Inclusive Recommendations and User Engagement: Experimental Evidence from Pinterest

MADHAV KUMAR*, Massachusetts Institute of Technology, USA
PEDRO SILVA, Pinterest, USA
ASHUDEEP SINGH, Pinterest, USA
ABHAY VARMARAJA, Pinterest, USA

We study the impact of diversifying recommendations for inclusivity on one of the largest visual content discovery platforms in the world, Pinterest. Pinterest re-designed its recommendation systems to improve the representation of all skin tones in recommended content and foster a more inclusive user experience. We describe the design of the new recommendation system and present results from a field experiment in which users across six countries were randomly assigned to receive a more diverse set of recommendations based on content skin tone. We find that the overall engagement rates remain stable and engagement with previously underrepresented content increases significantly. More broadly, users diversify their consumption by engaging with content from all skin tone ranges. We shed light on the mechanism driving these results using heterogeneous treatment effect analysis. We find that engagement for users with "*preference for deeper skin tone content*" increases significantly and engagement for users with "*preference for lighter skin tone content*" remains relatively stable. Finally, we analyze post-launch data to better understand the long-term implications of diversifying recommendations. Our research provides practical insights for platform managers and policymakers to create inclusive digital environments that promote engagement while catering to diverse user preferences.

Contents

*Corresponding author

Manuscript submitted for review to the 25th ACM Conference on Economics & Computation (EC'24).

1 INTRODUCTION

Over 4 billion people across the world use social media today[1] to discover new content, entertain themselves, and connect with friends and loved ones. Despite their global reach, unequal representation, reinforcement of stereotypes, and biased algorithmic outcomes are pervasive throughout online platforms [Kay et al., 2015, Lam et al., 2018, Lambrecht and Tucker, 2019]. Consequently, commensurate with the growth of these digital spaces have been calls to make them more representative of the users they serve [Albergotti, 2020, Shelton, 2021]. Search engines and recommendation systems, the key vehicles that deliver personalized content to users, are of particular relevance on these digital platforms. These systems are influential user interfaces that shape our digital diets. For example, recommendations drive 80% of the hours streamed on Netflix [Gomez-Uribe and Hunt, 2016], 70% of consumption on YouTube [Kiros, 2022], and 35% of sales on Amazon [MacKenzie et al., 2013].

In this paper, we present an "inclusive-by-design" solution to make online platforms more diverse and representative. We study the design and deployment of an inclusive recommendation system at one of the largest visual inspiration platforms in the world, Pinterest. Pinterest re-designed its recommendation system on its related Pins surface to promote more diverse and inclusive visual content.[2] Specifically, the new system strives for a more even representation of the skin tone ranges, as identified in the underlying content images while maintaining, or potentially improving, the quality of the content recommended. We study how this diversification of visual recommendations – 1) changes the distribution of content served to users, 2) influences top-line metrics such as engagement rates, and 3) shapes the diversity of content consumed over the long run.

The pursuit of inclusivity in recommendation systems has both a societal imperative and strategic underpinnings. This need is even more pronounced for visual-first platforms such as Pinterest. For example, in domains such as beauty and fashion, the representation of diverse skin tones can have profound implications on user satisfaction and content relevance. An exclusive focus on content with a narrow skin tone range could marginalize users who do not find content that resonates with their own identities.[3,4]

On the strategic front, recommendation systems that are designed to optimize for user engagement metrics, tend to create "feedback loops". These feedback loops can cause persistent homogeneous recommendations leading to "filter bubbles" [Pariser, 2011, Sunstein, 2001], which further reinforce homogeneity in recommendations, inadvertently sidelining diverse content [Chaney et al., 2018, Jiang et al., 2019]. Over time, narrowly focused content can impoverish the entire platform and potentially limit its market reach and appeal [Mansoury et al., 2020].

Both these perspectives offer strong motivation for diversifying recommendations. Research in academia, as well as industry, has recognized this need and has made substantial progress over the past decade in diversifying recommendations under different constraints [Agrawal et al., 2009, Chen et al., 2023, Wilhelm et al., 2018, Zhu et al., 2007]. However, much of the previous work has defined diversity as some variation of the number of distinct items, categories, or topics. A salient, but overlooked, dimension of diversity is the representation of people that are shown in the content. We extend the scope of this literature by introducing skin tone as a feature over which visual content can be effectively diversified to make platforms more inclusive.

We begin by describing the general class of recommendation systems employed on digital platforms and their limitations. We then outlay the design of a new system on Pinterest that

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

[2] The redesign was tested in Q2-2023.

[3] https://www.buzzfeednews.com/article/laurenstrapagiel/influencers-say-instagram-is-more-likely-to-remove-photos, (May, 2020)

[4] https://www.forbes.com/sites/janicegassam/2020/04/14/does-tiktok-have-a-race-problem/

promotes more diverse and inclusive visual content. Content items on Pinterest are either images or videos. Each item that includes people has an underlying skin tone signal that classifies it into one of four buckets based on the skin tone range of the people depicted in the image [Fawaz et al., 2020]. For the purposes of this work, we will call the skin tone ranges Lightest, Second Lightest, Second Darkest, and Darkest. The new system accounts for the similarity in skin tones and strives to surface content with a more even representation across the skin tone buckets while optimizing for engagement. This is operationalized using a Determinantal Point Process (DPP), a probabilistic model of repulsion that can be used to diversify a set of items [Kulesza and Taskar, 2012, Silva et al., 2023, Wilhelm et al., 2018]. DPP takes in utility scores and pairwise item similarities (or distances) from a candidate set of items to generate a "diversity-aware" set of recommendations, which in our case includes recommendations that are diversified over the underlying skin tone ranges.

With the new system in place, we study the impact of inclusive recommendations on user engagement using a large-scale field experiment on the platform. We randomize users on the platform to receive recommendations either from the old system (control) or the new re-designed system (treatment). Prior to the design of the new recommendation system, most of the surfaced recommendations belonged to "lighter" skin tone buckets (see Figure 4). The new system effectively increases the representation of deeper (darkest and second darkest) skin tone ranges, thereby achieving a more balanced representation across the four buckets (Figure 4b).

We report three sets of findings from the field experiment. First, on top-line metrics, we find that overall engagement, as measured by content items saved, is similar between the treatment and control groups. This result is robust to alternate measures of engagement and alternate model specifications.

Second, we find that consumption diversity goes up significantly. We measure consumption diversity using three metrics – a) engagement with deeper skin tone content, b) Shannon entropy over skin tone buckets, and c) proportion of users engaging with content from all four skin tone types. All three measures show a significant increase in the treatment group versus the control group.

Third, there is heterogeneity in the impact of diverse recommendations on overall engagement. We posit that users have a preference over skin tones in the content they see and because the initial set of recommendations were majorly concentrated within certain skin tone ranges, there might be users with preference for other skin tones whose needs weren't being adequately served. If there indeed are users who were being underserved by the older system, they are more likely to find relevant content with diversified recommendations and hence may experience an increase in overall engagement. We use historical data from the platform to segment users based on the diversity of content they are exposed to and engage with. Subsequently, we estimate treatment effects for these different user segments and find that users who had high diversity of exposure and high engagement with deeper skin tone content witness a ~10% increase in overall engagement.

Together, these findings are important and it is valuable to contextualize them in light of the broader literature. The "engagement-diversity" trade-off is well-documented and extant research has shown that an increase in engagement with personalized recommendations comes at the cost of reduced diversity in consumption, either at the individual level [Holtz et al., 2020] or at the aggregate level [Lee and Hosanagar, 2019]. We show that this conundrum of diversity need not hold universally and that platforms can meaningfully diversify recommendations without hurting top-line metrics. Crucially, we offer a practical blueprint for building inclusive algorithms and causal evidence endorsing their effectiveness. With growing scrutiny of potential algorithmic bias, our "inclusive-by-design" approach provides actionable insights for product managers. Our approach and findings are generalizable to other content-based digital platforms, especially those serving visual content. The recommendation system structure we describe is fairly standard in the industry

and the methodology we use for diversification is easily portable across contexts [Chen et al., 2017, Silva et al., 2023, Wilhelm et al., 2018].

1.1 Related work

Our work contributes to multiple strands of academic literature. Substantively, we contribute to a nascent but growing body of work on the design of inclusive products and on firm initiatives on diversity, equity, and inclusion. [Shulman and Gu, 2023] build an analytical framework to demonstrate how company culture and research bias influence investments in inclusive product design. [Aneja et al., 2023] show that making minority ownership of businesses salient on a restaurant review platform increases customer engagement and firm performance. [Hartmann et al., 2023] document an increase in demographic diversity in online display ads in the past years. They further show that this increase in diversity is associated with an increase in engagement with these ads. Our experiment extends this literature by providing real-world causal evidence on the effectiveness of inclusive product design on core engagement metrics.

Conceptually, we add to the literature in management that investigates the impact of recommendation systems and their design choices on consumer outcomes. For example, [Holtz et al., 2020] use a field experiment on Spotify to assess the impact of personalized recommendations on consumption diversity. They find an engagement-diversity trade-off: personalization increases engagement at a user level but reduces the diversity of content consumed. In a different setting, [Lee and Hosanagar, 2019] find the opposite effect where recommendations based on Collaborative Filtering increase individual-level consumption diversity but reduce aggregate sales diversity. More recently, [Chen et al., 2023] use a field experiment on an audio platform where randomly selected users are shown diverse recommendations. They find mixed results – for most users neither consumption diversity nor overall engagement went up. However, more active users saw an increase in their consumption diversity. Interestingly, all three papers described above define diversity in products/content as the number of unique categories. Our work directly broadens the scope of this literature by introducing a new dimension of content diversity – the skin tone of the images constituting the content.

Additionally, we connect with the literature on algorithmic design choices made by digital platforms and their implication on user outcomes. This line of work typically randomizes users into different versions of related algorithms to test the impact of a key data input on user behavior. For example, [Sun et al., 2023] modify the algorithm that powers Alibaba's home to exclude personal data for randomly selected users. Once personal data is removed, the algorithm switches to providing less relevant and more popular recommendations to users which eventually lowers engagement and decreases sales. Relatedly, [Lei et al., 2023], remove the use of external data provided by a leading search engine to a smaller engine. They find that removing the data access significantly lowers the click-through rate for the smaller search engine. In a slight departure from the trend, [Claussen et al., 2023] compare the performance of personalized recommendation engines with a human editor in the context of online news. They find that the recommendation system increases engagement as compared to a human editor when the system has sufficient data on user preferences. Finally, [Yang et al., 2023] evaluate the role of advertising information in ranking algorithms. They find that ranking algorithms that use advertising information can mitigate the cold-start problem in e-commerce and help improve platform outcomes. Our work is similar in spirit. We provide skin tone information to the recommendation system and explicitly include skin tone-based diversity as an optimization sub-routine to investigate the impact on engagement.

Finally and more broadly, we build on the vast literature in computer science and information retrieval on the diversification of recommendations and search results. Much of earlier work focused on building algorithms that can diversify content under different settings [Agrawal et al., 2009, Carbonell and Goldstein, 1998, Zhu et al., 2007], and evaluating the diversity of the generated results

[Clarke et al., 2008, Radlinski et al., 2009]. The earlier body of work reflected the problems faced by retrieval systems at the name – diversifying results based on sub-topics or sub-categories of the content. As social media gained popularity and the digital economy began to expand, researchers and practitioners began to notice diversity issues from a fairness and equity perspective [Geyik et al., 2019, Zehlike et al., 2017].

To summarize, we extend the notion of recommendation diversity beyond topics to include the skin tone of the images embedded in the content. We then describe the design of an inclusive recommendation system that accounts for this signal and strives for a more even representation across skin tone ranges. Finally, we present causal results on the impact of inclusive recommendations on user engagement and consumption diversity. To the best of our knowledge, there is no prior work that causally demonstrates the effectiveness of an "inclusive-by-design" product strategy in a real-world field setting.

2 EMPIRICAL SETTING

We study inclusive recommendations and how they influence user engagement on Pinterest, a global online platform used for content discovery, visual inspiration, and shopping. Pinterest is a visual discovery platform and each piece of content on it is called a Pin. Pins are bookmarks that people use to save content they like and they can be either images, videos, or products. Figure A1 in the Online Appendix shows examples of Pins served to a sample user. Pinterest serves over 489 million monthly active users[5] and has a corpus of over 12.5 billion Pins. To keep the discussion general, we refer to Pins as *items* or *content* in the paper.

Users on the platform consume content through multiple surfaces such as a personalized home feed, search, related Pins, and related products. All the surfaces are powered using in-house recommendation engines that provide personalized content to the users. In this paper, we focus on the related Pins surface and the recommendation system that powers it. The related Pins surface includes items that are recommended to a user after they click on a focal item[6]. It is the most popular surface on Pinterest and makes up for more than 50% of total impressions and ~40% of total engagement.

On the content side, each item on Pinterest that includes people has an underlying skin tone signal. This signal classifies the item into one of four buckets based on the skin tone range of the people depicted in the image [Fawaz et al., 2020]. For the purposes of this work, we will call the skin tone ranges Lightest, Second Lightest, Second Darkest, and Darkest. If an item does not have the image of a person, then the skin tone signal has no bucket assignment. Throughout the paper, deeper skin tone content refers to items that are classified as either having the "Second Darkest" or the "Darkest" skin tone bucket. Figure A2 shows examples of items from each skin tone bucket.

3 RECOMMENDATION SYSTEM (RE)DESIGN

Operating at the scale of millions of users and billions of items, modern recommendation systems rely on sophisticated machine learning pipelines to deliver personalized and relevant content to users. Furthermore, competing objectives such as short-term revenue vs. long-term retention necessitate complex designs with multiple components. Pinterest, like many other large digital platforms, powers its content feed using a state-of-the-art recommendation system. It is a two-stage deep neural network (DNN) based system, where the first stage is "retrieval" and the second stage "ranking". These two-stage type DNN-based systems are fairly common in large-scale industrial

[5]Pinterest Inc. 2024. Pinterest Announces Fourth Quarter and Full Year 2023 Results, Delivers Record High Users and Robust Margin Expansion. https://investor.pinterestinc.com/financial-results/quarterly-results/default.aspx

[6]Parallels from other platforms include the "products related to this item" carousel on Amazon product pages, and the playlist sidebar shown to a user while watching a video on YouTube.

applications [Covington et al., 2016, Huang et al., 2020, Meta, 2019, Wang et al., 2018, Zhang et al., 2020]. The widespread prevalence of similar systems makes our framework broadly applicable across platforms from different industries and imbues confidence in the generalizability of our findings. Figure 1 provides a pictorial depiction of such a system.

In the first stage (shown as the dark blue quadrilateral on the left), one or more candidate generation models filter a large corpus of items to a more relevant and manageable subset. This stage uses multiple DNN models that narrow the corpus from $10^6 - 10^{10}$ down to $10^2 - 10^3$. The DNNs are trained for coarse personalization with high recall. Speed is of the essence at this stage and hence the winnowing is done using fast similarity search methods such as approximate nearest-neighbors on high-dimensional embeddings. This effectively generates candidates with high visual similarity.

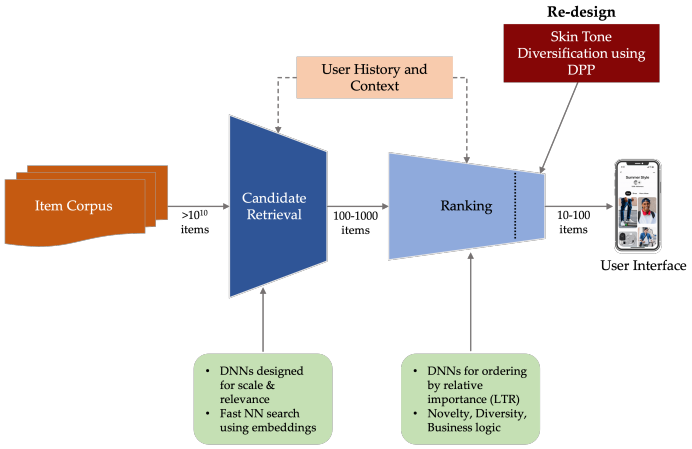


Fig. 1. Pinterest's recommendation systems and re-design using DPP

The second stage (shown as the light blue quadrilateral on the right in Figure 1) involves ranking where the goal is to order the candidate results from the previous stage in a way that maximizes a pre-specified objective or utility function. It is at this stage that layers with different objectives such as novelty, diversity, and business rules are applied. Though the specific utility objectives vary with the application, the ranking stage typically involves multiple learning-to-rank models for point-wise generation of a ranked list. Multiple models are required to optimize for different objectives such as the probability of an item's relevance to the query or the likelihood of user actions such as clicks, saves, or re-shares. The final ranking is produced using a multi-objective optimization that balances these objectives. We introduce our skin tone-based diversification in the ranking stage, as shown by the red quadrilateral on the top right of Figure 1. This augmentation succinctly captures the redesign of the recommendation system and our treatment in the field experiment that follows. We first describe the need for this layer and then explain the diversification process.

3.1 Need for diversification

The actual mechanics of the recommendation system described depend on the specific application. In our case, we are focused on producing related item recommendations. This is similar in spirit to producing recommendations for related products on Amazon or the playlist sidebar recommendations on YouTube. In such scenarios, there is a focal item/product/video for which

recommendations are sought and the recommended set tends to be similar to the focal item. This is because recommenders rely on correlated signals such as content similarity, which leads to redundant, non-diverse recommendations. In our study, the recommended items tend to be visually similar to the focal items. While the similarity between the focal and recommended items is a basic criterion of a good recommendation system, often the definition of similarity can be myopic and may overlook opportunities for relevant diversification. For instance, consider a user searching for fashion items such as summer dresses as shown in Figure A1. The user may expect to see recommendations that show similar items in terms of style, texture, fabric, and color. However, there may also be an appetite to see a variety of results based on the people wearing those clothes.

Figure 2a translates this anecdote to a data-driven argument. The figure shows the proportion of recommended impressions by skin tone conditional on the skin tone of the focal item. For example, the left-most stack shows the proportion of recommended impressions 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 focal items with the Darkest skin tone bucket, where the recommended set is more balanced. The need for diversification is motivated by Figure 2a and our goal is to strive for a more even representation of skin tones in visual recommendations.

3.2 Diversifying recommendations for inclusivity using DPP

The new recommendation system diversifies content using Determinantal Point Process (DPP) [Kulesza and Taskar, 2012]. Originating in physics, DPPs are now commonly used in machine learning for subset selection tasks while achieving some diversity criterion among the selected items. In the industry, they are also used to power YouTube's recommendation system [Wilhelm et al., 2018].

Before we describe DPP-based diversification, it is important to highlight that end-stage diversification is a thorny problem in recommendation systems. Simply promoting diversity as a secondary goal can hurt relevance as it completely ignores utility or item quality. Heuristics such as limiting the number of similar items or capping the number of items per category require manual tuning and are not adaptable to evolving systems. DPPs provide an elegant and scalable solution to the diversification problem. They are parameterized by a flexible kernel matrix that can encode complex item relationships easily, such as item embeddings learned through deep neural networks. We provide an intuitive description of DPP here, with some insights on design choices.[7]

Consider a set of items, $1, 2, \ldots N$ represented by $\mathbf{N}$. A DPP defines a probability measure over all subsets of $\mathbf{N}$ such that diverse subsets are more likely to be sampled. A DPP is parameterized by an $N \times N$ symmetric positive semi-definite kernel matrix, say $\mathbf{L}$, which encodes key information about item utility and diversity. Diagonal entries $\mathcal{L}_{ii}$ represent utility for item i and off-diagonal entries $\mathcal{L}_{ij}$ represent similarity between items i and j with respect to the chosen diversity dimension. The probability of a particular subset, $M \subseteq \mathbf{N}$ is proportional to the determinant of the kernel matrix $\mathbf{L}$ indexed by M. This is because, the determinant of sub-matrix $\mathcal{L}_M$ tends to be larger for subsets M that have high utility and low similarity, thereby promoting diversity when optimizing for utility. The following example will make this clear. Define the kernel matrix as follows:

$$L_{ij} = \begin{cases} q_i^2, & \text{if } i = j \\ q_i q_j f(d_{ij}), & \text{if } i \neq j \end{cases} \tag{1}$$

[7]For more in-depth explanations, please refer to the technical guide by [Kulesza and Taskar, 2012].

where, q_i is the utility score of item i, d_{ij} is the distance or dissimilarity between items i and j, and $f(\cdot)$ is a decreasing similarity function, e.g., Gaussian RBF: $f(d) = \exp(-d^2/2\sigma^2)$. For ease of exposition, consider the following kernel matrix with two items, 1 and 2:

$$L = \begin{bmatrix} q_1^2 & s_{12} \\ s_{21} & q_2^2 \end{bmatrix} \tag{2}$$

q_1 is the utility score for item 1, q_2 is the utility score for item 2, and $s_{12} = s_{21}$ captures the similarity between items 1 and 2 (they are equal because the kernel matrix is symmetric by construction). The probability of then selecting $M = \{1, 2\}$ is proportional to:[8]

$$P(\{1,2\}) \propto \det(L_M) = \det \begin{bmatrix} q_1^2 & s_{12} \\ s_{12} & q_2^2 \end{bmatrix} = q_1^2 \cdot q_2^2 - s_{12}^2 \tag{3}$$

The determinant balances utility ($q_1^2 \cdot q_2^2$) and diversity (s_{12}^2). If items 1 and 2 are very similar such that s_{12} is large, the probability of selecting them will be small. Conversely, if they are very different such that s_{12} is small, then the probability of selecting them is high. It is this property that allows DPP to balance the relevance-diversity trade-off.

In practical applications at scale, diversification methods such as DPP are typically applied at the second stage, i.e., the ranking stage after the candidate generation model has retrieved a set of potentially relevant items. In Pinterest's system, DPP takes the candidate items along with their relevance scores and similarity scores based on skin tone to produce a 'diversity-aware' rank-ordered list that balances utility and diversity scores. Figure 2b provides an overview of what happens after DPP-based diversification from the diversity perspective. Comparing with Figure 2a, we find that for a given focal item skin tone, DPP down-weights items from the same skin tone bucket and up-weights items from other skin tone buckets. This eventually leads to a more balanced representation across skin tone buckets.

4 FIELD EXPERIMENT AND DATA

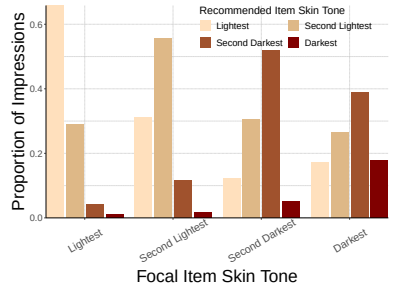
We ran a field experiment on the "related Pins surface" on Pinterest during April 2023 in which we randomized users on the platform into either the status quo recommendations or diversified recommendations from the re-designed system described in the previous section. The treatment was assigned at a user level for users in USA, Canada, Great Britain, Ireland, Australia, and New Zealand.

During the experiment period, users interacting with fashion-related focal items were triggered into either the control condition or the treatment condition. In the control condition, users were shown recommendations based on the current system, i.e., the status quo. These recommendations, as described above, heavily rely on the visual similarity between the candidate set and the focal item and hence tend to be more concentrated in terms of skin tone ranges displayed. In the treatment condition, users were shown more diverse content in terms of the visual skin tone of the underlying items. This diversification was achieved using DPP which, as described in the previous section, re-ordered the candidate items to produce a "diversity-aware" ranking that balances utility and diversity scores. Figure 3 presents an example of a focal item and the corresponding recommendations shown in the control and treatment conditions. Skin tone diversity among the recommendations served in the treatment condition is apparent.

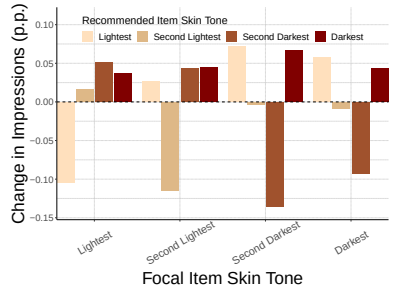
DPP diversifies each fashion-related recommendation query[9] made by users in treatment. This resulted in users' content becoming visually more inclusive and having a more even distribution

[8] See [Kulesza and Taskar, 2012] for the proof.

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



(a) Proportion of impressions by skin tone conditional on focal item



(b) Change in impressions (p.p.) by skin tone after DPP diversification

Fig. 2. Distribution of recommended content conditional on focal item skin tone

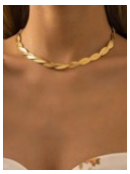
across skin tone ranges. We quantify the distributional impact of diversification on recommended items using the following metric

$$\text{Div@k(R)} = \frac{1}{|Q|} \sum_{q \in Q} \prod_{d_i \in \mathcal{D}} \mathcal{I}\left[d_i \in R_k(q)\right] \tag{4}$$

where q is a query belonging to Q, the entire set of queries. R is a ranking algorithm that produces an ordered set of results from which we select the top k results. $\mathcal{D}$ 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 recommended results. During our experiment, the proportion of queries in which all four skin tone buckets were represented among the top-20 recommended items increased 3-fold in the treatment group relative to the control group, as shown by the last bar in Figure 4a. Furthermore, in attempting to balance the representation across skin tones, DPP diversification also increased the exposure to content with deeper skin tones. In Figure 4b, we plot the proportion of impressions by each skin tone among the top 20 recommended items. We see that treatment increased the proportion of impressions for deeper skin tone content by ~33%. We investigate the impact of this provision of more visually inclusive recommendations on user engagement. In the Online Appendix, we show the shift in distribution between treatment and control using other measures of diversity.

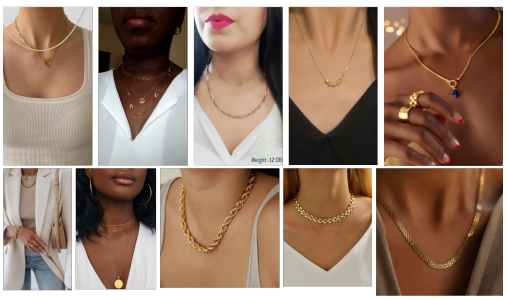
4.1 Data

Our working data contains a randomly drawn proportion of all users in the experiment. The sample contains ~320,000 users from each experimental condition from March 28, 2023, to April



(a) Example focal item for which recommendations are sought

(b) Recommendations in control

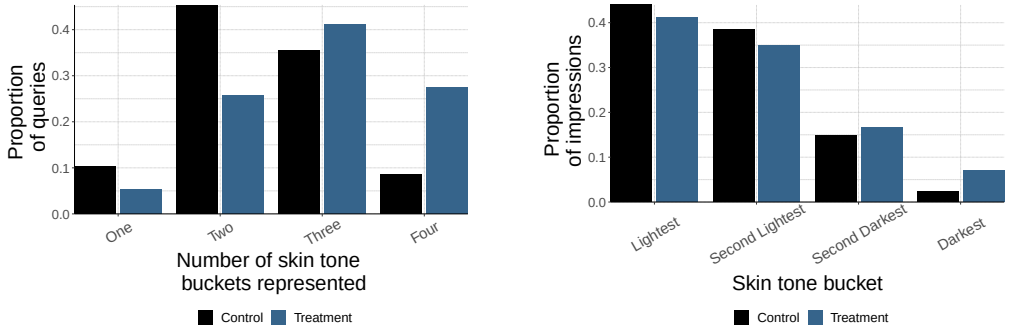


(c) Recommendations in treatment

Fig. 3. Recommendations: treatment vs. control

28, 2023. These users represent an undisclosed fraction of users who were randomized into the experiment. For each user, we observe their experimental assignment and their activity on the platform. Importantly, for each recommendation query generated by the user, we observe the focal items for which the recommendations are generated, the recommended items by position, and whether the user engaged with them. We also observe the time stamp of each query which we use for testing novelty effects.

Outcomes of interest. Our goal in this paper is to investigate how inclusive recommendations 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 outcome metrics, which collectively form our outcome variables in subsequent analysis.



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

(b) Proportion of impressions by each skin tone bucket

Fig. 4. Distribution of recommended content in treatment vs. control.

We measure overall engagement using the number of total items *Repinned* from the recommended set by each user. A Repin is equivalent to saving an item and hence for generality, we will use the term "saving" or "saved" henceforth. Like most other measures of online user activity, items saved has a long-tail [Brynjolfsson et al., 2006]. Consequently, we consider the top 20 recommended items for a given query, which collectively represent majority of the total engagement. Once the recommendations are surfaced, 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 the number of queries made by the user to get the total number of items saved. We mainly focus on the number of items saved since it is a key engagement metric used by the platform to make product decisions. In the Online Appendix, we check the robustness of our results with other measures of engagement.

Recall that our primary motivation behind designing an inclusive recommendation system is to promote the representation and consumption of more diverse content. In this spirit, we measure consumption diversity using three metrics that capture different aspects of diversity – 1) the number of deeper skin tone items saved, 2) the Shannon entropy of items saved computed over the four skin tone buckets, and 3) the proportion of users engaging with content from all four skin tone buckets.

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 proportion of impressions (relative to the control group) for deeper skin tone items (Figure 4b) and hence the increase in engagement with this content gives a direct measure of the impact of treatment on consumption diversity.

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

$$S^i = -\sum_{b \in \mathcal{B}} s_{bi} \ln(s_{bi}) \tag{5}$$

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 a simple and easily interpretable measure for what proportion of the user base is engaging with diverse content.[10] 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.

5 EMPIRICAL ANALYSIS

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 recommendations. The recommendation system then produces an ordered set of items that are relevant to the user's query. Once the recommendations 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. As mentioned earlier, given the long-tail nature of online user activity, we only consider the top 20 recommended items for a given query. 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

$$Y_i = \alpha + \beta T_i + \epsilon_i \tag{6}$$

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 among the top 20 recommended items across all queries made by the user. β captures the effect of getting exposed to more visually diversified recommendations by content skin tone 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 an 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 recommendations and so on. Hence, we focus our main analysis on aggregate user-level results that can answer a simple but important question, "How do the top-line metrics change once users are exposed to diverse and inclusive recommendations"? Nevertheless, we show the robustness of our results by re-estimating Model 6 at the query level, at the user-query-item level, and by using only the first query of each user. All robustness tables are present in the Online Appendix.

5.1 Average treatment effect

We estimate the treatment effect using Model 6 on different measures of engagement described in Section 4.1. We begin with overall engagement, which is the total number of items saved by a user during the course of the experiment. Like many other measures of online engagement, this measure is a long-tail outcome with a huge mass at zero. Hence, we use a Quasi-Poisson specification to

[10] 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.

estimate the treatment effect on a log scale. We check for robustness to functional form assumptions in the Online Appendix by re-estimating the model with OLS.[11]

Next, we focus on how the diversity of content engaged with changes, which we refer to as "consumption diversity". Focusing on consumption diversity is important for two reasons. First, it gives insight into how the engagement patterns are likely to change once the new system is launched platform-wide. This is valuable for managers as it helps inform subsequent content strategy for platform growth and monetization. Second, extant research has shown that diverse consumption is associated with better long-term outcomes such as retention [Anderson et al., 2020, Wang et al., 2022]. If our treatment causes consumption diversity to increase, then this could potentially have favorable long-term implications.[12] The first measure for consumption diversity is the number of deeper skin tone items saved. This, like overall engagement, is a long-tail outcome with a mass at zero and hence we estimate it on a log scale using a Quasi-Poisson specification. The second measure, Shannon entropy, accounts for the evenness in the distribution of engagement over the four skin tone buckets [Chen et al., 2023, Holtz et al., 2020]. The last measure captures whether users are engaging with content from all 4 skin tone buckets. This provides a simple and easily interpretable measure for what proportion of the user base is engaging with diverse content. This is a binary variable and we estimate it using a logistic regression. Robustness with OLS is shown in the Online Appendix.

The results are shown in Table 1. Each column corresponds to the different outcome measures described above. In all columns, we suppress the intercept to protect sensitive information. Since the regression specification is Quasi-Poisson, the coefficient can be directly read as an approximate percentage change. For ease of interpretation, in Column 3 for Shannon entropy, we normalize the raw outcome in the control group to 100 so that the coefficient on the treatment indicator can be read off as a percentage change. Heteroskedasticity-robust standard errors are shown in parentheses.

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. Our result is in line with previous research on the impact of diversification on engagement [Chen et al., 2023]. Additionally, the result is consistent if we define engagement using other common measures such as the number of successful queries (queries with at least 1 save) or the daily active user rate.

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$15%, and the proportion of users engaging with content from all four skin tone buckets increases by 70% ($\exp(0.5363) - 1 = .70$). These results are strong, significant, and important from a managerial perspective. Taken together, they indicate that platforms can meaningfully increase consumption diversity without hurting top-line engagement rates.

5.2 User preferences vs. mechanical effects

Admittedly, one may worry that the consumption diversity effects shown in Table 1 are largely mechanical. For example, the content that gets boosted up in the ranks by the algorithm gets more engagement. We believe that this is plausible but it is important to contextualize this statement in light of our motivation and the purpose behind building recommendation systems in the first

[11] Arguably, one can also estimate this model by $\log(\cdot + 1)$ transforming the outcome. While this does address the long-tail issue, there is still a huge mass at zero and a Quasi-Possion specification provides more statistical power in this scenario.

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

Table 1. Average treatment effect on engagement and consumption diversity

	Overall Engagement		Consumption Diversity	
Dependent Variables:	Total Saves	Saves of Deeper ST Content	Shannon Entropy	Saved All 4 ST (Binary)
Model:	(1)	(2)	(3)	(4)
	Poisson	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 diversified recommendations on user engagement. The first column shows estimates from a Quasi-Poisson regression of the total number of items saved on the treatment indicator (Results with OLS are in the Online Appendix). Coefficients in all columns can be interpreted as percentage change. Columns 2, 3, and 4 regress measures of consumption diversity on the treatment indicator. Heteroskedasticity-robust standard errors are shown in parentheses. Signif. Codes: **: 0.05, *: 0.1

place. An existentialist argument supporting the genesis of recommendation systems is precisely the results described above – recommendation systems should serve content that users would like to consume and hence a direct test of their performance is whether users actually consume the content recommended. In this section, we provide some analytical arguments to shed light on the debate between satisfying user preferences vs. documenting mechanical treatment effects.

First, it is worth pointing out that just because users are shown a certain type of content it does not mean that they need to necessarily engage with it. User agency is critical in separating passive consumption from active engagement. To this end, it is important to select the appropriate metric for analysis. For example, using "up-the-funnel" or less costly metrics such as clicks are likely to be more noisy and more prone to mechanical response. Hence, for this study, we use the number of items saved as the key outcome measure. The platform considers this to be among the most important top-line metrics that represent a "deeper" form of engagement.

We provide three pieces of data-based evidence to support this argument. First, if users don't prefer the recommended content, they can either scroll down the ranked list of recommendations or not engage altogether, both of which would result in lower overall engagement. If this were the case, we would have seen a lower overall engagement in the treatment group. However, as shown in Column 1 of Table 1, the overall engagement rate is stable.

Second, we estimate explicit user *dissatisfaction* to show that users prefer the content served to them. For each piece of content shown to the user, they have the option to *hide* the item using a single click. We estimate the impact of treatment on the number of items hidden in treatment and control. If users were being shown irrelevant content that does not satisfy their preferences, then we would see an increase in the *hide rate*. However, this is not the case. *Hide rate* is similar for users in treatment and control (See Table F8 in the Online Appendix for results).

Third, we ameliorate the concern that the treatment effect is essentially an effect driven by position bias, where deeper skin tone content is moved up the ranking and hence gets more engagement due to its higher position. To test this, we use the user-query-item level data to estimate treatment effects for each position separately. Figure E4 in the Online Appendix echoes the result from Table 1. Across all top 20 positions, we find that the overall engagement remains stable and engagement with deeper skin tone content goes up significantly. In addition to the points above, we assess longer-term persistence in the predicted effects. We use data five months after the platform-wide launch of the system to track consumption diversity. The approach and results are described in Section 6.1.

5.3 Treatment effect by user characteristics

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

About 77% of the users report their gender to be female and ~ 79% of the users are from the US. We do not find any meaningful differences across user demographics. Table G9 in the Online Appendix shows the estimated coefficients. We find consistent results of stability in overall engagement and an increase in engagement with deeper skin tone content. Similarly, we estimate treatment effects for new vs. old platform users, where new users are those who joined the platform during the experiment period. New users make up ~ 4% of our sample. We do not find any substantial difference in the impact of the treatment on engagement. The results are shown in Table G10. We explore heterogeneity in treatment effects further in Section 6 where we decompose the changes in engagement across key behavioral dimensions.

6 MECHANISM AND IMPLICATIONS

The results in Section 5 show that exposure to inclusive recommendations keeps user engagement, as measured by items saved, stable. Here, we further dig deeper to understand the mechanism driving these results. An aggregate *null effect* could either be driven by a uniform near zero effect across the user spectrum or it could be masking potential heterogeneity in user response. To uncover the underlying mechanism, we posit whether the diversified recommendation system improves the match quality between recommended 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 in recommendations, users with a preference for this type of content may have been underserved. The new recommendation algorithm pushes for a more equal representation of content which then lowers the costs of accessing deeper skin tone content for users with a strong preference for it.

To test for this, we consider users' prior exposure to and engagement with deeper skin tone content. We segment users in our experiment on two dimensions - diversity of content exposed to in the 3-month window prior to the experiment and engagement (saves) with deeper skin tone content in the 3-months prior to the experiment.[13] We include both exposure and engagement in this segmentation to account for users who were "actively searching" for diverse content before the experiment. Since recommendations from the older algorithm were largely concentrated in the two lighter skin tone buckets, users with high exposure to deeper skin tone content would have had to manually search for it. Accounting for this information then allows us to better segment users based on their preferences.

We quantify the diversity of exposed content by computing the Shannon Entropy of exposure for each user using pre-experiment data [Chen et al., 2023, Holtz et al., 2020]. Historical engagement with deeper skin tone content is calculated as the number of deeper skin tone content items saved divided by the total number of impressions. This gives us a direct measure of how much the user prefers deeper skin tone content normalized by all the items they have been exposed to. We then use the 80-20 rule to create a 2x2 matrix based on whether a user was in the top-pentile of these

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

dimensions or not. Summary statistics as per this segmentation are shown in Table H11 in the Online Appendix.

After segmenting the users, we classify those who are in the top-pentile of exposure to diverse content and the top-pentile of engagement with deeper skin tone content as users with *preference for deeper skin tone content*. The remaining set of users is classified as users with *preference for lighter skin tone content*. Under this classification scheme, we re-estimate the treatment effects for both user groups separately. The results are shown in Figure 5. We indeed find that the treatment effect on overall engagement for some users is significantly greater than zero. Users with *preference for deeper skin tone content* see about a 10% increase in overall engagement. The treatment effect on overall engagement for users with *preference for lighter skin tone content* is negative, but statistically indistinguishable from zero. When aggregated and weighted by the corresponding number of users in each group, this gives a total effect that is close to zero. We provide robustness checks for this specification in the Online Appendix.

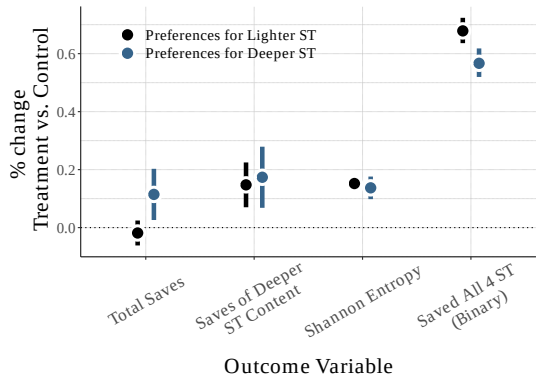


Fig. 5. Heterogeneous treatment effects on engagement and consumption diversity

6.1 Longer-term persistence

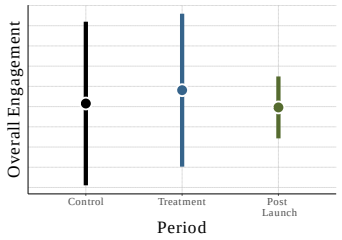
The new recommendation system was launched to all users on the platform in May 2023. We track activity and engagement data for a random sample of ~800,000 users five months after the launch in September 2023. We use this data to assess longer-term persistence in the predicted effects. The examination of user engagement and consumption patterns five months post the deployment of the new system provides critical longitudinal insights into the lasting impact of inclusive content recommendations.

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

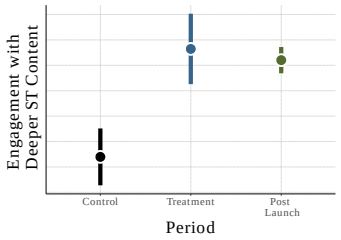
For this analysis, we compute our four outcome variables – 1) overall engagement, 2) engagement with deeper skin tone content, 3) Shannon entropy, and 4) engagement with content from all 4 skin tone buckets. We plot the four outcome variables in separate panels in Figure 6. For each variable, we show their mean values in control and treatment during the experiment. Additionally, we show the mean five months after the platform-wide launch of the new diversified recommendation system. In each case, we also plot the 95% confidence intervals. The intervals for the post-launch

period are much shorter since we have substantially more data, both cross-sectionally and over time.

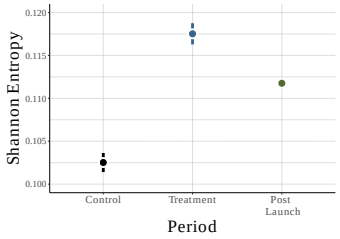
We find that overall engagement is relatively stable. Further, while the increase in consumption diversity experienced during the experiment period has somewhat moderated over time, it continues to be substantially higher than the pre-launch baseline (control group during the experiment period). Note that the post-launch sample is a randomly selected subset of users *not* included in the experiment.[14] Consequently, this sustained longer-term persistence in consumption diversity is encouraging and helps inform the platform's subsequent content strategy while achieving broader social goals of fostering inclusive digital environments.



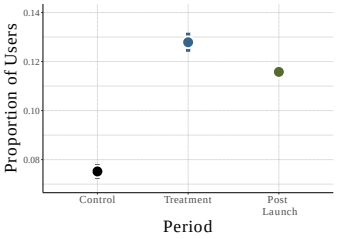
(a) Overall engagement



(b) Engagement with deeper ST content



(c) Shannon entropy across skin tone buckets



(d) Engagement with content from all 4 ST buckets

Fig. 6. Overall engagement and consumption diversity four months post-launch of the diversified recommendation system to all users

7 ROBUSTNESS

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 Appendix.

These include – 1) alternate functional forms (OLS), 2) alternate outcome variables (number of queries with at least one save and number of unique login dates), 3) additional control (gender, country, and user activity index), 4) multiple data aggregations (user-query level, user-query-item level, only first query for each user after being triggered into the experiment), 5) novelty effects, 6) potential algorithmic spillover across treatment arms, and 7) robustness for heterogeneous treatment effects (interactive form, estimates across all pentiles).

[14]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 five months after the launch are likely to be more engaged and hence their outcome measures will not be representative of the entire user base. Hence, 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.

8 DISCUSSION

We study the design and deployment of an inclusive recommendation system at one of the largest visual content discovery platforms in the world, Pinterest. The new system is designed to achieve a more even representation of skin tones in its recommendations. Items on Pinterest are largely made up of images or videos and items that include people have an underlying skin tone signal. This signal classifies the item into one of four buckets, which we label as Lightest, Second Lightest, Second Darkest, and Darkest. Diversification in the new recommendation system accounts for the similarity in skin tones to surface content that is more balanced in representation across these ranges. This is operationalized using a Determinantal Point Process that takes in item relevance and similarity scores to generate a "diversity-aware" ranking that balances utility with diversity. As designed, the system increases the proportion of queries where all four skin tones are represented in the top 20 recommended items by more than 200%. Since deeper skin tone content was initially under-represented, the diversification process increased their impressions by 33%.

We then investigate the impact of inclusive recommendations on user engagement by running a field experiment where we randomize users into receiving either status quo recommendations (control) or more visually inclusive recommendations (treatment). We find that users in the treatment group respond favorably to the new system, overall engagement rates are stable and consumption diversity increases significantly. Specifically, engagement with deeper skin tone content, Shannon entropy of engagement across skin tone buckets, and proportion of users engaging with content from all four skin tone buckets increase by 15%, 15%, and 53% respectively. We provide evidence that these results show satisfaction of user preferences by the new system rather than just mechanical effects.

Additionally, we use historical data to uncover the mechanism behind the results. We create user segments based on pre-treatment exposure to diverse content and pre-treatment engagement with deeper skin tone content. We classify users in the top-pentile of these segments as users with "preference for deeper skin tone content" and the others as users with "preference for lighter skin tone content". Subsequently, we estimate the treatment effects for these users separately and find that overall engagement for users with "preference for deeper skin tone content" increases by ∼10%. Users with "preference for lighter skin tone content" have a small negative treatment effect that is statistically indistinguishable from zero. Since overall engagement is stable for the majority of the users and it goes up substantially for users with a strong preference for deeper skin tone content, we posit that the increased diversity of recommendations in the new system increases the match value of the content for these users.

Finally, we study user activity and engagement five months after the platform-wide launch of the new recommendation system. Tracking persistence in the effects witnessed during the experiment period allows us to better understand longer-term implications for the platform's content strategy. We find that overall engagement levels continue to remain stable and consumption diversity continues to remain elevated as compared to the pre-launch baseline. Not only does this complement the experimental findings, but it also indicates the platform's potential role in fostering a more inclusive digital culture. Platforms, by virtue of their recommendation algorithms, have the power to shape societal narratives and user perceptions. A consistently diverse content consumption pattern means that users are being exposed to varied perspectives, cultures, and ideas, promoting inclusivity and improving representation.

Our research illuminates the transformative potential of inclusive recommendation systems, offering invaluable insights for platform managers and policymakers. By embracing diversity in their recommendations, platforms can create digital spaces that resonate with users' multifaceted preferences and aspirations, while simultaneously cultivating a sense of inclusivity and belonging.

We hope that this study paves the way for a future where equitable algorithms drive engagement and foster an enriched user experience.

REFERENCES

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

Albergotti, Reed, "Black creators sue YouTube, alleging racial discrimination," Technical Report 2020.

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

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

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

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

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

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

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

Clarke, Charles L.A., Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon, "Novelty and Diversity in Information Retrieval Evaluation," in "Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval" SIGIR '08 Association for Computing Machinery New York, NY, USA 2008, p. 659–666.

Claussen, Jörg, Christian Peukert, and Ananya Sen, "The Editor and the Algorithm: Recommendation Technology in Online News," *Forthcoming, Management Science*, 2023.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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.

Meta, 2019.

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

Radlinski, Filip, Paul N. Bennett, Ben Carterette, and Thorsten Joachims, "Redundancy, Diversity and Interdependent Document Relevance," *SIGIR Forum*, dec 2009, *43* (2), 46–52.

Shelton, Deborah, "Charged with racial discrimination, Airbnb promised to 'mitigate the bias'," Technical Report 2021.

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

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

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, 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.

Yang, Joonhyuk, Navdeep S. Sahni, Harikesh S. Nair, and Xi Xiong, "Advertising as Information for Ranking E-Commerce Search Listings," *Marketing Science*, 2023.

Zehlike, Meike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates, "FA*IR: A Fair Top-k Ranking Algorithm," in "Proceedings of the 2017 ACM on Conference on Information and Knowledge Management" CIKM '17 Association for Computing Machinery New York, NY, USA 2017, p. 1569–1578.

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.

ONLINE APPENDIX

A CONTENT ON PINTEREST

A.1 Example items for a representative user

Figure A1 shows the search results shown to a user when the search bar in the top shows the type “summer dress” in the search bar on the top. The search engine produces a mix of items in order of relevance. Say the user clicks on the first item from the second row, which we call the “focal item”. This navigates the user to the related Pins surface where they are recommended related items. In this paper, we study the recommendation system that powers the Related Pins surface.

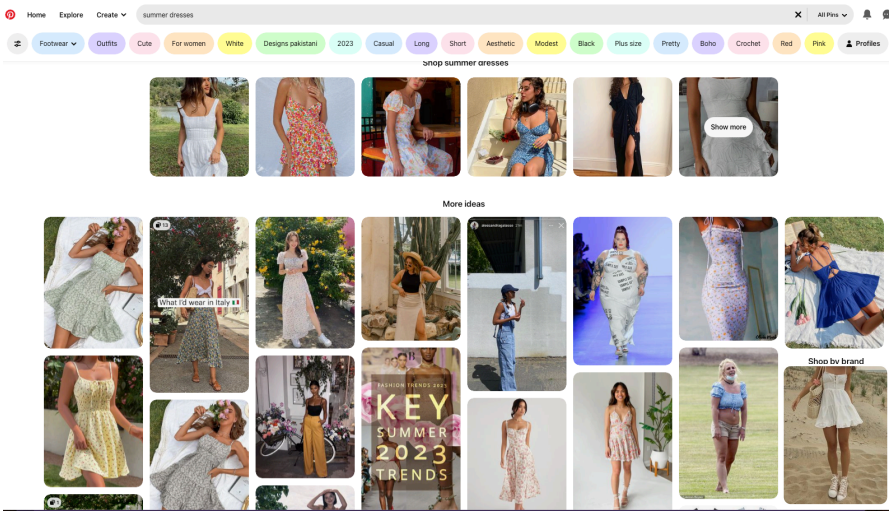


Fig. A1. Example Pins/items shown to a user

A.2 Example items from each skin tone bucket

Every content item on Pinterest that includes people has an underlying skin tone signal. This signal classifies the item into one of four buckets based on the skin tone range of the people depicted in the image [Fawaz et al., 2020]. For the purposes of this work, we will call the skin tone ranges Lightest, Second Lightest, Second Darkest, and Darkest. If the item does not have the image of a person, then the skin tone signal has no bucket assignment. Figure A2 shows examples of items from each skin tone bucket.

(a) Sample Pin from Lightest skin tone bucket

(b) Sample Pin from Second Lightest skin tone bucket

(c) Sample Pin from Second Darkest skin tone bucket

(d) Sample Pin from Darkest skin tone bucket

Fig. A2. Example Pins from each skin tone bucket

B RECOMMENDATION DIVERSITY

In the main text we use two measures of diversity to quantify the change in the distribution of content recommended after diversification based on skin tone. These include – 1) Div@20(R), i.e., the proportion of queries with all four skin tone buckets represented in the top 20 results, and 2) the proportion of recommended impressions by skin tone bucket. Here we show how the Shannon entropy of exposure changed in treatment and control. Figure B3 shows how the Shannon entropy shifts to the right in the treatment group, indicating that exposure is more evenly distributed across skin tone buckets.

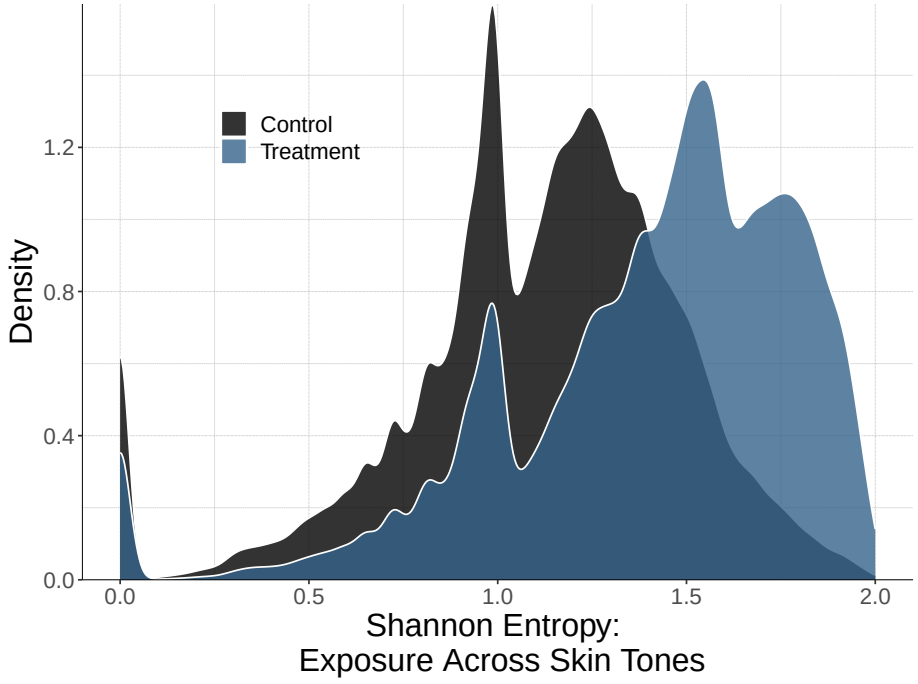


Fig. B3. Shannon entropy of exposure computed over the 4 skin tone buckets

C RANDOMIZATION CHECKS

In addition to the data during the experiment, we also observe three months of pre-experiment data. This includes their historical activity and engagement on the platform. We present randomization checks for key pre-treatment variables, including historical overall engagement and consumption diversity described above in the top panel of Table C1. We only show the difference in means between the two conditions (in percentage terms relative to control) to protect confidential information. Standard errors are shown in parentheses and the third column reports p-values from a t-test. We find no evidence of imbalance across these variables.

The bottom panel of Table C1 shows the difference in means of exposure to deeper skin tone content between treatment and control during the experiment period. We formally test whether the treatment works as intended. The p-value in the last column confirms this is indeed the case.

Table C1. Relative difference in means for key variables (treatment vs. control)

Variable	Difference in means (%)	p-value
Pre-treatment variables - balance check		
Gender (Female = 1)	0.0022	0.1067
	(0.0013)	
Number of queries	−0.0053	0.5342
	(0.0085)	
Exposure to deeper skin tone content	−0.0005	0.9078
	(0.0046)	
Overall engagement rate	0.0048	0.3554
	(0.0052)	
Engagement rate with deeper skin tone content	0.0128	0.2083
	(0.0102)	
Shannon entropy	0.0014	0.6221
	(0.0029)	
Proportion of users engaging with all 4 ST	0.0014	0.5916
	(0.0026)	
Experiment period		
Proportion of queries with all 4 ST buckets	2.3057	< 0.001
	(0.0089)	
Exposure to deeper skin tone content	0.3369	< 0.001
	(0.0141)	

1) The top panel shows user-level balance checks for key variables between the treatment and control conditions. There are $\sim 320,000$ *users in each condition. "Deeper skin tone" content includes items that are classified as either having the "Second Darkest" or the "Darkest" skin tone bucket. 2) The second column shows the difference in means between the two conditions. Standard errors are in parentheses. 3) The bottom panel formally tests whether the treatment increases exposure to deeper skin tone content, as intended. 4) ST is an abbreviation for skin tone.*

D ROBUSTNESS CHECKS: ATE MEASURED USING ALTERNATE MODEL SPECIFICATIONS

Our first robustness check involves re-estimating Model 6 using OLS. The results are shown in Table D2. We report estimates for overall engagement measured using total items saved and for three measures of consumption diversity. Our results are consistent – 1) overall engagement rates are stable and the treatment effect is indistinguishable from zero and 2) consumption diversity increases significantly.

Table D2. Average treatment effect on engagement and consumption diversity (Estimated using OLS)

	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)
Variables				
Treatment	0.4604	16.09**	14.63**	70.09**
	(1.974)	(3.443)	(0.8259)	(3.336)
Fit statistics				
Observations	638,683	638,683	638,683	638,683
Adjusted R^2	-1.48×10^{-6}	3.26×10^{-5}	0.00049	0.00069

This table checks the robustness of our main results to functional form. We estimate the impact of treatment on engagement and consumption diversity using OLS. The first column shows estimates from OLS of the total number of items saved on the treatment indicator. The mean value of respective outcomes has been normalized to 100 in the control group. Hence, coefficients in all columns can be interpreted as percentage change. Columns 2, 3, and 4 regress measures of consumption diversity on the treatment indicator. Heteroskedasticity-robust standard errors are shown in parentheses. Signif. Codes: **: 0.05, *: 0.1

Further, we estimate the overall impact on engagement using other common measures of engagement employed by the platform. These include – 1) the number of successful queries made by the user, where a query is considered successful if the user saves at least one of the recommended items, and 2) the number of unique dates a user visited the platform and made at least one query. The results are shown in Table D3. Both measures show similar results – overall engagement rates are stable and the treatment effect is not different from zero.

Table D3. Robustness check: Average treatment effect using alternate measures of engagement

Dependent Variables:	Number of Successful Queries	Days Logged In
Model:	(1)	(2)
	Poisson	Poisson
Variables		
Treatment	0.0078	-0.0038
	(0.0169)	(0.0030)
Fit statistics		
Observations	638,683	638,683
Log-Likelihood	-2,265,987.7	-1,852,324.8

This table checks the robustness of our main results with alternate measures of overall engagement. In Column 1, engagement is defined as the number of successful queries made by a user, where a query is considered successful if the user saves at least one item from the top 20 recommended items. In Column 2, we use daily average users (DAU) as a measure of engagement which we tabulate as the number of unique days a user logs onto the platform and makes at least one query. Both are count variables estimated on a log scale using a Quasi-Poisson specification. Heteroskedasticity-robust standard errors are shown in parentheses. Signif. Codes: **: 0.05, *: 0.1

Lastly, we re-estimate treatment effects while controlling for use demographics (self-reported gender and country), user cohort, and user activity index. The user activity index is Pinterest's internal classification of users based on their activity history. The results are reported in Table D4. The results are consistent with the previous findings.

Table D4. Robustness check: Average treatment effect on engagement and consumption diversity after controlling for user demographics, cohort, and activity index

	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)
	Poisson	Poisson	OLS	Logit
Variables				
Treatment	0.0032	0.1481**	0.1456**	0.5441**
	(0.0195)	(0.0316)	(0.0078)	(0.0261)
Demographics	✓	✓	✓	✓
User Activity Index	✓	✓	✓	✓
Cohort	✓	✓	✓	✓
Fit statistics				
Observations	638,683	638,683	638,683	638,683
Log-Likelihood	-130,493,786.6	-204,417,594.7	-4,575,235.3	-30,822.0

This table checks the robustness of our main results to alternate model specifications. We estimate the impact of treatment on engagement and consumption diversity after controlling for baseline pre-treatment variables such as user demographics (self-reported gender and country), cohort, and the user activity index. The user activity index is an internal classification made by the platform to segment users based on how active they are on the platform. The first column shows estimates from a Quasi-Poisson regression of the total number of items saved on the treatment indicator. Coefficients in all columns can be interpreted as 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

E ROBUSTNESS CHECKS: ATE MEASURED AT DIFFERENT DATA AGGREGATION LEVELS

Our main specification, Model 6 is estimated at a user level, i.e., we aggregate all observations for a user and then estimate the impact of the treatment on different measures of engagement. Here, we check the robustness of our results by estimating treatment effects at the user-query level. Specifically, we estimate the following regression:

$$Y_{iq} = \alpha + \beta T_i + \epsilon_{iq} \tag{7}$$

where i is a user and q is a query made by the user. A query means clicking on a focal item and generating a request for surfacing recommendations. T_i is a binary variable indicating the user's treatment status. β captures the effect of getting exposed to more visually diversified recommendations by content skin tone on user engagement. Y_{iq} represents the engagement by user i for query q. This could mean either the total number of saves among the top 20 recommended items in case of overall engagement or the total number of saves of deeper skin tone content among the top 20 recommended items in case of consumption diversity. Note that, in this case, we only focus on engagement with deeper skin tone content as the measure of consumption diversity. This is because the other two measures, Shannon entropy and proportion of users engaging with content from all four skin tone buckets, are user-level metrics and don't have a direct correspondence for query-level results. The results are shown in Table E5.

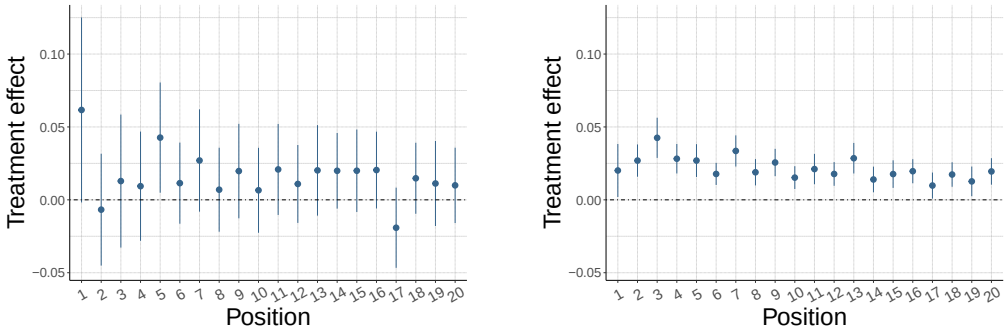
We find that the estimates from query-level regression are qualitatively similar to the estimates from the user-level regression. Overall engagement rates are similar between treatment and control groups and the treatment effect is not statistically different from zero. Consumption diversity, as measured by engagement with deeper skin tone content goes up significantly.

Table E5. Treatment effects on engagement with a Quasi-Poisson specification

Dependent Variables:	Engagement			Engagement with Deeper ST		
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Treatment	0.0200	0.0179	0.0171	0.1647***	0.1608***	0.1602***
	(0.0173)	(0.0173)	(0.0171)	(0.0302)	(0.0301)	(0.0301)
Demographics		✓	✓		✓	✓
Cohort			✓			✓
Date			✓			✓
Fit statistics						
Observations	9,252,921	9,252,921	9,252,921	9,252,921	9,252,921	9,252,921
Log-Likelihood	-3,866,003.1	-3,831,810.9	-3,827,025.1	-925,124.2	-915,627.9	-913,391.1

*1) The table shown regression results from estimating Model 7 at a query level as a Quasi-Poisson model. 2) Demographics include user's self-reported gender and their country. 3) Overall engagement is measured using total saves among the top 20 recommended items. Engagement with Deeper ST is measured using saves of only deeper skin tone content. 4) Columns 1-3 show the treatment effect on overall engagement with various controls. Columns 4-6 repeat the analysis for engagement with deeper skin tone content. 5) Clustered (User) standard errors in parentheses. Signif. Codes: **: 0.05, *: 0.1*

In addition to query-level data, we estimate the treatment effects using query-item-position-level data. This is the most granular form of data available which records which item was recommended at what position and whether the user saved it. We estimate treatment effects in 3 ways with this data – 1) in a simple position agnostic way, 2) linearly accounting for position in a regression, and 3) separately estimating for each position. The third specification is our preferred one and its results are shown in Figure E4. The left panel shows treatment effects for overall engagement across the top 20 positions and the right panel shows it for engagement with deeper skin tone content. We find consistent results in both cases. Results for specifications 1 and 2 are shown in Table E6.



(a) Overall engagement

(b) Engagement with deeper skin tone content

Fig. E4. Treatment effect on overall engagement measured at each position separately using query-item-position level data

Table E6. Average treatment effect on engagement using query-item-position level data

Dependent Variables:	Engagement		Engagement with Deeper ST	
Model:	(1)	(2)	(3)	(4)
Variables				
Treatment	0.0174	0.0175	0.0232^{***}	0.0185^{***}
	(0.0155)	(0.0154)	(0.0043)	(0.0041)
Position		✓		✓
Fit statistics				
Observations	128,551,286	128,551,286	128,551,286	128,551,286
Adjusted R^2	8.22×10^{-7}	0.00146	9.27×10^{-6}	0.00184

*1) The table shown regression results from estimating Model 7 at a query-item-position level using OLS. 2) Column 1 shows the treatment effect of inclusive recommendations on overall engagement and Column 2 linearly controls for the position. Columns 3 and 4 repeat the analysis for engagement with deeper skin tone content. 3) Overall engagement is measured using total saves among the top 20 recommended items. 4) Engagement with Deeper ST is measured using saves of only deeper skin tone content. 5) Clustered (User) standard errors in parentheses. Signif. Codes: **: 0.05, *: 0.1*

Finally, we estimate the treatment effects using only the first query for each user. As mentioned in the main text, there could be potential path dependence in the queries due to the treatment such that what a user sees at time $t + 1$ is influenced by what she clicks on at time t. Given that the new recommendation system changes what the user is exposed to initially, this could ultimately influence what the user gets exposed to subsequently. We check for robustness against this concern by only using the first query made by each user. We estimate the treatment effect on overall engagement and engagement with deeper skin tone content since these measures can be defined at a query level. The results are shown in Table E7 and are consistent with our main specification.

Table E7. Robustness check: ATE using only first query for each user

Dependent Variables: Model:	Overall Engagement (1)	Engagement with Deeper ST (2)
Variables		
Treatment	-0.0092	0.1584**
	(0.0111)	(0.0256)
Fit statistics		
Observations	638,683	638,683
Log-Likelihood	-264,319.3	-52,522.5

*1) The table shown regression results from estimating Model 6 using data from only the first query for each user. 2) Column 1 shows the treatment effect of inclusive recommendations on overall engagement and Column 2 shows the effect on engagement with deeper skin tone content. Both columns show estimates from a Quasi-Poisson regression and hence the treatment coefficient can be interpreted as an approximate percentage change. 3) We do not show results for Shannon entropy or engagement with all four skin tone buckets since they are defined at an aggregate level for a user and not for a single query. 4) Heteroskedasticity-robust standard errors in parentheses. Signif. Codes: **: 0.05, *: 0.1*

F TREATMENT EFFECT ON USER DISSATISFACTION

The platform provides a single-click option for users to express their dissatisfaction with the content they are being served. The option allows them *hide* the content and this signal is eventually fed back into the recommendation system's training routine. We estimate the treatment effect of diversified recommendations on the number of items hidden by each user. This allows us to gauge whether users are being exposed to irrelevant content in the pursuit of diversity. The results are shown in Table F8. We find that the number of hides has in fact slightly reduced in the treatment group, although the effect is not statistically significant. This evidence, at least in part, assures us that the new recommendation system is catering to user preferences.

Table F8. Average treatment effect on user dissatisfaction measured by items hidden

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

*Heteroskedasticity-robust standard errors in parentheses. Signif. Codes: **: 0.05, *: 0.1*

G HETEROGENEOUS TREATMENT EFFECTS ACROSS USER CHARACTERISTICS

We modify Model 7 from the main text to test for heterogeneous treatment effects across two key demographic variables – the user's self-reported gender and their country. We estimate the following regression for this purpose:

$$Y_{iq} = \alpha + \beta_1 T_i + \gamma X_i + \beta_2 T_i \times X_i + \epsilon_{iq} \tag{8}$$

where β_1 captures the average treatment effect and β_2 potential heterogeneity in the treatment effect. The results are shown in Table G9. We find no substantial difference in results across demographics.

Table G9. Heterogeneous treatment effects on engagement by consumer demographics

	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)	(5)	(6)	(7)	(8)
	Poisson	Poisson	Poisson	Poisson	OLS	OLS	Logit	Logit
Variables								
Treatment	0.0078	0.0012	0.2123**	0.4475**	0.0107**	0.0214**	0.4124**	0.4599*
	(0.0605)	(0.1559)	(0.0880)	(0.2243)	(0.0016)	(0.0095)	(0.0584)	(0.2683)
Treatment × Gender (F=1)	-0.0044		-0.0761		0.0055**		0.1534**	
	(0.0637)		(0.0942)		(0.0019)		(0.0651)	
Treatment × Country CA		0.0651		-0.0695		-0.0090		0.0271
		(0.1708)		(0.2581)		(0.0099)		(0.2841)
Treatment × Country GB-IE		-0.0230		-0.2570		-0.0044		0.0732
		(0.1625)		(0.2346)		(0.0098)		(0.2766)
Treatment × Country US		0.0005		-0.3262		-0.0065		0.0828
		(0.1576)		(0.2272)		(0.0095)		(0.2699)
Fit statistics								
Observations	638,683	638,683	638,683	638,683	638,683	638,683	638,683	638,683
Log-Likelihood	-170,486,419.5	-170,586,026.7	-249,081,086.3	-249,713,041.9	-213,506.1	-214,036.9	-35,968.5	-35,967.3

*Heteroskedasticity-robust standard errors in parentheses. Signif. Codes: **: 0.05, *: 0.1*

Further, we estimate the treatment effects for new vs. older platform users, where new users are those who joined the platform during the experiment. About 4% of the users in our sample are new. The results are shown in Table G10. The treatment effects are largely similar for new and old users.

Table G10. Treatment effects on engagement for new users

Dependent Variables:	Total Saves	Saves of Deeper ST Content	Shannon Entropy	Saved All 4 ST (Binary)
Model:	(1)	(2)	(3)	(4)
	Poisson	Poisson	OLS	Logit
Variables				
Treatment	0.0027	0.1526**	14.92**	0.5395**
	(0.0198)	(0.0321)	(0.8468)	(0.0262)
Treatment × New User (Binary)	0.0521	-0.1279	-9.162**	-0.1182
	(0.1366)	(0.1751)	(3.617)	(0.1545)
Fit statistics				
Observations	638,683	638,683	638,683	638,683
Log-Likelihood	-170,701,597.0	-249,795,741.2	-4,609,965.2	-35,985.4

*Heteroskedasticity-robust standard errors in parentheses. Signif. Codes: **: 0.05, *: 0.1*

H USER SEGMENTATION BASED ON PRE-TREATMENT EXPOSURE AND ENGAGEMENT

We segment users based on 3-month pre-treatment data. The segmentation buckets users based on their pre-treatment exposure to diverse skin tone content and engagement with deeper skin tone content.

Table H11. User classification based on pre-treatment exposure to and engagement with diverse content

Shannon Entropy (E)	Engagement with Deeper ST (R)	Users	Shannon Entropy Mean	Engagement Rate Overall*
Low	Low	422,807	1.40	0.64
Low	High	69,000	1.53	0.80
High	Low	69,015	1.75	0.64
High	High	54,038	1.79	0.79

**Overall engagement numbers are masked by multiplying with a common random number between 0 and 100.*

We test for heterogeneity in treatment effects for the user segments defined above. We hypothesize that, while overall engagement rates are stable, users with a stronger preference for deeper skin tone content are likely to benefit more from the diversified recommendation system. To test this, we run the following regression:

$$Y_{iq} = \alpha + \beta T_i + \gamma \mathbf{X}_i + \beta^H T_i \times \mathbf{X}_i + \epsilon_{iq} \tag{9}$$

where Y_{iq} is the total number of saves made by user i in query q, and $\mathbf{X}_i$ is the user segment that classifies user i into one of four buckets based on pre-treatment exposure diversity and engagement with deeper skin tone content. The segments are – 1) low exposure and low engagement (E:Low-R:Low), 2) low exposure and high engagement (E:Low-R:High), 3) high exposure and low engagement (E:high-R:Low), and 4) high exposure and high engagement (E:High-R:High). We are interested in the set of coefficients β^H in Model 9, which show how the treatment effect varies for these segments. The results are shown in Table H12. Standard errors are clustered at the user level.

We find that the treatment effect is indistinguishable from zero for 3 out of the 4 user segments. However, users with high pre-treatment exposure diversity and high engagement rates with deeper skin tone content show a strong and positive increment in engagement. This result lends credibility to our hypothesis that the diversified recommendation system is able to cater to these user's preferences by surfacing relevant content.

Table H12. Heterogeneous treatment effects on engagement for new users

Dependent Variables:	Total Saves	Saves of Deeper ST Content	Shannon Entropy	Saved All 4 ST (Binary)
Model:	(1)	(2)	(3)	(4)
	Poisson	Poisson	OLS	Logit
Variables				
Treatment	-0.0079	0.3850**	11.51**	0.7678**
	(0.0243)	(0.0305)	(0.8828)	(0.0490)
Treatment × E:Low-R:High	0.0366	-0.3576**	13.79**	-0.1676*
	(0.0542)	(0.0575)	(2.952)	(0.0878)
Treatment × E:High-R:Low	-0.0829	-0.3450**	5.099*	-0.3507**
	(0.0730)	(0.1106)	(2.956)	(0.0748)
Treatment × E:High-R:High	0.1224**	-0.2114**	17.61**	-0.3357**
	(0.0512)	(0.0618)	(4.310)	(0.0660)
Fit statistics				
Observations	614,860	614,860	614,860	614,860
Log-Likelihood	-160,618,466.0	-202,216,257.3	-4,437,037.6	-32,393.8

*Heteroskedasticity-robust standard errors in parentheses. Signif. Codes: **: 0.05, *: 0.1*

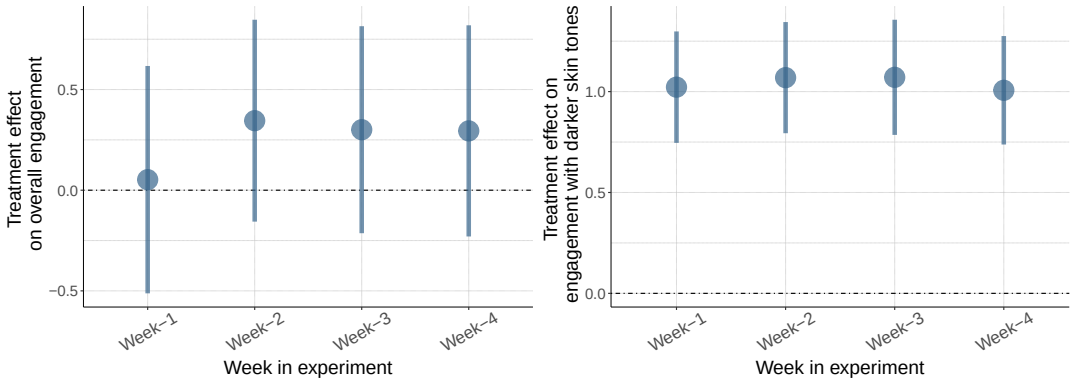
Table H13. Heterogeneous treatment effects on overall engagement across ventiles based on pre-treatment engagement with deeper skin tone content

Dependent Variable:	Total Saves	
	User Segments Based on	
	Historical Engagement	Historical Exposure
Model:	(1)	(2)
Variables		
Treatment	-0.0675	-0.0358
	(0.0469)	(0.0531)
Treatment × Pentile - 2	0.0328	0.0075
	(0.0581)	(0.0644)
Treatment × Pentile - 3	0.1057*	0.0603
	(0.0573)	(0.0698)
Treatment × Pentile - 4	0.0292	-0.0367
	(0.0652)	(0.0670)
Treatment × Pentile - 5	0.1345**	0.1527**
	(0.0554)	(0.0633)
Fit statistics		
Observations	614,860	614,860
Log-Likelihood	-155,855,903.4	-158,780,732.2

*Heteroskedasticity-robust standard errors in parentheses. Signif. Codes: **: 0.05, *: 0.1*

I ROBUSTNESS CHECKS AGAINST NOVELTY EFFECTS

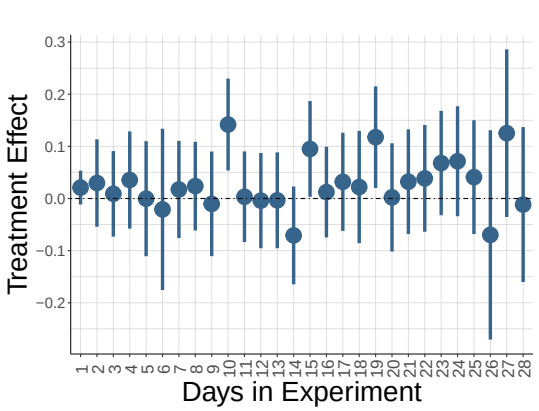
We check the robustness of our results against potential novelty effects [Kohavi et al., 2020]. We leverage query-level data for this task and re-estimate our main results using two approaches. First, we split the data into 4 parts, one for each week for which the experiment was run. We use the timestamp associated with each query to assign it to its corresponding "experiment week". We then estimate the average treatment effect on each week of data separately. The results are shown in Figure I5. Each point corresponds to the ATE estimate for each week and the bars represent 95% confidence intervals. We find that the effects are stable and consistent during the entire experiment period. Note that we only estimate the treatment effects for overall engagement and engagement with deeper skin tone content. This is because Shannon entropy and engagement with all four skin tone buckets are user-level outcomes and are not defined for an individual query.



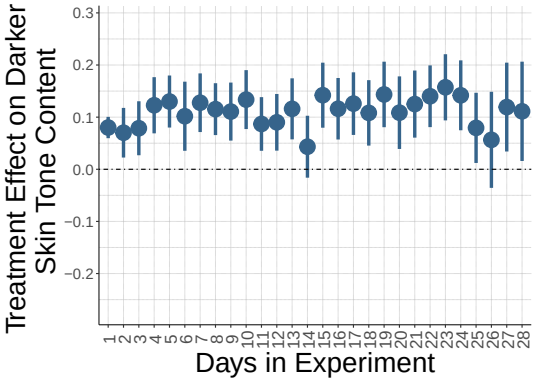
(a) Overall engagement (b) Engagement with deeper skin tone content

Fig. I5. Treatment effects during each week of the experiment

In our second approach, we segment user-query pairs by the number of days a user has been triggered into the experiment. This allows us to directly account for the amount of time a user has spent in the experiment and investigate whether there is any attenuation of the effect over time. The results are shown in Figure I6. The left panel shows estimates for overall engagement and the right panel shows estimates for engagement with deeper skin tone content. There is no evidence of a systematic increase or decrease in treatment effects over time.



(a) Overall engagement



(b) Engagement with deeper skin tone content (normalized by exposure)

Fig. I6. Treatment effects split by the number of days a user spent in the experiment