferences for the item characteristics correlated with it.

Content Creators. Our analysis focuses entirely on the demand side, i.e., how users respond to changes in content recommendations. We do not examine effects on content creators or the sources of content on the platform. This reflects an institutional feature of Pinterest: unlike other social platforms where individual creators produce and monetize original content, Pinterest primarily aggregates visual content from the broader web, and the platform does not maintain comprehensive data on creators' identities or demographics. Whether increased engagement with content featuring deeper skin tones translates into meaningful outcomes for specific content creators or businesses is an empirical question we cannot address with our data. This boundary does not diminish the demand-side impact: establishing that a platform can substantially broaden representation without reducing user engagement is a necessary condition for any supply-side benefits to materialize, and it directly addresses the concern, widely held among platform operators, that inclusive design comes at a cost to engagement.

9 Discussion and Implications

This paper provides large-scale causal evidence that inclusive platform design, specifically, rebalancing the representation of skin tones in algorithmically curated content, can substantially increase the diversity of content consumed without reducing overall user engagement. A modest intervention, striving for minimum visibility for each of the four skin tone categories rather than imposing strict distributional parity, increased the proportion of queries with full skin tone representation by over 200% while leaving aggregate engagement unchanged. Even under treatment, lighter skin tone content retained its majority share of impressions, meaning our results speak to the feasibility of a representational floor, not a radical redistribution.

The heterogeneity in consumer responses to more inclusive content reveals important insights about digital platform design and how algorithmic concentration shapes content access. Users

whose historical behavior revealed a preference for diverse content, but who faced higher search frictions to find it under the baseline algorithm, increased overall engagement by 11% under the new design. This suggests that narrowly concentrated content curation can leave measurable demand on the table: there exist identifiable user segments whose engagement can be recovered through relatively simple design changes. Importantly, this recovery comes without reducing engagement for other users, meaning inclusive design need not involve cross-segment tradeoffs. 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.

These findings generalize along at least two dimensions. The persistence of elevated consumption diversity nine months after platform-wide implementation suggests that the effects reflect durable shifts in content discovery rather than transient novelty, though the absence of a concurrent control group limits causal interpretation of the post-launch data. The successful extension to body type diversity shows that platforms can address multiple representation dimensions simultaneously, with compounding rather than competing benefits. We note, however, that our evidence comes from fashion-related content on a visual discovery platform. This is a setting where the visual characteristics of people depicted in content influence users' assessment of self-relevance. The extent to which these results generalize to content categories where such self-relevance is weaker, or to platforms with different content modalities, remains an open question for future work.

More broadly, our results suggest that the perceived tradeoff between representation and engagement may, in some settings, reflect narrow design choices rather than a fundamental economic constraint. Platforms that optimize purely for content similarity risk concentrating recommendations along dimensions they do not explicitly target, inadvertently creating frictions for segments of their user base. Inclusive design offers a practical path to correcting these inefficiencies, one where improving the platform's match between content and the full breadth of user preferences and advancing equity objectives are not competing goals but complementary ones.

References

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius, "A Model of Online Misinformation," *The Review of Economic Studies*, 12 2023, *91* (6), 3117–3150.

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.

Aguiar, Luis and Joel Waldfogel, "Quality Predictability and the Welfare Benefits from New Products: Evidence from the Digitization of Recorded Music," *Journal of Political Economy*, 2018, *126* (2), 492–524.

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," *American Economic Review*, February 2025, *115* (2), 660–89.

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

Athey, Susan, Dean Karlan, Emil Palikot, and Yuan, "Smiles in Profiles: Improving Fairness and Efficiency Using Estimates of User Preferences in Online Marketplaces," Working Paper 30633, National Bureau of Economic Research November 2022.

Babar, Yash, Ali Mahdavi Adeli, and Gordon Burtch, "The Effects of Online Social Identity Signals on Retailer Demand," *Management Science*, 2023, *69* (12), 7335–7346.

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

Banerjee, Shrabastee, Ishita Chakraborty, Hana Choi, Hannes Datta, Remi Daviet, Chiara Farronato, Minkyung Kim, Anja Lambrecht, Puneet Manchanda, Aniko Öry, Ananya Sen, Marshall Van Alstyne, Prasad Vana, Kenneth C. Wilbur, Xu Zhang, and Bobby Zhou, "Digital Platforms 2.0: Emerging Topics, Opportunities, and Challenges," *International Journal of Research in Marketing*, 2026.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth, "A convex framework for fair regression," *arXiv preprint arXiv:1706.02409*, 2017.

Bohren, J. Aislinn, Alex Imas, and Michael Rosenberg, "The Dynamics of Discrimination: Theory and Evidence," *American Economic Review*, October 2019, *109* (10), 3395–3436.

Brynjolfsson, Erik, Yu (Jeffrey) Hu, and Michael D. Smith, "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers," *Management Science*, 2003, *49* (11), 1580–1596.

_ , 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.

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.

Eckles, Dean, "Algorithmic transparency and assessing effects of algorithmic ranking," 2022.

Edelman, Benjamin, Michael Luca, and Dan Svirsky, "Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment," *American Economic Journal: Applied Economics*, April 2017, *9* (2), 1–22.

Evans, D.S. and R. Schmalensee, *Matchmakers: The New Economics of Multisided Platforms* G - Reference,Information and Interdisciplinary Subjects Series, Harvard Business Review Press, 2016.

Evsyukova, Yulia, Felix Rusche, and Wladislaw Mill, "LinkedOut? A Field Experiment on Discrimination in Job Network Formation*," *The Quarterly Journal of Economics*, 10 2024, *140* (1), 283–334.

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.

Fleder, Daniel and Kartik Hosanagar, "Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity," *Management Science*, 2009, *55* (5), 697–712.

Fong, Jessica, Puneet Manchanda, and Yu Song, "How effective is suggested pricing?: Experimental evidence from an E-commerce platform," *SSRN Electron. J.*, 2024.

Fradkin, Andrey and David Holtz, "Do Incentives to Review Help the Market? Evidence from a Field Experiment on Airbnb," *Marketing Science*, 2023, *42* (5), 853–865.

Ghose, Anindya, Panagiotis G. Ipeirotis, and Beibei Li, "Examining the Impact of Ranking on Consumer Behavior and Search Engine Revenue," *Management Science*, 2014, *60* (7), 1632–1654.

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

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.

Hosanagar, Kartik, Daniel Fleder, Dokyun Lee, and Andreas Buja, "Will the Global Village Fracture Into Tribes? Recommender Systems and Their Effects on Consumer Fragmentation," *Management Science*, 2014, *60* (4), 805–823.

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.

Kahn, Barbara E., "Consumer variety-seeking among goods and services: An integrative review," *Journal of Retailing and Consumer Services*, 1995, *2* (3), 139–148.

Kahn, Barbara E and Therese A Louie, "Effects of retraction of price promotions on brand choice behavior for variety-seeking and last-purchase-loyal consumers," *Journal of Marketing research*, 1990, *27* (3), 279–289.

Kasy, Maximilian and Rediet Abebe, "Fairness, Equality, and Power in Algorithmic Decision-Making," in "Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency" FAccT '21 Association for Computing Machinery New York, NY, USA 2021, p. 576–586.

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.

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.

Lee, Dokyun and Kartik Hosanagar, "How Do Recommender Systems Affect Sales Diversity? A Cross-Category Investigation via Randomized Field Experiment," *Information Systems Research*, 2019, *30* (1), 239–259.

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.

Lepri, Bruno, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck, "Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges," *Philosophy & Technology*, 2018, *31* (4), 611–627.

Luca, Michael, "Designing Online Marketplaces: Trust and Reputation Mechanisms," *Innovation Policy and the Economy*, 2017, *17*, 77–93.

_ , Elizaveta Pronkina, and Michelangelo Rossi, "The Evolution of Discrimination in Online Markets: How the Rise in Anti-Asian Bias Affected Airbnb During the Pandemic," *Marketing Science*, 2026, *45* (1), 108–122.

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.

Melendez, Steven, "To get skin tones right, Pinterest's AI went way beyond facial recognition," *Fast Company*, August 2020.

Meta, 2019.

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

Rhue, Lauren and Jessica Clark, "Automatically signaling quality? A study of the fairness-economic trade-offs in reducing bias through AI/ML on digital platforms," *A Study of the Fairness-Economic Tradeoffs in Reducing Bias through AI/ML on Digital Platforms (January 10, 2020). NYU Stern School of Business*, 2020.

Sahni, Navdeep S and Harikesh S Nair, "Does Advertising Serve as a Signal? Evidence from a Field Experiment in Mobile Search," *The Review of Economic Studies*, 10 2019, *87* (3), 1529–1564.

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.

Simonson, Itamar, "The Effect of Purchase Quantity and Timing on Variety-Seeking Behavior," *Journal of Marketing Research*, 1990, *27* (2), 150–162.

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.

Williams, Betsy Anne, Catherine F Brooks, and Yotam Shmargad, "How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications," *Journal of Information Policy*, 2018, *8*, 78–115.

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.