Optimizing Multi-Stage Personalization in the Customer Journey

Yu Song*

March 2026

Abstract

Firms increasingly apply personalization to influence product discovery and engagement along the customer journey. However, effective multi-stage personalization is challenging because customers' later-stage decisions depend on earlier experiences, creating cross-stage spillovers. I study personalization at two stages on an e-commerce platform: an early, firm-driven discovery stage and a later, customer-initiated query stage. Using two field experiments, I find that personalizing the discovery stage increases immediate engagement but reduces engagement in the query stage, producing no net gain. When the query stage is also personalized, total revenue rises without cannibalizing discovery-stage revenue. Personalization has stronger effects for customers with more consistent preferences. To identify the revenue-maximizing policy across stages and customer types, I build and estimate a structural search model that uses experimental variation for identification and a neural-network estimator for computational tractability. The results show that personalizing only at the query stage maximizes total revenue. Discovery-stage personalization shapes beliefs about the value of querying and can trigger early exit when matches are weak, whereas query-stage personalization conditions on query-revealed signals to deliver more relevant results. Moreover, targeting personalization to customers with more consistent preferences further increases revenue. These findings offer firms guidance on optimizing personalization across stages and customers.

Keywords: Multi-Stage Personalization, Personalized Recommendations, Personalized Search Ranking, Field Experiments, Consumer Search, Customer Journey

*I am a Ph.D. candidate in Marketing at the Stephen M. Ross School of Business, University of Michigan. Email: yyusong@umich.edu. I am deeply grateful to Jessica Fong and Puneet Manchanda for their invaluable mentorship and guidance. I thank Anocha Aribarg, Zach Brown, Jun Li, Yeşim Orhun, S. Sriram, Candice Wang, and Qingliang Wang for their helpful feedback and suggestions. I also thank the Mercari team, especially Ajay Daptardar, Michael Manzon and Kumon Takuma Yamaguchi, for their tremendous support in implementing the experiment and for the many insightful conversations. The views expressed in this paper are solely my own and do not necessarily reflect those of Mercari, Inc.

1 Introduction

As digital markets mature, personalization has become a central tool influencing how customers discover and engage with products. Firms that excel at personalization generate 40% more revenue than their peers.[1] Platforms such as Amazon and Spotify attribute 35% of transactions and a large share of a 1,000% increase in revenue, respectively, to personalized recommendation systems (MacKenzie et al., 2013; Abraham and Edelman, 2024). Given its potential, firms such as Netflix and eBay have begun to apply personalization to multiple stages of the customer journey (Ostuni et al., 2023).[2] In many digital settings, customers begin their journey by discovering recommendations on a landing page, and then decide whether to initiate search queries to retrieve additional products relevant to their queries. These two stages, which I refer to as the discovery stage and the query stage, are prevalent across various platforms, including Amazon and eBay in e-commerce, TikTok and YouTube in video sharing, and Uber Eats and DoorDash in food delivery.

Many firms manage each stage of the customer journey in isolation, seeking to maximize returns within a single stage. Organizational design and measurement systems can reinforce this siloed approach by prioritizing stage-level metrics over journey-level outcomes.[3] However, this approach can overlook spillovers across stages and lead to suboptimal outcomes. The discovery stage is particularly salient because it sets the trajectory for the remainder of the journey. On the one hand, personalizing the discovery stage may help customers find relevant products quickly, freeing attention and time for broader exploration later in the journey (Song, 2021). On the other hand, if discovery surfaces satisfactory options too early, it may weaken incentives to engage in deeper search (Fong, 2017; Chen et al., 2025), potentially lowering total revenue (Yuan et al., 2025). Moreover, when discovery fails to produce a good match (for example, because customer preferences shift across sessions), customers may interpret the discovery results as evidence that the platform does not offer relevant products and exit without purchasing or continuing the journey.

[1] McKinsey & Company (2022), "The Value of Getting Personalization Right—or Wrong—is Multiplying," `https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/the-value-of-getting-personalization-right-or-wrong-is-multiplying`, accessed March 2026.

[2] Modern Retail (2023), "eBay Steps Up Personalization Efforts to Improve Search, Ads and Experience," `https://www.modernretail.co/retailers/ebay-steps-up-personalization-efforts-to-improve-search-ads-and-experience/`, accessed March 2026.

[3] McKinsey & Company (2016), "From touchpoints to journeys: Seeing the world as customers do," `https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/from-touchpoints-to-journeys-seeing-the-world-as-customers-do?`, accessed March 2026.

In this paper, I study how personalization affects customer search and whether stage-by-stage optimization is suboptimal when cross-stage spillovers are present. Specifically, I aim to answer two research questions. First, how do the effects of personalization vary across different stages of the customer journey? I focus on its effects on customer behavior (e.g., clicks and purchases) and platform revenue. Second, how do these effects differ across customers? If personalization is effective for some customers but generic (e.g., popularity-based) experiences work better for others, firms may be able to leverage this heterogeneity to "personalize personalization"—that is, to target personalization efforts toward those most likely to benefit while leaving others with a more generic experience.

To study these questions, I collaborate with Mercari, Inc. (Mercari), an e-commerce platform that enables individuals to buy and sell items across a wide range of categories. Founded in Japan, it has become Japan's largest community-powered marketplace and expanded its operations to the United States (US) in 2014. This paper focuses on Mercari US. I first analyze a large-scale field experiment that only personalizes the discovery stage. I find that the personalized discovery stage increases user engagement (clicks, orders, and revenue) within that stage but reduces the likelihood that customers proceed to the subsequent query stage. These findings indicate that personalizing discovery can generate negative spillovers on downstream search. When the personalized discovery stage offers a good match, customers purchase immediately and thus do not proceed to the query stage. When it offers a poor match, customers may interpret the weak match as evidence that the platform lacks relevant products, infer a low value of querying, and exit early without purchasing.

I then design and implement a subsequent field experiment that personalizes the query stage.[4] I develop two versions of a ranking algorithm: a non-personalized version that uses only item-level features (e.g., brand) and contextual features (e.g., inventory of similar items at the time of the query), and a personalized version that additionally incorporates user-level features (e.g., past purchases in the last 30 days). To ensure that outcome differences are driven by personalization rather than by differences in algorithms, I apply the same ranking algorithm, LambdaMART, to both conditions. LambdaMART operates by using many small decision trees to compare item pairs

[4] All users in this experiment receive a personalized discovery stage. A full factorial design would allow me to identify the optimal way to personalize across stages, but it is often difficult to implement in practice. The company faces operational constraints, especially when the two stages are managed by different teams.

and learn how to order the query results. Deploying LambdaMART-based ranking models in a field experiment on Mercari, I find that users exposed to personalized product rankings at the query stage click on more items and make more purchases, with no spillover effects on the discovery stage. As a result, total revenue across both stages increases.

Across both experiments, personalization is more effective for customers with more consistent preferences (who tend to click similar items over time). By contrast, discovery-stage personalization reduces revenue for customers with more dynamic preferences (who click across a broader set of products). Together, these results suggest that firms may "personalize personalization" to increase total revenue. For example, they could intensify personalization for customers with more consistent preferences while offering a more generic experience for others.

While my two experiments offer valuable insights, they cannot evaluate all combinations of multi-stage personalization (e.g., comparing personalization only at the query stage versus personalizing at both stages), nor can they assess alternative strategies such as personalized personalization. To fill these gaps, I develop a structural search model building on Weitzman's (1979) optimal sequential search framework. The model captures customers' inspection (i.e., clicking) and purchase decisions as they transition from an initial discovery stage to a subsequent query stage. In the discovery stage, customers observe limited product information. They can choose whether to inspect an item by clicking to reveal additional product information or to submit a query to view a new set of products in the query stage; both actions incur costs. Customers use the products they see in the discovery stage to form beliefs about the value of querying, and they place more weight on this signal when the discovery set aligns with their purchase intent. If the customer conducts a query, they switch from the discovery stage to the query stage. At the query stage, the customer can similarly incur a cost to inspect an item. The customer can terminate the process at any point, either by selecting the outside option or by purchasing one of the inspected products.

I exploit experimental variation in customers' propensity to transition to the query stage to estimate the structural search model. I apply the neural network estimator (NNE) developed by Wei and Jiang (2025), which trains a neural network to learn the mapping from observed data moments to the model's structural parameters. This approach avoids direct integration over unobservables and overcomes computational challenges.

Using the estimated model, I conduct counterfactual simulations to evaluate a range of multi-

stage personalization strategies. In the first counterfactual simulation, I consider a utility-based ranking framework (Ursu, 2018). Using the same framework at both stages allows me to isolate the role of personalization from differences in the underlying ranking algorithms across stages. The results show that although personalization generally increases platform revenue, its effectiveness depends on where it is applied. Applying personalization to more stages does not necessarily improve total outcomes. In fact, personalizing only the query stage maximizes total revenue: query-only personalization generates 2.9% higher revenue than personalizing both stages and 12% higher revenue than personalizing only the discovery stage. Discovery-stage personalization surfaces items that are more aligned with a customer's purchase intent, which makes customers more likely to treat the discovery experience as informative when forming beliefs about the value of querying. However, because the discovery set is limited and may reflect preferences imperfectly, a weak preference fit can lead customers to update downward their expected gains from querying, causing some to exit early rather than continue the journey. By contrast, a popularity-based (non-personalized) discovery stage matches intent less precisely, so customers place less weight on it when forming beliefs, which attenuates these negative continuation effects. Furthermore, the query stage reveals richer information about the customer's current intent and preferences (e.g., brand). This information allows the ranking to condition on that information and deliver more relevant results with a higher likelihood of satisfying demand. In the second counterfactual simulation, I evaluate "personalizing personalization" based on customers' preference consistency. I find that applying personalization only to customers with consistent preferences increases platform revenue by 14.7% relative to no personalization and by 3.3% relative to applying it to all customers.

This paper offers several managerial implications. First, as firms increasingly coordinate personalization across multiple stages of the customer journey, they should account for both within-stage effects and cross-stage spillovers. Given evidence of negative spillovers from the discovery stage to the query stage, myopic personalization that optimizes each stage in isolation can be suboptimal. Managers should therefore evaluate personalization decisions jointly across stages, weighing incremental benefits against implementation costs when deciding whether, and at which stages, to personalize. A holistic approach can also help preserve product diversity. For example, keeping the discovery stage less personalized can expose customers to a broader assortment, which may mitigate concerns about "filter bubbles," where highly tailored content limits exposure to new or

diverse products (Pariser, 2011). Second, firms can tailor personalization intensity at the individual level. A simple starting point is to reduce or shut down personalization for some customers. For instance, platforms might provide more tailored recommendations to customers with consistent preferences while offering a broader, more generic experience to customers whose preferences are more dynamic. Finally, although I study a single platform, the challenge I document is broadly relevant: stage-level metrics can obscure cross-stage trade-offs. The empirical approach in this paper provides a practical template for identifying whether such spillovers arise in other settings and for quantifying the net value of personalization across the full journey. In this sense, the paper offers a foundation for designing experiments and decision rules that optimize personalization end to end rather than locally within individual stages. In practice, firms can apply the same logic by designing experiments that measure both direct effects and downstream outcomes. This enables them to detect spillovers, quantify net effects on engagement and revenue, and choose personalization policies that optimize performance at the journey level.

The remainder of the paper is structured as follows. Section 2 situates my work within the existing literature. Section 3 describes the empirical setting and data. Section 4 details two field experiments and shows the experimental results. Section 5 presents the structural search model and outlines the identification and estimation procedures. Section 6 reports the search model estimation results and counterfactual simulations. Section 7 concludes with implications and limitations.

2 Literature Review

This paper relates closely to the literature on personalized recommendations and personalized search ranking. Personalized recommendations at the discovery stage have been shown to increase consumer engagement (e.g., Hosanagar et al., 2014) and purchase propensity (e.g., Li et al., 2022). However, they may also reduce consumption diversity and diminish long-term user engagement (Holtz et al., 2020; Chen et al., 2024). Personalized search ranking has been found to improve platform outcomes and consumer welfare, even when platforms take revenue into account (Donnelly et al., 2024). While prior work focuses on a single stage of the customer journey, this paper is, to the best of my knowledge, the first to examine the combined effects of personalization across multiple stages.

My paper also relates to empirical work on the interplay between firm recommendations and consumer search. Song (2021) documents a positive spillover effect: personalized recommendations improve search efficiency in the essential product category and stimulate broader exploration in other categories. Wan et al. (2024) find that recommendations help consumers identify higher-value products, through lower prices or better preference alignment. In contrast, Yuan et al. (2025) show that when product discovery relevance declines, consumers compensate by increasing their search activity. Fang et al. (2026) demonstrate that the relationship between recommender and non-recommender searches depends on the number of information cues provided. My paper adds to the literature by examining multi-stage personalization using a consumer search model.

This paper contributes to the growing literature on the customer journey. Prior research has employed a wide range of approaches to modeling the customer journey, including probabilistic machine learning techniques (Padilla et al., 2024), Poisson point process models (Goić et al., 2022), and transformer-based architectures (Lu and Kannan, 2026). My approach is most closely related to work that employs structural search models to analyze consumer decision-making. Existing research has examined how consumers progress from awareness to consideration and choice (Honka et al., 2017; Greminger, 2022), explored how they navigate product discovery pathways (Zhang et al., 2023), and investigated post-search behaviors such as revisits (Dang et al., 2025) and retargeting responses (Jiang et al., 2021). I develop a sequential search model to understand the impact of multi-stage personalization on consumer decisions.

Furthermore, my work adds to the literature on the empirical search model (Honka et al., 2019; Honka et al., 2024). There is growing empirical search literature built on the tractable solution offered by Weitzman (1979), such as modeling how consumers use refinement (Chen and Yao, 2017; De los Santos and Koulayev, 2017; Gu and Wang, 2022; Wang and Liu, 2026), incorporating consumer learning (Hodgson and Lewis, 2025; Korganbekova and Zuber, 2023), considering browsing behaviors (Greminger, 2022; Choi and Mela, 2019; Zhang et al., 2023), modeling which product attributes customers search over (Compiani et al., 2024), and so on. Search models typically assume that the decision to conduct a search is exogenous, but this assumption may not hold when prior experience with landing-page recommendations influences that decision. I contribute to the literature by endogenizing consumers' decisions to conduct search queries. Modeling these decisions is important when firms seek to coordinate consumers' active search (initiating a query) with passive

search (exposure to firm-generated recommendations).

3 Empirical Setting and Data

3.1 Empirical Setting

The empirical setting of this paper is Mercari US. As of 2024, it has 5 million monthly active users, 0.6 million sellers, 11 million new listings per month, and over 100 million app downloads.[5] Users can purchase new or used items from a wide range of categories (e.g., women's fashion, men's fashion, office supplies, home goods, etc.). Mercari emphasizes ease of use through its app-based platform, positioning itself as "a C2C marketplace app that allows anyone with a smartphone to easily sell items they no longer need."[6] This paper focuses on the Mercari app, including the field experiments and data.

In Figure I.1 in the web appendix, I illustrate the platform layout with examples of Mercari's discovery page, query results page, and product page. When users enter the app, they first see product recommendations on the discovery page, which I refer to as the discovery stage. This stage is pivotal because it receives 100% of user traffic. On average, customers spend over one minute in the discovery stage before entering the query stage. This suggests that customers engage with both stages of the journey, making each an important component.[7] At a high level, the platform constructs discovery-stage discoveries based on popularity, trendiness, and, when personalization is applied, the user's historical activity. Section 4.1 provides a detailed description of the discovery-stage algorithm used in this study.

Users can conduct a query; when they do, the platform returns a ranked list of results relevant to the query.[8] The order from the discovery stage to the query stage is mechanically fixed. Mercari employs a three-stage ranking process to generate query results, following a standard industry pipeline documented in the academic literature (e.g., Zhan et al., 2024; Chen et al., 2024). The

[5] Sources: `https://about.mercari.com/en/press/news/articles/20230201_mercari10th/`; `https://thesmallbusinessblog.net/mercari-users/`; `https://www.mercari.com/about/`. All accessed March 2026.

[6] Source: `https://careers.mercari.com/en/services/`, accessed March 2026.

[7] Users also see "similar items" at the bottom of the product detail page after clicking on a product. Due to scope, I do not study this recommendation surface and leave it for future research.

[8] Users may also conduct non-query-based searches by choosing product categories. In such cases, item rankings follow the platform's default algorithm and are unaffected by my experiment. As the majority of searches on the platform are query-based, this paper focuses exclusively on query-based search.

process consists of: (1) a retrieval stage, where the platform retrieves a broad set of candidate products based on the query, typically ranging from 400 to 10,000 items depending on relevant product availability; (2) a ranking stage, where items are primarily sorted by relevance and recency, with more relevant or recent items appearing higher in the list; and (3) a re-ranking stage, where the top 100 items from the ranking stage are re-ranked using a machine learning model. Section 4.2 provides further details on how products are re-ranked in this study.

3.2 Data Description

For each user session, I observe the full sequence of consumer actions in both the discovery and query stages, along with their timestamps. These actions include (1) conducting a query, (2) inspecting a product, as measured by the user clicking on the product, (3) making a purchase, and (4) exiting the platform. In both stages, I observe the set of displayed products and their positions.

I observe each user's historical activity on the platform, including product clicks, purchases, and search queries. For each product, I record attributes such as brand, category, physical condition, item age, number of photos, price, and seller attributes. I also track downstream selling outcomes, including whether the product is sold and, if so, the sale date, transaction price, and buyer.

4 Personalization Field Experiments

In this section, I present two field experiments. The first experiment (Section 4.1) varies whether the discovery stage is personalized, while the second experiment (Section 4.2) varies whether the query stage is personalized. Section 4.3 summarizes the experimental results and motivates the structural search model.

As users may self-select into future sessions based on their first session's experience, I focus on each user's first session during the experiment to mitigate selection bias, following the approach of Fong et al. (2026) and Fong (2024). User-level outcomes across all sessions are similar and are reported in Tables I.1 and I.2 in the web appendix. Both experiments pass the randomization check (see Web Appendix A.1) and manipulation check (see Web Appendix A.2). Finally, I rule out the potential stable unit treatment value assumption (SUTVA) violations using several checks in Web Appendix A.3.

4.1 Personalizing Discovery Stage

The personalized discovery stage experiment was conducted for two weeks in Fall 2022. The experiment was implemented on the Mercari mobile app; users accessing the platform via the web continued to receive non-personalized discovery stage. My analyses focus on user activity on the Mercari app. In this experiment, 50% of users were assigned to the control condition that received a non-personalized discovery stage, and 50% of users were assigned to the treatment condition that received a personalized discovery stage. In the non-personalized condition, the platform recommended products based on product popularity and trendiness. Product popularity was measured by the cumulative historical revenue of its category-brand pair, and trendiness was defined as the maximum week-over-week revenue increase for that same pair.[9] In the personalized condition, the platform used a neural-network architecture. Specifically, it constructed a latent preference profile for each user based on their recent activity, such as search queries and item clicks. Products the user interacted with were simultaneously embedded into the same latent space using item attributes such as category and title. User and item representations were trained jointly so that items similar to those a user had previously engaged with were positioned closer to their profile.

For new or inactive users (i.e., those without activity in the past 6 months), the platform applied a cold-start (non-personalized, popularity-based) algorithm to generate product discoveries. Thus, the estimated effects should be interpreted as intent-to-treat.

4.1.1 Summary Statistics

During the experimental period, 3,050,146 customers viewed the discovery-stage recommendations. 19.5% of customers clicked on at least one recommended item at the discovery page; among them, the median number of clicks was 1, and the mean was 2.2. 0.9% of customers made a purchase from the recommended products.[10] Customers spent an average of 72 seconds in the discovery stage before initiating a query. This pattern suggests that the discovery stage is a meaningful part of the customer journey, because many users engage with the discovery page rather than immediately

[9] As products on the platform are single-unit, popularity and trendiness are measured at the category–brand-pair level.

[10] A product purchase is attributed to the current session if the customer eventually completes the transaction for that product. This practice is consistent with prior work (e.g., Donnelly et al., 2024; Korganbekova and Zuber, 2023). I adopt this definition of purchase throughout the paper.

submitting a query. 56.6% of customers initiated a search query. Conditional on viewing query results, customers clicked on a median and mean of 3 and 5.6 items, and 2.8% of them made a purchase from the query results.[11]

4.1.2 Treatment Effects of Personalizing Discovery Stage

Because users were randomly assigned to either the personalized or non-personalized discovery-stage condition, differences in outcomes between the two conditions identify the causal effect of personalizing the discovery stage. I examine the number of clicks, orders, and revenue in both the discovery and query stages, as well as whether the user initiates a query.

Table 1 reports the treatment effects. Personalizing products at the discovery stage increases discovery-stage engagement: users click on 12.1% more items, place 8.4% more orders, and generate 8.2% more revenue compared to the non-personalized condition. However, there is a reduction in engagement in the subsequent query stage. Users exposed to personalized product discoveries are 4.2% less likely to conduct a search query, click on 4.7% fewer items, place 3.5% fewer orders, and generate 4.1% less revenue from the query stage. I further establish the substitution effect between product recommendations at the discovery stage and consumer search by leveraging experimental variation as an instrumental variable in Web Appendix B.1.

Since the query stage presents more relevant products, has a higher purchase rate and contributes a larger share of total revenue, even a modest percentage drop in its revenue offsets the revenue gains from personalizing the discovery stage. As a result, total orders and total revenue (combining both stages) do not differ significantly between experimental conditions.

I next examine heterogeneity in the effects of personalizing the discovery stage. Because preferences can evolve over time (Yoon and Simonson, 2008), the extent to which past behavior predicts preferences in the current session likely varies across users, which in turn affects personalization performance. I test whether the benefit of personalized discovery differs for customers with relatively consistent versus dynamic preferences. As a proxy for preference consistency, I use clicking behavior in the 30 days prior to the experiment and compute the average number of unique brands and price levels clicked in each category. Higher values indicate more diverse clicking and thus

[11] I only consider the top 100 items in the query results, as items ranked lower than 100 are rarely clicked.

Table 1: Treatment Effects of Personalizing Discovery Stage

Variable	Non-Personalized	Personalized	Change (%)	t-statistic	p-value
Discovery Clicks	0.404	0.453	12.135	29.018	< 0.001
Discovery Orders	0.009	0.010	8.380	6.629	< 0.001
Discovery Revenue	0.379	0.411	8.222	4.341	< 0.001
I(Initiate a Query)	0.578	0.554	-4.160	-42.344	< 0.001
Query Clicks	3.243	3.090	-4.717	-22.865	< 0.001
Query Orders	0.018	0.017	-3.517	-3.733	< 0.001
Query Revenue	0.576	0.552	-4.055	-2.361	0.018
Total Clicks	3.647	3.543	-2.850	-15.123	< 0.001
Total Orders	0.027	0.027	0.621	0.813	0.416
Total Revenue	0.955	0.963	0.821	0.638	0.524

Notes. This table shows the treatment effects of personalizing the discovery stage. 1,528,471 (50%) users are in the personalized condition. I report the mean of outcomes in both conditions. The difference is measured using the outcomes in the personalized condition minus the outcomes in the non-personalized condition. The last two columns report t-statistics and p-values from the corresponding two-tailed tests for differences in means.

more dynamic preferences. I classify customers as dynamic if this measure is above the median.[12]

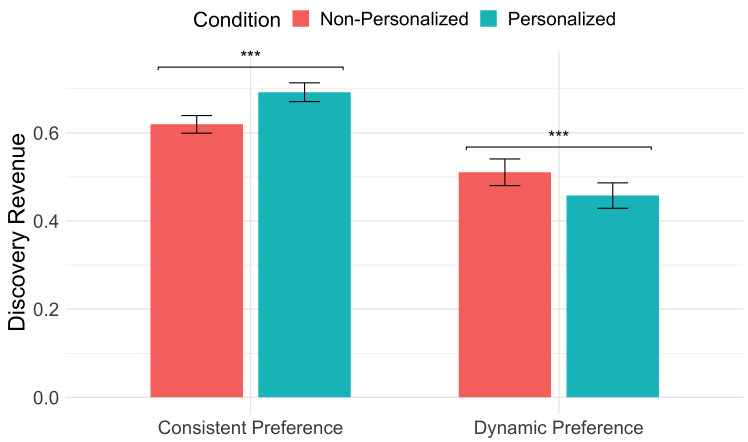
As shown in Figure 1, discovery-stage personalization is more effective for customers with consistent preferences, whereas the non-personalized discovery stage performs better for customers who tend to click a broader range of products. This heterogeneity suggests that firms could tailor personalization to users' preference consistency. For example, firms could personalize the discovery stage only for customers with consistent preferences while keeping broader exposure for customers with more dynamic preferences. Because this "personalized personalization" approach is not tested in the experiment, I evaluate its incremental value using counterfactual simulations.

4.1.3 Additional Evidence and Discussion

Discovery versus Query Stages Personalization can work differently in the discovery and query stages because the firm's information set and the customer's decision context differ across the two. In the discovery stage, the firm mainly relies on a user's historical activity to curate an assortment of items it expects the user will like, and users are passively exposed to these recommendations. In the query stage, the firm observes not only past behavior but also a contemporaneous signal of intent and preference through the submitted query, so personalization operates within a narrower, query-relevant set of items. Additionally, users actively initiate the query.

[12] The results are robust to alternative definitions of the measure and to different cutoff thresholds.

Figure 1: Heterogeneous Effects of Personalizing Discovery Stage



Notes. This figure plots the revenue at the discovery stage for customers in the personalized and non-personalized conditions, separately for those with consistent and dynamic preferences. The sample is restricted to customers who clicked more than one item in the 30 days prior to the experiment. Error bars represent 95% confidence intervals.

To characterize these stages, I summarize descriptive evidence on the products shown at each stage along three dimensions: price, item condition, and product diversity. As shown in Web Appendix B.2.1, discovery-stage items are more likely to be in new condition and have slightly higher average prices. The discovery stage also exposes users to a broader set of products than the query stage (see Web Appendix B.2.2 for details).

Composition of Customers Customers on Mercari may arrive with a fairly well-defined purchase intent (demand fulfillment) or may browse to discover what they like (demand formation). Although I do not directly observe a user's underlying intent, I present suggestive evidence that behavior in my setting is largely demand fulfillment. First, applying the framework of Yuan et al. (2025), I find that personalized discovery-stage recommendations lead users to submit less specific queries (Web Appendix B.3.1). Following Yuan et al. (2025), this pattern is consistent with users arriving with an existing demand and using discovery as a substitute for query refinement.

Second, I test whether discovery-stage exposure shapes subsequent query behavior by measuring the semantic alignment between the session's query and the products shown in the discovery stage (Web Appendix B.3.2). If users mainly come to be inspired by broad, non-personalized discoveries, alignment should be higher in the non-personalized condition. Instead, I find that alignment is higher under personalization (Web Appendix B.3.2).

Third, internal interviews at Mercari indicate that many users treat the platform as a destination for targeted shopping rather than purely open-ended browsing. Taken together, these patterns suggest that customers in my setting often arrive with specific intent, which helps rationalize the observed substitution between discovery and query behavior.[13]

Why Discovery-Stage Personalization May Not Improve Total Outcomes The discovery stage at the beginning of the journey may shift users' beliefs about the continuation value of querying. After observing the discovery list, users may update their expectations about how likely the platform is to surface attractive alternatives. For instance, if users view a personalized discovery list as informative about what the platform can find for them, then a close but unsatisfactory match may lower the perceived gains from submitting a query, reducing query initiation and increasing early exit. To probe this possibility, I stratify users by whether they make a purchase at the discovery stage and compare outcomes between treatment and control within each stratum (see Web Appendix B.4). Among users who purchase at the discovery stage, those in the personalized condition are 4.2% less likely to initiate a query, with no significant change in total revenue, consistent with substitution between discovery and query stages. More importantly, among users who do not purchase at the discovery stage, those in the personalized condition are also 4.1% less likely to initiate a query and spend less overall across both stages.[14] These results suggest that discovery-stage personalization can harm a subset of users, potentially by discouraging further query after an unsatisfactory initial set of recommendations.

4.2 Personalizing Query Stage

Recall that personalization occurs in the re-ranking stage, where the top 100 products identified by the ranking stage are re-ordered. I first describe the ranking model, LambdaMART, used in this re-ranking stage. LambdaMART (Burges, 2010) is a widely used learning-to-rank algorithm that

[13] Future work could measure customer composition more directly, for example, by pairing experiments with survey-based intent elicitation or by leveraging richer session-level signals of goal-directed versus exploratory browsing.

[14] There are several reasons why the platform may fail to deliver a good match at the discovery stage. For instance, one possibility is that the personalization algorithm is imperfect. However, I find similar patterns in a separate field experiment that uses a more advanced personalization approach, suggesting that algorithm improvements alone may not eliminate this channel. Another possibility is that users' preferences vary across sessions (see Section 4.1.2), which reduces the predictive value of historical activity at the moment a user arrives.

optimizes item ordering by minimizing a ranking loss. It combines gradient-boosted decision trees with a pairwise loss. LambdaMART has been widely used in practice (e.g., Microsoft's Bing search engine) and has shown strong empirical performance (Yoganarasimhan, 2020; Kaye, 2024). In my experiments, I use LambdaMART for both the personalized and non-personalized rankings. The two models are trained on the same set of product and contextual features, and the personalized model additionally includes user-specific features. Item features include product attributes (e.g., brand) and past engagement measures (e.g., item likes in the past 30 days). Contextual features capture market conditions at the time the query is submitted (e.g., real-time inventory levels). User features include recent activity over the past 30, 60, and 365 days, inferred preferences for brand, price, and category, past purchases, and other user-level signals (e.g., ratings and tenure). Further details on ranking models are in Web Appendix C. Since I hold the ranking algorithm fixed and vary only the inclusion of user features, any differences between the personalized and non-personalized conditions can be attributed to personalization features rather than to the algorithm itself.

The personalized query stage field experiment lasted for 12 days in Fall 2024. I randomly assigned 10% of users to the personalized query condition and another 10% to the non-personalized query condition. The remaining 70% of users received the status quo non-personalized ranking and 10% of users received the personalized version of the status quo ranking (i.e., the same model augmented with individual features).[15]

4.2.1 Summary Statistics

I record user activity on the Mercari app during the experiment.[16] The experiment included 360,679 individuals, of whom 187,818 (52.1%) conducted a search query. Among those who entered the query stage, users clicked 4.2 items on average (with a median of 3), and 4.5% made a purchase from the items in the query results.

To validate that the treatment condition indeed produces personalized rankings at the query

[15] I use LambdaMART with and without user features as my experimental conditions, because the only distinction between these two conditions is the inclusion of user-level inputs. While the remaining conditions also compare personalized and non-personalized rankings, they employ different training approaches, making their results less straightforward to interpret. I use these remaining 80% of users to examine the robustness of personalization effects under an alternative ranking algorithm in Web Appendix E.

[16] Users were assigned to treatment conditions at the user level rather than the device level. Thus, a user assigned to the personalized condition received personalized rankings on both the app and the web (including mobile web).

stage, I examine the top 12 items displayed on the initial screen of the query results. Using price preference as an illustrative example, Figure 2 plots the price distribution of these items separately for customers inferred to prefer high versus low prices based on past purchases.[17] The figure shows that, under personalization, users with a high-price preference are more likely to see higher-priced items, while users with a low-price preference are shown lower-priced options. At the same time, the average item price is similar across the two experimental conditions. In addition to price preferences, Web Appendix A.2 shows that the personalized condition also increases the share of top-ranked items that match users' inferred brand preferences.

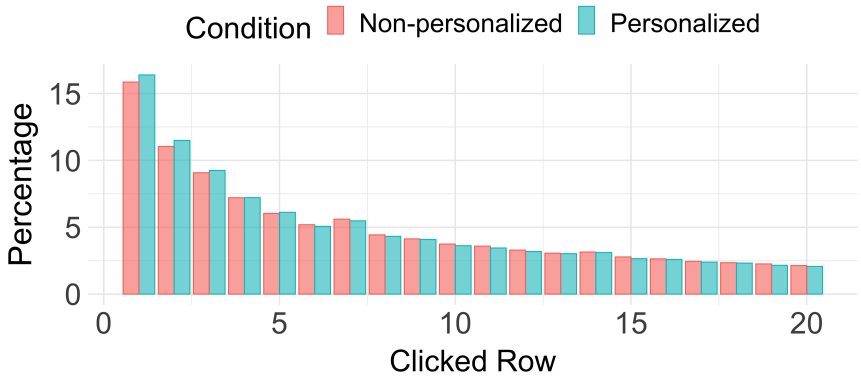
Figure 2: Price Distribution of Products at Query Stage by Customer Price Preference

Notes. This figure shows the price distribution of products displayed on the initial screen at the query stage (top 12 items), segmented by experimental condition ("personalized" vs. "non-personalized") and customers' price preferences inferred from past purchase histories ("high" vs. "low").

Figure 3 presents the distribution of clicks across item positions by experimental condition. Items ranked higher in the list receive more clicks, and this pattern is more pronounced under the personalized query-stage condition. This suggests that personalization effectively elevates preferred products to more prominent positions.

[17] For example, I classify a user as preferring high-priced products if the maximum price among their past purchases exceeds $50. I choose this threshold based on historical data.

Figure 3: Distribution of Clicked Positions



Notes. This figure plots the distribution of positions of clicked items across experimental conditions, focusing on the top 20 rows of the query results. Each row contains three items. The x-axis represents the row number, and the y-axis represents the percentage of total clicks.

4.2.2 Treatment Effects of Personalizing Query Stage

I present the treatment effect of personalizing the query stage.[18] I examine how personalizing the query stage influences customer engagement and purchases by measuring item clicks, orders, and revenue. I find in Table 2 that customers in the personalized condition click on 3% more items, make 9.2% more purchases, and generate 10.6% higher revenue. Table I.3 in the web appendix shows no changes in discovery-stage outcomes, while aggregate outcomes across the two stages are higher in the personalized condition.

Table 2: Treatment Effects of Personalizing Query Stage

Variable	Non-Personalized	Personalized	Change (%)	t-statistic	p-value
Clicks	2.240	2.307	2.995	4.867	< 0.001
Orders	0.025	0.028	9.165	3.882	< 0.001
Revenue	0.838	0.927	10.622	2.870	0.004

Notes. This table shows the treatment effects of the personalizing query stage on query-stage outcomes. There were 360,679 customers in the experimental conditions, and 180,438 (50%) customers in the personalized condition. I report the mean of outcomes in both personalized and non-personalized conditions. The difference is measured using the outcomes in the personalized condition minus the outcomes in the non-personalized condition. The last two columns report t-statistics and p-values from the corresponding two-tailed tests for differences in means.

However, personalization can also impose costs on customers. Because it requires additional computation, it can increase search latency. Search latency is a critical concern for online platforms

[18] Users were randomly assigned to an experimental condition upon arriving at the platform. By including users who did and did not submit a search query, I estimate the overall impact of later-stage personalization under an intent-to-treat framework.

because even small delays can degrade user experience and reduce key performance indicators.[19] Reflecting this concern, many companies have prioritized latency reduction, including Netflix (Ostuni et al., 2023), Airbnb[20] and Booking.com.[21] In Web Appendix D, I present experimental evidence that average search latency is higher in the personalized condition (274 milliseconds) than in the non-personalized condition (259 milliseconds). Moreover, a 1% increase in search latency reduces customer clicks, orders, and revenue by 4.6%, 0.4%, and 1.2%, respectively.

I also examine how the effectiveness of the personalized query stage varies with customers' preference consistency. Using the same procedure outlined in Section 4.1.2, I present the results in Figure I.2 in the web appendix. The findings indicate that personalized query rankings have a stronger positive effect for customers with more consistent preferences (i.e., those who tended to click on similar items prior to the experiment). In contrast, there is no significant effect for customers with more dynamic preferences. Given the potential latency costs associated with personalization, the non-personalized query stage might work better for these customers.

4.3 Experimental Results Summary and Structural Model Motivation

To summarize, the experimental evidence shows that the personalized discovery stage boosts revenue in that stage but crowds out revenue in the subsequent query stage, leaving total revenue unchanged. When the query stage, in addition to the discovery stage, becomes personalized, customers make more purchases and increase their total spending.

While these experiments shed light on multi-stage personalization, they have limitations. First, they do not on their own establish which stage(s) should be personalized to maximize revenue. The experiments do not cover all relevant combinations of multi-stage personalization. In particular, they do not test a setting in which only the query stage is personalized, so its performance relative to personalizing both stages depends on the size of cross-stage spillovers. In addition, because the

[19] For example, a Deloitte study reports that a 0.1-second improvement in site speed increases retail conversions by 8.4% and average cart value by 9.2%. Source: Deloitte Consulting LLP (2021), "Milliseconds Make Millions: The Economic Impact of Site Speed," `https://www.deloitte.com/ie/en/services/consulting/research/milliseconds-make-millions.html` , accessed March 2026.

[20] Airbnb Engineering (2023), "Embedding-Based Retrieval for Airbnb Search," `https://medium.com/airbnb-engineering/embedding-based-retrieval-for-airbnb-search-aabebfc85839`, accessed March 2026.

[21] Booking.com Development (2022), "The Engineering Behind Booking.com's Ranking Platform: A System Overview," `https://medium.com/booking-com-development/the-engineering-behind-booking-coms-ranking-platform-a-system-overview-2fb222003ca6`, accessed March 2026.

two experiments were run in different time periods, it is hard to compare them directly without concern about time trends or other platform-wide changes. The algorithms in the two stages also differ in two field experiments. Although it is standard industry practice to use distinct algorithms at each stage,[22] this variation makes it difficult to disentangle the effect of personalization from the effect of the underlying algorithms.

Second, one might worry that the effects I document are specific to the particular algorithms used in these experiments. To assess robustness, I run additional field experiments that test alternative personalization approaches. I summarize the main takeaways here, with details in Web Appendix E. First, because of substitution between stages, improving discovery-stage personalization does not translate into higher total outcomes. Second, when I hold the ranking algorithm fixed, adding individual-level features generally increases platform revenue. Even with these additional experiments and analyses, the set of experiments on Mercari cannot cover all relevant algorithmic designs. For example, some platforms (such as Amazon) blend individual preferences with aggregate popularity signals (e.g., "Best Seller" lists). Moreover, no experiment evaluates the value of "personalized personalization."

These limitations motivate a structural model that enables counterfactual simulations. The model allows me to: (i) hold the underlying ranking models fixed while varying only the stage at which personalization is applied; (ii) compare alternative stage-specific personalization regimes, including regimes not directly tested in the field; and (iii) evaluate alternative algorithmic designs, including hybrids that combine individual-level and aggregate signals. More broadly, the structural model complements these experiments by serving two purposes. First, it provides a mechanism-based framework that clarifies why personalization can generate cross-stage spillovers, which helps interpret the experimental results and organize the underlying economic forces. Second, it functions as an offline decision-support tool that allows the platform to evaluate counterfactual policies that would be costly to test exhaustively in the field.[23]

Importantly, the experiments remain essential even with the structural model. The first experi-

[22] For example, Netflix employs Large Language Models (LLMs) for personalized discoveries and neural networks to rank query results. See Netflix Tech Blog at `https://netflixtechblog.com/foundation-model-for-personalized-recommendation-1a0bd8e02d39`, accessed March 2026.

[23] The structural model is not intended for production deployment. Instead, it is intended to inform high-level design choices about which stage and for whom personalization should be applied.

ment provides motivating evidence that adding personalization to more stages does not necessarily improve total outcomes, while the second experiment provides exogenous variation that supports identification of the model.

5 Structural Search Model

I begin by laying out the structural search model and its key components: the utility function, the information set and beliefs, and the costs. I then formalize the decision rules that govern customer behavior. Finally, I describe the estimation approach and discuss how the parameters are identified.

5.1 Model Setup

An overview of the structural search model is presented in Figure 4. The model consists of two stages: the discovery stage and the query stage. In each stage, customers engage in sequential search. I assume that the customer's decision to visit the platform is exogenous (e.g., it is independent of whether the discovery stage is personalized).[24] Additionally, customers are assumed to have a preexisting unit demand and stable preferences that remain fixed throughout the platform visit in each session (e.g., viewing or clicking products does not change their preferences).[25,26]

In the discovery stage, the customer is presented with a list of recommended products. They can costlessly observe the visible (i.e., pre-inspection) attributes of these products. In my setting, I consider price, brand, and category as visible attributes.[27] Upon viewing the discovery list, the customer faces three options: (1) inspect an item from the discovery list to access its product page, (2) conduct a search query to switch from the discovery stage to the query stage, or (3) leave the platform by opting for the outside option (e.g., no purchase or purchase on competing platforms). If the customer inspects an item, they fully uncover its hidden attributes and (post-inspection) match value. Hidden attributes in my setting include the number of photos, the item's physical condition,

[24] As personalization is now widespread and many users even expect platforms to provide personalized recommendations at the discovery (landing) page, I assume that whether the discovery stage is personalized does not affect overall traffic to the platform.

[25] As shown in Section 4.1.2, customers visit the platform to fulfill an existing demand, thereby supporting my model assumption.

[26] In my estimation sample, I focus on customers whose intent is to purchase a product in the "Video Games and Consoles" category. Details on how I construct the estimation sample are provided in Section 5.4.1.

[27] As shown in Figure I.1 in the web appendix, customers observe each product's price and thumbnail image. I assume that customers infer the product's category and brand from the thumbnail, which typically displays a brand logo.

item age, and the seller's past sales. The inspected item enters the customer's consideration set. After inspecting an item, a customer can either terminate the process by purchasing an item in their consideration set or choosing the outside option, or continue by inspecting another item or conducting a search query. Customers can purchase only from their consideration set because, on Mercari, an item can be purchased only from its product page (i.e., after inspection).

If the customer conducts a search query, they exit the discovery stage and enter the query stage. In the query stage, they are presented with a new list of products relevant to the query, and can costlessly view the visible attributes of these products. The set of visible and hidden attributes is consistent across both stages. After viewing the query results, the customer faces two options: (1) inspect an item from the query result list, or (2) leave the platform by choosing the outside option.[28] Similarly, by inspecting an item in the query stage, the customer learns its hidden attributes and (post-inspection) match value. These inspected items expand the customer's consideration set.

I adopt the standard assumptions of a sequential search model. Specifically, I assume that it is costless for customers to navigate from a product page back to the discovery list or the query results. In addition, within each session, customers have perfect recall: they remember the utility from previously inspected products (or can costlessly revisit them). I do not model refinement actions such as sorting or filtering because, consistent with Korganbekova and Zuber (2023), I find no experimental evidence that sorting or filtering significantly affects purchase behavior (see Table I.4 in the web appendix).

5.2 Model Specification

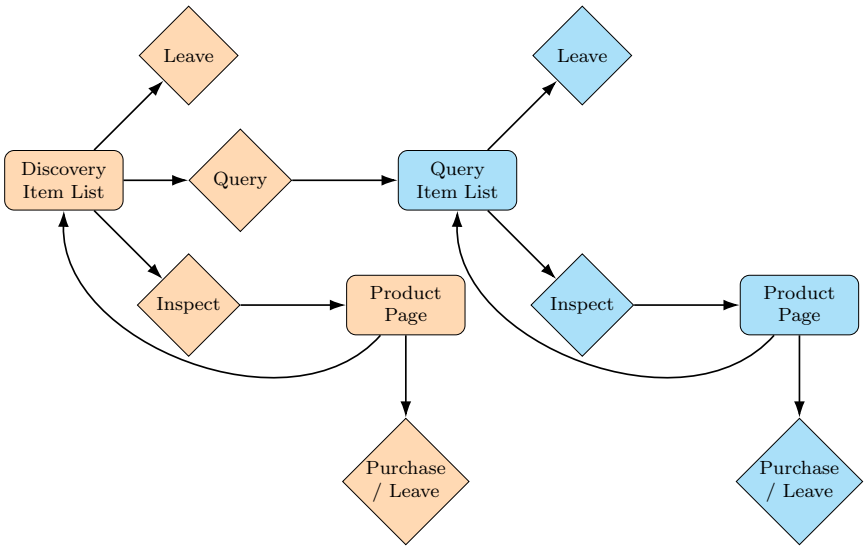
5.2.1 Utility

Customer i derives the following utility from purchasing product j:

$$u_{ij} = \boldsymbol{X_j'\alpha_i} + \boldsymbol{Z_j'\beta_i} + \varepsilon_{ij}^{pre} + \varepsilon_{ij}^{post}. \tag{1}$$

[28] In practice, customers may also return to the discovery stage or submit another search query. In this case, the model could be extended to allow for two additional actions in the query stage: returning to the discovery stage to inspect previously uninspected items, or submitting another query. Both actions mean that customers will enter a new sequential search problem, in which they update cost structures and carry over their current highest realized utility. In this paper, I abstract away from these behaviors to preserve model tractability. In my estimation, I keep only the first query if customers submit multiple queries within a session (16.5% of sessions) and only the first discovery-stage activity if the user returns to the discovery stage after entering the query stage (3.5% of sessions).

Figure 4: Structural Search Model Overview



Notes. Consumers begin the process at the discovery stage, where they are shown a list of discovery items. If they choose to conduct a query to enter the subsequent query stage, they will be presented with a list of query results.

$\boldsymbol{X_j}$ denotes visible attributes (i.e., attributes revealed before inspection), and $\boldsymbol{Z_j}$ denotes hidden attributes (i.e., attributes revealed after inspection). The parameters $\boldsymbol{\alpha_i}$ and $\boldsymbol{\beta_i}$ capture customer i's preferences for visible and hidden attributes respectively. Customer i's product-specific pre-inspection idiosyncratic taste shock ε_{ij}^{pre} is known to the customer before inspection, and post-inspection taste shock ε_{ij}^{post} is known to the customer after inspection. Both taste shocks are unknown to the researcher. Examples of ε_{ij}^{pre} include the i's perceived attractiveness of product j's thumbnail image, and examples of ε_{ij}^{post} include the match between the customer i and the seller of product j. The utility from the outside option u_{i0} is normalized to be ε_{i0}^{post}.

I assume that the idiosyncratic taste shocks (ε_{ij}^{pre} and ε_{i0}^{post}) follow normal distributions, consistent with the commonly adopted assumptions in Weitzman-type search models (e.g., Kim et al., 2010; Chen and Yao, 2017). The standard deviation of the pre-inspection taste shock is set to one as a scale normalization (i.e., $\varepsilon_{ij}^{pre} \sim \mathcal{N}(0,1)$). I assume that the post-inspection taste shock is independently and identically distributed across customers and products, i.e., $\varepsilon_{ij}^{post} \sim \mathcal{N}(0, \sigma^2_{\varepsilon^{post}})$, where its variance $\sigma^2_{\varepsilon^{post}}$ is estimated empirically.[29]

I allow for customer heterogeneity in price sensitivity and hidden attributes through random

[29] See Ursu et al. (2024) for detailed discussions of why fixing the variance of the pre-inspection shock does not lead to a loss of generality, but why both taste shock standard deviations cannot be fixed simultaneously if one aims to monetize the inspection cost.

coefficients, such that $\alpha_i^{price} \sim \mathcal{N}(\alpha^{price}, \Omega_{\alpha^{price}})$ and $\boldsymbol{\beta_i} \sim \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{\Omega_\beta})$, where α^{price} and $\Omega_{\alpha^{price}}$ represent the mean and variance of price sensitivity, and $\boldsymbol{\beta}$ and $\boldsymbol{\Omega_\beta}$ denote the mean and variance of preferences for hidden attributes.

5.2.2 Information Set and Beliefs

I use $\boldsymbol{A_i^D}$ and $\boldsymbol{A_i^Q}$ to denote the list of products seen by customer i in the discovery stage D and query stage Q, respectively. Customer i's initial information set consists of item attributes $\boldsymbol{X_j}$, ε_{ij}^{pre} and position pos_{ij} of $\boldsymbol{A_i^D}$, their preferences $\{\boldsymbol{\alpha_i}, \boldsymbol{\beta_i}\}$, the search query q_i, and the value of the outside option u_{i0}.

Conducting a Query While still in the discovery stage, customer i cannot observe the query-stage results that would appear after submitting a query. Instead, the customer forms beliefs about the value of conducting query q, denoted by u_i^q. If the discovery-stage products do not match the customer's purchase intent, these beliefs are based only on the expected payoff from the query-stage environment. If the discovery-stage products match the customer's purchase intent, the customer additionally incorporates information revealed in the discovery stage when forming beliefs.

I first describe the case in which the discovery-stage products do not match the customer's purchase intent. Consistent with prior work (Koulayev, 2014; De los Santos and Koulayev, 2017; Gu and Wang, 2022), I summarize the value of entering the query stage by the maximum utility among the products the customer expects to *inspect* after submitting query q. This is because when conducting a query, the customer cares only about the product with the highest utility in the query results. Furthermore, focusing on inspected products is important because personalization changes the ordering of the same set of query-relevant items, and inspection is position-dependent (see Figure 3).[30] Formally, for a query-result list $\tilde{\boldsymbol{A}}^Q$, let $\bar{\tilde{\boldsymbol{A}}}^Q$ denote the set of inspected products. I define the query-stage signal as $u_{iq}^Q \equiv \max\{u_{ij} : j \in \bar{\tilde{\boldsymbol{A}}}^Q\}$.

To operationalize this belief, I estimate the distribution of query-stage lists using the empirical distribution. For each query q, I construct $H(\tilde{\boldsymbol{A}}^Q \mid q, treatment, segment)$, where *treatment* indicates the personalization treatment and *segment* (only for treated users) corresponds to the

[30] Personalized and non-personalized query-stage rankings contain the same set of products in the top 100 positions for a given query but differ in their ordering. In the absence of inspection costs, the customer would expect to reach the same maximum utility with and without personalization.

behavioral inputs used by the personalized ranking algorithm (e.g., user price sensitivity, brand preferences, and activity). If customer i is in the non-personalized condition, then H is formed using all observed query-result lists (including positions) shown to users who are in the non-personalized condition and submit query q. If customer i is in the personalized condition, I further restrict H to query-result lists shown to users who are in the personalized condition, submit query q and belong to the same segment as i. Conditioning on segment captures systematic differences in what users expect to see under personalization. For example, price-sensitive users may expect to see lower-priced items ranked near the top (Figure 2). Additional details are provided in Web Appendix F.1.

Because computing the distribution of $\max\{u_{ij} : j \in \bar{\tilde{\boldsymbol{A}}}^Q\}$ is computationally costly, I approximate it via bootstrap simulation (see Koulayev, 2014 and Ghose et al., 2019). For each user i, I repeatedly (i) draw a query-result list $\tilde{\boldsymbol{A}}^Q$ from H given $(q, treatment, segment)$, (ii) simulate which products the user inspects, yielding $\bar{\tilde{\boldsymbol{A}}}^Q$, and (iii) compute $\max_{j \in \bar{\tilde{\boldsymbol{A}}}^Q} u_{ij}$. Across bootstrap draws, I obtain the mean and variance and summarize the query-stage signal as

$$u_{iq}^Q \sim \mathcal{N}\big(\mu_{iq}^Q, (\sigma_{iq}^Q)^2\big), \qquad \mu_{iq}^Q = \mathbb{E}\left[\max_{j \in \bar{\tilde{\boldsymbol{A}}}^Q} u_{ij}\right], \qquad (\sigma_{iq}^Q)^2 = \mathrm{Var}\left(\max_{j \in \bar{\tilde{\boldsymbol{A}}}^Q} u_{ij}\right). \tag{2}$$

Next, when the discovery-stage products match the customer's intent, I treat the discovery stage as an additional signal about the latent payoff from querying. I capture the discovery-stage information relevant for beliefs by constructing an analogous signal from the products the customer inspects in the discovery stage. Let $\bar{\tilde{\boldsymbol{A}}}^D$ denote the set of inspected products in the discovery stage. I define

$$u_{iq}^D \equiv \max\{u_{ij} : j \in \bar{\tilde{\boldsymbol{A}}}^D\}, \qquad u_{iq}^D \sim \mathcal{N}\big(\mu_{iq}^D, (\sigma_{iq}^D)^2\big). \tag{3}$$

When the customer's intent does not match the discovery-stage content, the customer relies only on the expected query-stage payoff, so $u_i^q = u_{iq}^Q$. When the customer's intent matches the discovery-stage content, both u_{iq}^Q and u_{iq}^D are noisy but informative signals about u_i^q. Let their precisions be $\tau_{iq}^Q = 1/(\sigma_{iq}^Q)^2$ and $\tau_{iq}^D = 1/(\sigma_{iq}^D)^2$. Assuming conditional independence, the customer combines the two signals using Bayes' rule. The resulting posterior belief is

$$u_i^q \sim \mathcal{N}\big(\mu_i^q, (\sigma_i^q)^2\big), \qquad \mu_i^q = \frac{\tau_{iq}^Q \mu_{iq}^Q + \tau_{iq}^D \mu_{iq}^D}{\tau_{iq}^Q + \tau_{iq}^D}, \qquad (\sigma_i^q)^2 = \frac{1}{\tau_{iq}^Q + \tau_{iq}^D}. \tag{4}$$

In this case, discovery- and query-stage information jointly shape the customer's beliefs about the value of conducting a query.

Inspecting a Product Before inspecting product j, customer i observes j's visible attributes $\boldsymbol{X}_j$, its rank position pos_{ij}, and its stage s_{ij}. These observables can affect the customer's beliefs about the product's utility in two ways. First, visible attributes may be informative about hidden attributes. For example, a higher price may be correlated with better physical condition. Ignoring this correlation can bias estimates of preferences. Second, both rank position (Fong et al., 2024; Kaye, 2024) and stage can shift beliefs. For instance, conditional on the same visible attributes, items shown in the discovery stage tend to be associated with different seller characteristics than items shown in the query stage (see Figure F.2 in the Web Appendix).[31]

Putting these channels together, I represent customer i's pre-inspection beliefs about product j's utility by a conditional distribution G_i and write $u_{ij} \mid (\boldsymbol{X}_j, s_{ij}, pos_{ij}) \sim G_i$. I assume that G_i is normal, with mean and variance that depend on (i) the item's visible attributes $\boldsymbol{X}_j$, (ii) its stage s_{ij}, and (iii) its rank position pos_{ij}. To discipline beliefs about hidden attributes, I model the relationship between hidden attributes $\boldsymbol{Z}_j$ and visible attributes $\boldsymbol{X}_j$ using the multivariate linear regression $\boldsymbol{Z}_j = \boldsymbol{\Phi}_0 + \boldsymbol{\Phi}_1 \boldsymbol{X}_j + \boldsymbol{\eta}_j$. In addition, to capture stage- and position-specific information, I estimate the empirical distribution of product attributes within each (s, pos) cell by pooling items observed at that stage and position across sessions.[32] Details are presented in Web Appendix F.2.

Under these assumptions, the conditional mean of G_i can then be written as $\boldsymbol{X_j'\alpha_i} + \varepsilon_{ij}^{pre} + (\hat{\boldsymbol{\Phi}}_\mathbf{0} + \hat{\boldsymbol{\Phi}}_\mathbf{1}\boldsymbol{X_j})'\boldsymbol{\beta_i}$. The conditional variance of G_i is $\boldsymbol{\beta_i}^\top \hat{\Sigma}_\eta \boldsymbol{\beta_i} + \sigma^2_{\phi_{ij}}$, where $\hat{\Sigma}_\eta$ is the estimated variance-covariance matrix of the regression residuals $\boldsymbol{\eta}_j$, and $\sigma^2_{\phi_{ij}}$ is the customer's belief about the variance of the post-inspection idiosyncratic taste shock ε_{ij}^{post}. I assume $\sigma^2_{\phi_{ij}}$ equals its true variance $\sigma^2_{\varepsilon_{ij}^{post}}$.

[31] One plausible reason is that the platform showcases products from different sets.

[32] In principle, one could allow these beliefs to vary by personalization condition or customer segment. In my setting, allowing the attribute distributions to vary by personalization condition does not improve model fit. To preserve precision, I impose a common prediction rule across personalization conditions. While the prediction rule for hidden attributes is common across customers, the implied distribution of utility remains heterogeneous through customer-specific preference coefficients.

5.2.3 Costs

Cost of Conducting a Query I model the cost of conducting a query as:

$$c_i^{query} = exp(\tau_0 + \tau_1 Personalized_i), \tag{5}$$

where $Personalized_i$ is an indicator that equals 1 if customer i is in the personalized query-stage ranking condition, and 0 otherwise. I use this term to capture any potential costs borne by customers as a result of personalization. For example, as presented in Section 4.2.2, personalization increases search latency, which results in longer query result loading times and subsequently reduces product inspection decisions.[33] Similar to Gu and Wang (2022), who assume that customers form rational expectations about loading time costs, I assume that customers form rational expectations about the costs introduced by personalization. The baseline cost of conducting a search query is captured by τ_0, which can also be interpreted as the cost incurred when the query-stage ranking is not personalized. The exponential functional form is standard in the literature (e.g., Kim et al., 2010; Ghose et al., 2012; Chen and Yao, 2017) and ensures that costs are positive.

Inspection Cost Since the layout of product lists in the discovery and query stages is identical (see Figure I.1 in the web appendix), I adopt the same inspection cost structure in both stages. Inspection costs are specified as:

$$c_{ijs}^{inspect} = exp(\delta_{0s} + \delta_1 pos_{ijs}), \quad s \in \{D, Q\}, \tag{6}$$

where pos_{ijs} is the position of product j in customer i's list at stage s. As individuals typically read and process product information sequentially from top left to bottom right (Rayner, 1998), items in lower positions can be viewed as incurring higher time or cognitive costs.[34] This interpretation is

[33] One might expect discovery-stage personalization to increase latency and impose additional costs as well. However, I do not find evidence of this in my setting: discovery-stage latency does not differ statistically between the personalized and non-personalized conditions. A plausible explanation is that query-stage ranking requires real-time computation based on the user's query, whereas discovery-stage recommendations can be largely precomputed and served from the backend.

[34] Jameei Osgouei et al. (2026) indicate that, when presented with a matrix, two reading patterns yield qualitatively similar results: (1) reading each column from top to bottom before moving to the next column, and (2) reading each row from left to right before proceeding to the next row.

consistent with Figure I.3 in the Web Appendix, which shows that items in lower positions receive fewer clicks and are inspected later in the session. The parameter δ_{0s} is the baseline inspection cost at stage s. I allow this baseline cost to differ between the discovery and query stages, as, for instance, inspection in the query stage may be subject to tighter time constraints given its later point in the customer journey.

5.3 Optimal Customer Behavior

The model builds on the classic cost-benefit framework of Weitzman (1979), in which individuals evaluate each action by weighing its expected gains against the costs. Optimal decision rules are characterized by reservation utilities, which define the threshold at which an individual is indifferent between continuing the process and stopping.

The reservation utility of a product j equates the marginal gain from inspecting product j with the marginal cost of doing so. Intuitively, it represents the utility level a product must offer for the customer to be just indifferent between inspecting that product and stopping. For each uninspected product j, the customer computes its reservation utility z_{ij} as:

$$c_{ijs}^{inspect} = \int_{z_{ij}}^{\infty} (u_{ij} - z_{ij})\, dG_i(u_{ij} \mid \boldsymbol{X}_j, s_{ij}, pos_{ijs}) \tag{7}$$

Similarly, the reservation utility for conducting a search query q equates its marginal gain with its expected marginal cost. The customer computes the reservation utility z_i^q as:

$$c_i^{query} = \int_{z_i^q}^{\infty} (u_i^q - z_i^q)\, dF_i(u_i^q). \tag{8}$$

Optimal behavior is determined by the relative values of reservation utilities. Under the assumption that customers' beliefs follow a normal distribution, I use the solution derived by Kim et al. (2010) and implement a look-up table to recover reservation utilities (see detailed implementation discussion in Ursu et al. (2024)). Let $R(n)$ denotes the index of the selection option with the n-th largest reservation utility $z_{R(n)}$.[35] Let $\boldsymbol{O} = \{\boldsymbol{A}, q\}$ denotes the set of available options, which includes both the set of uninspected products $\boldsymbol{A}$ and the unvisited query q. Let $\bar{\boldsymbol{A}}$ be the set

[35] For notational simplicity, I drop the customer-specific subscript i in what follows.

of products that have been inspected across both stages. The customer's current highest realized utility, denoted by u^*, is defined as the maximum utility among the inspected products and the outside option: $u^* = \max\left\{\max_{j \in \tilde{\boldsymbol{A}}} u_j,\ u_0\right\}$. I outline the optimal rules as follows.

Selection Rule: The customer ranks all available options (including uninspected products and the query if in the discovery stage) in decreasing order of reservation utilities, and then sequentially inspects each option in that order.

Stopping Rule: The customer stops the process if the current highest realized utility exceeds the maximum reservation utility of all remaining available options (i.e., uninspected products and the unvisited query if in the discovery stage). Thus, stopping at step k implies: $u^* \geq \max_{n=k+1}^{\boldsymbol{O}} z_{R(n)}$.

Choice Rule: The customer purchases the product with the highest realized utility among the inspected products and the outside option. Therefore, purchasing product j implies: $u_j = u^*$.

5.4 Estimation and Identification

5.4.1 Estimation Sample Construction

In the structural search model estimation, I focus on consumer sessions that reveal purchase intent for items in the "Video Games and Consoles" category. I choose this category because its products are relatively standardized, meaning that, for example, product attributes like photos provide limited incremental information. In addition, it is one of Mercari's highest-volume categories in terms of queries, inspections, and sales. I define a consumer session as exhibiting purchase intent in the "Video Games and Consoles" category if it meets any of the following criteria: (i) the customer submits a relevant search query; (ii) if no query is submitted, the customer purchases a product from that category in the discovery stage; or (iii) if neither a query is submitted nor a purchase is made, the customer inspects only items from that category. In cases (ii) and (iii) where no query is conducted, I proxy for the customer's intended query using their most recent query in the "Video Games and Consoles" category.[36]

Because discovery-stage products can come from multiple categories depending on a customer's prior activity and many items are never inspected, it is difficult to identify category-specific co-

[36] Note that for a product in the "Video Games and Consoles" category to appear in the discovery stage, the customer must have previously submitted a related query.

efficients in the utility function. Thus, instead of including individual product categories, I use a dummy variable to indicate whether product j belongs to the same category as customer i's purchase intent (i.e., "Video Games and Consoles" category in my model estimation).

5.4.2 Model Estimation Approach

The parameters to be estimated $\boldsymbol{\Theta}$ include the preference parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \Omega_{\alpha^{price}}, \boldsymbol{\Omega}_{\boldsymbol{\beta}}\}$, cost parameters $\{\boldsymbol{\delta}, \boldsymbol{\tau}\}$, and the post-inspection taste shock standard deviation $\sigma_{\varepsilon^{post}}$. I estimate the model using data from the query-stage personalization experiment. The random assignment in the experiment introduces exogenous variation in customers' propensity to conduct a query. The dataset contains, for each customer, the full list of products shown at both stages (with their attributes and positions), the submitted query, the inspected and purchased items (including the outside option), and the order of inspections. I also observe each customer's personalization condition and segment.

I estimate the model using NNE proposed by Wei and Jiang (2025). The core idea behind NNE is to train a deep learning model to learn the mapping from data to the underlying structural parameters. In many structural econometric models, including the sequential search model, data moments are sufficient for pinning down parameters. This approach offers several advantages over traditional methods such as the simulated method of moments (SMM) and maximum likelihood estimation (MLE). First, NNE eliminates the need to evaluate high-dimensional integrals over unobservables, which are often required in SMM and MLE. Instead, the neural network learns the mapping from data to parameters. Second, NNE is robust to redundant or noisy moments, as the learning algorithm automatically prioritizes the most informative features for parameter identification. Third, this approach offers computational efficiency: once trained, the estimator can generate point estimates and statistical uncertainty measures with minimal additional computation. Finally, NNE is relatively robust to the choice of neural network architecture and hyperparameters.

I estimate my model parameters using NNE as follows. First, given the products and their positions in my actual estimation sample, I simulate consumer decisions over a grid of candidate parameter values. The simulated decisions include whether to inspect a product, conduct a query, and purchase a product. Second, for each parameter vector, I summarize simulated outcomes into a set of data moments (e.g., average purchase rate), computed at the product, session, and individual levels. Third, I train a neural network using the simulated data moments and their associated

parameter values as the training set. These training steps learn an inverse mapping from moments to parameters: given a set of observed data moments, what parameter values are most likely to have generated them? Finally, I compute the same set of moments from the actual estimation sample and feed them into the neural network. The network then outputs the estimated model parameters and covariance matrix. Further details on NNE are provided in Web Appendix G.1.

5.4.3 Identification

This section presents the informal identification of parameters. I discuss the identification of position effects δ_1, preference parameters $\{\boldsymbol{\alpha_i}, \boldsymbol{\beta_i}\}$, price parameter α^{price}, cost parameters $\{\tau_0, \tau_1, \delta_0\}$, and standard deviation of post-inspection taste shock $\sigma_{\varepsilon^{post}}$. I perform a Monte Carlo simulation to verify parameter identification and find that the estimation procedure precisely recovers all parameters (see Web Appendix G.2).

Position Effects $\boldsymbol{\delta_1}$ At the discovery stage, products are selected and ranked based on customers' past histories. Although the proprietary discovery-stage algorithm is not directly observed, internal discussions with Mercari reveal its key input variables, such as historical search queries and item clicks. Similar to De los Santos and Koulayev (2017), I use these inputs to capture the systematic component of discovery products. Concretely, I estimate a predictive model of position using these algorithm inputs and obtain the fitted value. I then include this value as a control function in the utility specification and treat the resulting residual variation as exogenous to individual preferences.[37]

At the query stage, as detailed in Section 3.1, when a customer conducts a query, the platform first retrieves a set of items relevant to that query, then orders them by relevance, and finally re-orders the top 100 items using either a personalized or a non-personalized LambdaMART model. This procedure implies that the candidate set is primarily determined by query relevance, while the final ordering reflects the re-ranking model. I leverage full knowledge of the deployed LambdaMART specification and its input features to compute each item's ranking score in each session. These

[37] Since the recommendation algorithms at the discovery stage were designed to encourage serendipitous exploration during the period from which my estimation sample is drawn, the residual variation can be interpreted as the system's exploratory component, which is plausibly exogenous to individual preferences.

ranking scores determine an item's relative position in the list, and I include them as a control for position.

At both stages, I observe customers' repeated inspection decisions across positions, and these decisions are informative about the customer's sensitivity to position.

Preference Parameters $\boldsymbol{\alpha_i, \beta_i}$ As discussed above, product attributes are not endogenous in the query stage. However, they may be endogenous in the discovery stage, since the algorithm selects and ranks products based on customers' past behaviors. I correct for this endogeneity using a control-function approach that leverages the discovery-stage algorithm's input variables, following the same approach used above.

I identify customer preferences based on patterns in purchase and inspection. As I observe purchases directly, the identification of the preference parameters is as in a conditional choice model. Similar to how purchase decisions among a set of products reveal mean preference coefficients in a traditional discrete choice model, purchase decisions among inspected products inform the mean preference coefficients in my model. Given the low purchase incidence (3.1% of sessions), I further incorporate customers' inspection decisions. The selection of which product to inspect helps to pin down the preference parameters (e.g., Bronnenberg et al., 2016; Chen and Yao, 2017; Honka et al., 2017). For example, a tendency to inspect more low-priced items indicates larger price sensitivity.

The heterogeneity of the preference coefficients across customers can be recovered since I observe a sequence of decisions for each customer, including product inspection, query submission, and purchase. Repeated observations for each customer are informative for customer-specific preference coefficients, while the variation in these coefficients across individuals enables me to identify preference heterogeneity.

Price Parameter α^{price} Estimating the price parameter is prone to endogeneity, as prices may correlate with unobserved product-level taste shocks. To address this, I follow a standard approach using instruments constructed from the product attributes of competing products (see Gandhi and Houde (2019)). I define a "market" in my setting as the set of products that belong to the same category-brand cluster and are listed in the same week.

Cost Parameters δ_0, τ_0, τ_1 Conditional on the consideration set (i.e., a set of inspected items), the purchase decision depends only on consumer preferences and not on inspection costs. Therefore, I can separate inspection costs from preference parameters by using conditional purchase probabilities. The joint variation in inspection and purchase enables the identification of the distribution of consumer preferences. Conditional on those preferences, observed inspection decisions reveal the distribution of inspection costs. I identify the baseline inspection costs δ_0 using the number of product inspections in each stage.

To identify the effect of personalization on customers' cost of conducting a query τ_1, I exploit variation in customers' experimental assignment (i.e., whether they are in the personalized or non-personalized condition). The baseline cost of conducting a search query τ_0 is inferred from the frequency with which customers conduct a query. In my estimation sample, 84.7% of sessions involve a query while 15.3% do not.

Post-Inspection Taste Shock Standard Deviation $\sigma_{\varepsilon^{post}}$ The variation in the number of items a customer inspects pins down the ratio of the inspection cost to the post-inspection taste shock standard deviation. These two parameters can be separately identified by the parametric form. As the product position shifts inspection costs but not post-inspection utility, this inspection cost shifter allows me to identify the standard deviation of post-inspection taste shock (Yavorsky et al., 2021).

6 Estimation Results and Counterfactual Simulations

6.1 Estimation Results

Table 3 presents the estimation results of the sequential search model. The comparisons between observed and estimated data patterns in Web Appendix G.3 suggest that my model fits the data well. To express the coefficients in monetary terms, I normalize each estimate by dividing it by the absolute value of the price coefficient. I find that customers derive higher utility from items that are lower priced, have fewer photos, are newly listed (i.e., smaller item age), are in better physical condition, and are sold by more experienced sellers. The estimation results also reveal heterogeneity in these preferences. The positive coefficient of "Video Games and Consoles" category

is not surprising as customers prefer products from categories that align with their purchase intent. Relative to unbranded products, customers are willing to pay premia for four major brands: Sony ($0.54), Microsoft ($0.65), Nintendo ($0.34), and Pokémon ($0.47).

In terms of costs, I estimate that the baseline inspection cost is $0.24 in the discovery stage and $0.17 in the query stage. The baseline cost of conducting a search query is $0.16. To quantify position effects, I compute the dollar equivalent of the change in inspection cost from a one-unit increase in position. The estimated position effect is $0.06 in the discovery stage and $0.04 in the query stage. For the initiation screen of 12 products, these inspection costs correspond to $0.72 for the discovery list and $0.48 for the query list. Personalization introduces frictions in transitioning to the query stage, with an estimated cost equivalent to about $0.16.

Table 3: Estimation Results of Structural Search Model

Variables	Mean	SE	Heterogeneity (SD)	SE	Dollarized Value ($)
Preferences					
Constant	-3.4608	0.8528			
Price	-0.2453	0.0485	0.1134	0.0437	
1(Video Game Category)	0.8369	0.2501			0.6548
Brand: Sony	0.6929	0.1554			0.5422
Brand: Microsoft	0.8365	0.1558			0.6545
Brand: Nintendo	0.4343	0.1319			0.3398
Brand: Pokemon	0.5985	0.2166			0.4683
Brand: Others	0.2365	0.0965			0.1851
N Photos	-0.4013	0.0851	0.0284	0.0179	-0.3140
Item Age	-0.7931	0.3176	0.0367	0.0124	-0.4284
Physical Condition	0.1283	0.0305	0.0284	0.0476	0.1004
Seller Past Sales	0.0912	0.6099	0.0770	0.0477	0.0747
Costs					
Inspection Base (Discovery)	-1.1786	0.1034			0.2408
Inspection Base (Query)	-1.5134	0.0771			0.1723
Position	0.2157	0.0705			0.0588 / 0.0415
Query Base	-1.5985	0.5616			0.1582
Personalization	0.8237	0.2762			0.1584

Notes. The estimation sample includes 8,138 customers and 17,884 sessions. The last column reports the monetary value of each variable, calculated by dividing the coefficient estimate by the absolute value of the logged price coefficient. For example, the implied dollarized value of the preference is given by $\frac{\hat{\alpha}}{\exp(\hat{\alpha}^{price})}$ and the baseline cost is calculated as $\frac{\exp(\hat{\delta})}{\exp(\hat{\alpha}^{price})}$. Item age and seller past sales are (log+1)-transformed. The omitted brand is unbranded.

6.2 Counterfactual Simulations

As discussed in Section 4.3, some questions cannot be fully answered by field experiments alone and instead require counterfactual simulations. The first counterfactual simulation explores how

personalization at different stages of the customer journey affects total revenue. The second counterfactual simulation quantifies the value of personalized personalization.

6.2.1 Multi-Stage Personalization

To isolate stage-specific effects from algorithm effects, I rank products in both the discovery and query stages using the same utility-based scoring rule. The ranking score is a convex combination of (i) predicted utility under an individual customer's preferences and (ii) predicted utility under average population preferences (see Web Appendix H for details). I vary the weight on (i) versus (ii) to capture different personalization intensities: popularity-based ranking (no personalization), full personalization, and intermediate policies that blend individual-level and population-level signals.[38] Using these designs, I evaluate how personalization at each stage affects total revenue by comparing policies that apply personalization (or partial personalization) at neither stage, at exactly one stage (discovery or query), or at both stages.

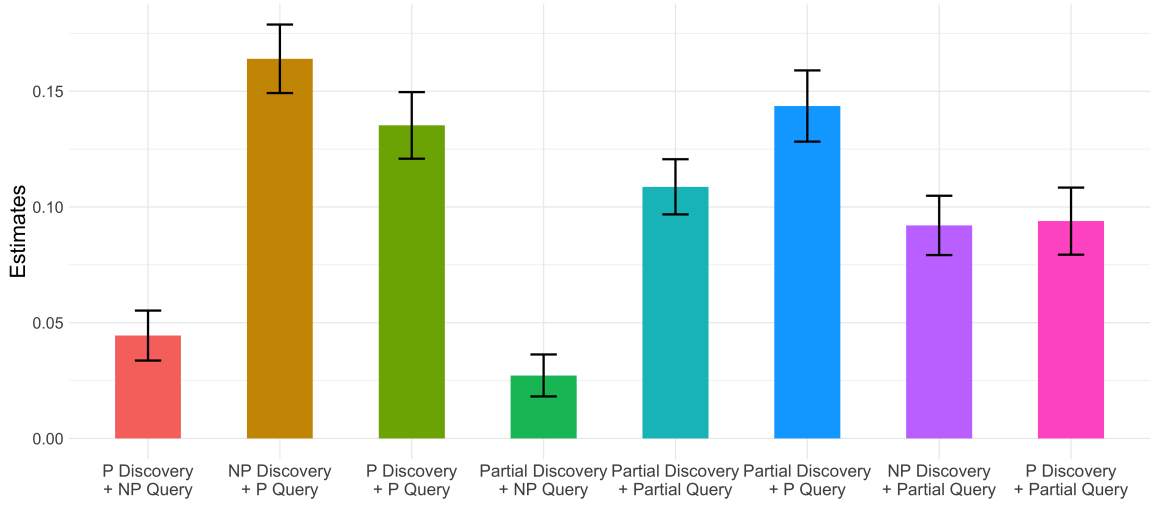
The results, presented in Figure 5, highlight the importance of selecting the appropriate stage for personalization. Personalizing only the early discovery stage yields a modest increase (4.4%) in total revenue relative to no personalization at either stage. Extending personalization to both stages does not maximize revenue. Instead, personalizing only the query stage delivers the largest uplift, generating 2.9% more revenue than personalizing both stages. To understand this pattern, I examine how discovery-stage personalization affects continuation to the query stage. Discovery-stage personalization improves intent matching (by 33%), making customers more likely to use their discovery-stage experience to update beliefs about the expected gains from querying. However, the discovery stage may not always provide a strong fit with the customer's current demand. Customers who use this weak match to infer the value of querying may update toward a lower expected payoff, reducing continuation value and increasing early exit.

Partial personalization, which blends popularity-based ranking with individual utility, performs well despite slightly underperforming full personalization at the query stage or at both stages. This approach can be attractive when firms wish to preserve product variety for long-run performance (Chen et al., 2024), or when data and resource constraints limit the accuracy of individual-level

[38] I assume that τ_1, which captures how personalization affects the cost of querying, is invariant to the degree of personalization at the query stage.

preference inference.

Figure 5: Total Revenue of Multi-Stage Personalization



Notes. This figure reports estimates from the specification $\log(Revenue_i + 1) = \alpha_0 + \alpha_1 T_i + \epsilon_i$, where T_i denotes the personalization strategy. The omitted strategy is no personalization at either stage. The x-axis displays the personalization strategies. The y-axis reports the estimated percentage change in revenue relative to the omitted condition. Error bars represent 95% confidence intervals. "NP" denotes non-personalized and "P" denotes personalized.

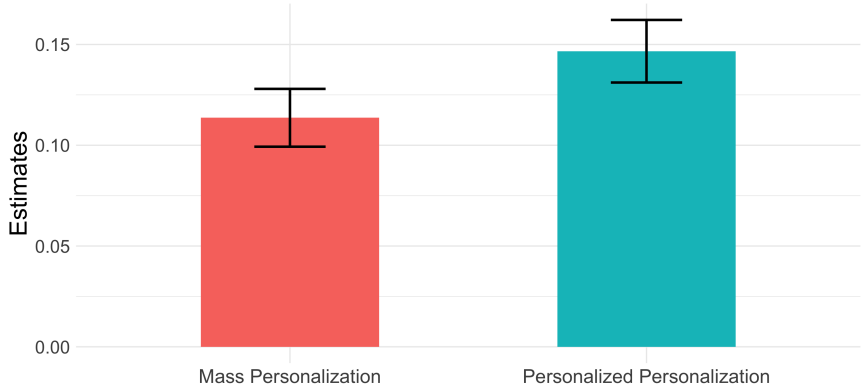
6.2.2 Personalized Personalization

The experimental results in Section 4.1.2 show that personalization delivers larger gains for customers with consistent preferences, whereas the benefits attenuate and can even become negative for customers with more dynamic preferences. This heterogeneity suggests that a platform may improve performance by tailoring personalization according to preference consistency. To quantify the value of this strategy, I simulate an environment in which 20% of customers are randomly assigned to a "dynamic-preference" segment and 80% to a "consistent-preference" segment.[39] For customers in the dynamic-preference segment, I introduce a 10% random perturbation to their estimated preference parameters to mimic temporal shifts in preference, while customers in the consistent-preference segment experience no shocks. I then compare (i) universal personalization, which personalizes both stages for all customers, with (ii) personalized personalization, which personalizes both stages only for customers in the consistent-preference segment. As shown in Figure 6, personalized personalization increases total individual-level revenue by 3.3% relative to universal

[39] This split best matches the empirical patterns; alternative splits such as 25%/75% yield similar results.

personalization and by 14.7% relative to applying no personalization at either stage. In practice, firms can infer preference consistency from historical behavior such as purchases. For example, a customer who repeatedly purchases low-priced or new-condition items is likely to benefit from personalization, whereas a customer whose choices frequently shift across price tiers or product attributes may be better served by a less personalized, popularity-based experience.

Figure 6: Revenue of Personalized Personalization Based on Preference Consistency



Notes: This figure reports estimates from the specification $\log(Revenue_i + 1) = \alpha_0 + \alpha_1 T_i + \epsilon_i$, where T_i denotes the personalization strategy. The omitted strategy is no personalization at either stage. The x-axis displays the personalization strategies. The y-axis reports the estimated percentage change in revenue relative to the omitted condition. Error bars represent 95% confidence intervals.

7 Conclusion

Personalization is central to the operations of many digital platforms. For instance, Spotify curates discovery feeds, Netflix tailors thumbnails, and Amazon customizes search query results. From discovery-page recommendations that capture initial engagement to query-stage rankings that showcase additional relevant products, personalization influences multiple stages of the customer journey. Although many firms make significant investments to optimize each touchpoint, this siloed approach could create tensions when gains from personalization in one stage offset benefits in another. Ignoring important spillovers between stages may lead to suboptimal outcomes. Designing effective multi-stage personalization thus requires a systematic understanding of both within-stage effects and cross-stage spillovers. This paper studies two stages where personalization is central—the firm-driven discovery stage and the customer-initiated query stage—and examines how outcomes change depending on which stage or stages are personalized.

Using two field experiments, I find that although personalizing the discovery stage can boost initial engagement, it can reduce user engagement in the subsequent query stage, ultimately leaving total revenue flat. When both stages are personalized, total revenue increases without cannibalizing discovery-stage revenue. These findings suggest a one-way negative spillover from the discovery stage to the query stage. I also find that the impact of personalization is more pronounced for customers with consistent preferences. In fact, personalizing the discovery stage can reduce revenue for those with more dynamic preferences.

To identify the revenue-maximizing multi-stage personalization strategy and quantify the value of personalized personalization, I develop and estimate a sequential search model. Counterfactual simulations based on my model estimates show that personalizing only the query stage generates the highest total revenue, outperforming both no personalization and personalization at both stages. The model highlights that discovery-stage personalization can improve intent matching by surfacing items aligned with a customer's purchase intent, and customers use this early experience to form expectations about the benefits of continuing to the query stage. When discovery-stage personalization fails to deliver a strong preference fit, these expectations can decline, increasing early exit. Many of these customers would have found better matches had they proceeded to the query stage, where the submitted query reveals richer information about current preferences and rankings can condition on that signal to return better results. Finally, the model implies that personalization can be further improved by targeting it: offering personalization only to users with consistent preferences increases platform revenue by 3.3% relative to personalizing all customers.

Overall, this paper shows that expanding personalization to additional touchpoints does not necessarily improve aggregate performance, challenging the view that more personalization is always better. The findings caution against stage-by-stage optimization, which is common when product discovery and query are managed and evaluated separately. When cross-stage spillovers are present or the objective function is misspecified, optimizing each stage in isolation can be counterproductive; in some cases, turning off personalization at a stage improves overall outcomes. Moreover, when the platform cannot observe why a user abandoned a prior session or cannot reliably predict preference shifts, personalization based on stale signals can mis-rank products and reduce the perceived value of continuing the journey. Personalization should therefore be treated as a coordinated, multi-stage strategy embedded in the broader consumer decision process rather

than as a sequence of independent interventions. A practical implication is that platforms should design objective functions and learning systems that evaluate and optimize personalization at the level of the full customer journey, instead of maximizing stage-level metrics. More broadly, this paper highlights these challenges and provides a framework for firms to decide whether, where, and how much to personalize in the customer journey.

Beyond coordinating personalization across stages, firms can improve performance by exploiting heterogeneity in responses to personalization. I illustrate this idea using preference consistency, but it is only a starting point. More broadly, firms can use behavioral histories to implement "personalized personalization" by directing personalization toward users with the highest expected gains and adjusting its intensity across users and stages.

Although this paper studies a single platform (Mercari), the setting shares key features with other e-commerce platforms such as eBay, Depop, Poshmark, and Etsy. To gauge how broadly these patterns may extend, I use cross-category variation in market environments as suggestive evidence. For instance, while many Mercari listings are single-unit, some categories have deeper inventories and many close substitutes, resembling multi-unit retail. In these larger-inventory, more homogeneous categories, personalization effects are stronger, consistent with greater gains from improved matching when substitution opportunities are richer. In addition, Mercari primarily features pre-owned items whereas other platforms may offer a higher share of goods in new condition. I compare effects across categories that differ in the share of items shown in new (versus used) condition and find stronger effects in categories with a higher prevalence of new items.

This paper has several limitations. First, the field experiments last for less than two weeks, so the results primarily capture short-run effects of personalization. Longer-run impacts may differ if market participants adjust their behavior over time. For example, customers may learn how the platform's personalization works and adapt their behavior, and sellers may respond strategically by changing prices (Kaye, 2024) or product quality, which could in turn affect consumer welfare and long-term platform revenue. Second, I focus on two stages of the customer journey, but companies often personalize additional touchpoints such as push notifications, promotions, and checkout incentives. Studying these interventions jointly with discovery- and query-stage personalization is a natural extension. Relatedly, I use preference consistency to illustrate personalized personalization, but this is just one way to operationalize the idea. With richer data, platforms could vary personal-

ization intensity across users, tailor it at the category level, or adapt it dynamically as preferences evolve. Third, my empirical setting is an e-commerce marketplace where users often arrive with relatively well-defined purchase intent. The implications may therefore differ in contexts with more exploratory or serendipitous consumption, such as social media platforms. Finally, the structural model is estimated using data from a single product category and all analyses are conducted on a single platform, so the generalizability of the findings should be interpreted with caution. Future research could examine multi-stage personalization over longer horizons, across platforms with different market environments and customer compositions, and under more advanced algorithms.

References

Abraham, Mark and Edelman, David C (2024). "Personalization Done Right The five dimensions to consider and how AI can help". *Harvard Business Review*, 103 (11-12), 104–115.

Bronnenberg, Bart J, Kim, Jun B, and Mela, Carl F (2016). "Zooming in on choice: How do consumers search for cameras online?" *Marketing Science*, 35 (5), 693–712.

Burges, Christopher JC (2010). "From ranknet to lambdarank to lambdamart: An overview". *Learning*, 11 (23-581), 81.

Chen, Guangying, Chan, Tat, Zhang, Dennis, Liu, Senmao, and Wu, Yuxiang (2024). "The effects of diversity in algorithmic recommendations on digital content consumption: A field experiment". *Available at SSRN 4365121*.

Chen, Yuxin and Yao, Song (2017). "Sequential search with refinement: Model and application with click-stream data". *Management Science*, 63 (12), 4345–4365.

Chen, Zirou, Shi, Mengze, and Zhong, Zemin (Zachary) (2025). "Predictive Accuracy, Consumer Search, and Personalized Recommendation". *Available at SSRN 4298841*.

Choi, Hana and Mela, Carl F (2019). "Monetizing online marketplaces". *Marketing Science*, 38 (6), 948–972.

Compiani, Giovanni, Lewis, Gregory, Peng, Sida, and Wang, Peichun (2024). "Online search and optimal product rankings: An empirical framework". *Marketing Science*, 43 (3), 615–636.

Dang, Chu (Ivy), Ursu, Raluca, and Chintagunta, Pradeep K. (2025). "Going Back to Move Forward? How Search Revisits on a Website We Built, and in Field Data, Inform Us about Search Outcomes". *Available at SSRN 3626451*.

De los Santos, Babur and Koulayev, Sergei (2017). "Optimizing click-through in online rankings with endogenous search refinement". *Marketing Science*, 36 (4), 542–564.

Donnelly, Robert, Kanodia, Ayush, and Morozov, Ilya (2024). "Welfare effects of personalized rankings". *Marketing Science*, 43 (1), 92–113.

Fang, Xing, Kim, SunAh, and Chintagunta, Pradeep K (2026). "Too many or too few? Information cues in recommender systems and consequences for search and purchase behavior". *Journal of Marketing*, 90 (1), 9–28.

Fong, Jessica (2024). "Effects of market size and competition in two-sided markets: Evidence from online dating". *Marketing Science*, 43 (5), 971–985.

Fong, Jessica, Manchanda, Puneet, and Song, Yu (2026). "How Effective is Suggested Pricing?: Experimental Evidence from an E-Commerce Platform". *Available at SSRN 4993457*.

Fong, Jessica, Natan, Olivia, and Pantle, Ranmit (2024). "Consumer Inferences from Product Rankings: The Role of Beliefs in Search Behavior". *Available at SSRN*.

Fong, Nathan M (2017). "How targeting affects customer search: A field experiment". *Management Science*, 63 (7), 2353–2364.

Gandhi, Amit and Houde, Jean-François (2019). "Measuring substitution patterns in differentiated-products industries". *NBER Working paper*, (w26375).

Ghose, Anindya, Ipeirotis, Panagiotis G, and Li, Beibei (2012). "Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content". *Marketing Science*, 31 (3), 493–520.

Ghose, Anindya, Ipeirotis, Panagiotis G, and Li, Beibei (2019). "Modeling consumer footprints on search engines: An interplay with social media". *Management Science*, 65 (3), 1363–1385.

Goić, Marcel, Jerath, Kinshuk, and Kalyanam, Kirthi (2022). "The roles of multiple channels in predicting website visits and purchases: Engagers versus closers". *International Journal of Research in Marketing*, 39 (3), 656–677.

Greminger, Rafael P (2022). "Optimal search and discovery". *Management Science*, 68 (5), 3904–3924.

Gu, Chris and Wang, Yike (2022). "Consumer online search with partially revealed information". *Management Science*, 68 (6), 4215–4235.

Hodgson, Charles and Lewis, Gregory (2025). "You can lead a horse to water: Spatial learning and path dependence in consumer search". *Econometrica*, 93 (4), 1299–1332.

Holtz, David, Carterette, Ben, Chandar, Praveen, Nazari, Zahra, Cramer, Henriette, and Aral, Sinan (2020). "The engagement-diversity connection: Evidence from a field experiment on spotify". In "Proceedings of the 21st ACM Conference on Economics and Computation", 75–76.

Honka, Elisabeth, Hortaçsu, Ali, and Vitorino, Maria Ana (2017). "Advertising, consumer awareness, and choice: Evidence from the US banking industry". *The RAND Journal of Economics*, 48 (3), 611–646.

Honka, Elisabeth, Hortaçsu, Ali, and Wildenbeest, Matthijs (2019). "Empirical search and consideration sets". In "Handbook of the Economics of Marketing", volume 1, 193–257. Elsevier.

Honka, Elisabeth, Seiler, Stephan, and Ursu, Raluca (2024). "Consumer search: What can we learn from pre-purchase data?" *Journal of Retailing*, 100 (1), 114–129.

Hosanagar, Kartik, Fleder, Daniel, Lee, Dokyun, and Buja, Andreas (2014). "Will the global village fracture into tribes? Recommender systems and their effects on consumer fragmentation". *Management Science*, 60 (4), 805–823.

Jameei Osgouei, Ata, Ching, Andrew T, Ratchford, Brian T, and Shahrokhi Tehrani, Shervin (2026). "Estimating Position and Social Influence Effects in Online Search". *Marketing Science*, 45 (1), 203–223.

Jiang, Zhenling, Chan, Tat, Che, Hai, and Wang, Youwei (2021). "Consumer search and purchase: An empirical investigation of retargeting based on consumer online behaviors". *Marketing Science*, 40 (2), 219–240.

Kaye, Aaron P (2024). "The Personalization Paradox: Welfare Effects of Personalized Recommendations in Two-Sided Digital Markets". *Working Paper*.

Kim, Jun B, Albuquerque, Paulo, and Bronnenberg, Bart J (2010). "Online demand under limited consumer search". *Marketing science*, 29 (6), 1001–1023.

Korganbekova, Malika and Zuber, Cole (2023). "Balancing user privacy and personalization". *Working Paper*.

Koulayev, Sergei (2014). "Search for differentiated products: identification and estimation". *The RAND Journal of Economics*, 45 (3), 553–575.

Li, Xitong, Grahl, Jörn, and Hinz, Oliver (2022). "How do recommender systems lead to consumer purchases? A causal mediation analysis of a field experiment". *Information Systems Research*, 33 (2), 620–637.

Lu, Zipei and Kannan, PK (2026). "AI for Customer Journeys: A Transformer Approach". *Journal of Marketing Research*, 63 (1), 1–26.

MacKenzie, Ian, Meyer, Chris, and Noble, Steve (2013). "How retailers can keep up with consumers". *McKinsey & Company*, 18 (1), 1–10.

Ostuni, Vito, Kofler, Christoph, Nilange, Manjesh, Lamkhede, Sudarshan, and Zylberglejd, Dan (2023). "Search personalization at netflix". In "Companion Proceedings of the ACM Web Conference 2023", 756–758.

Padilla, Nicolas, Ascarza, Eva, and Netzer, Oded (2024). "The customer journey as a source of information". *Quantitative Marketing and Economics*, 1–40.

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

Rayner, Keith (1998). "Eye movements in reading and information processing: 20 years of research." *Psychological bulletin*, 124 (3), 372.

Song, Michelle (2021). "How Do Personalized Recommendations Affect Consumer Exploration: A Field Experiment". *Working paper*.

Ursu, Raluca, Seiler, Stephan, and Honka, Elisabeth (2024). "The sequential search model: A framework for empirical research". *Quantitative Marketing and Economics*, 1–49.

Ursu, Raluca M (2018). "The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions". *Marketing Science*, 37 (4), 530–552.

Wan, Xiang, Kumar, Anuj, and Li, Xitong (2024). "How do product recommendations help consumers search? Evidence from a field experiment". *Management Science*, 70 (9), 5776–5794.

Wang, Xuewen and Liu, Jia (2026). "Modeling Consumer Search with Refinement under Non-Compensatory Preference Rules". *Working paper*.

Wei, Yanhao and Jiang, Zhenling (2025). "Estimating parameters of structural models using neural networks". *Marketing Science*, 44 (1), 102–128.

Weitzman, Martin L. (1979). "Optimal Search for the Best Alternative". *Econometrica*, 47 (3), 641–654.

Yavorsky, Dan, Honka, Elisabeth, and Chen, Keith (2021). "Consumer search in the US auto industry: The role of dealership visits". *Quantitative Marketing and Economics*, 19, 1–52.

Yoganarasimhan, Hema (2020). "Search personalization using machine learning". *Management Science*, 66 (3), 1045–1070.

Yoon, Song-Oh and Simonson, Itamar (2008). "Choice set configuration as a determinant of preference attribution and strength". *Journal of Consumer Research*, 35 (2), 324–336.

Yuan, Zhe, Chen, AJ Yuan, Wang, Yitong, and Sun, Tianshu (2025). "How recommendation affects customer search: A field experiment". *Information Systems Research*, 36 (1), 84–106.

Zhan, Ruohan, Han, Shichao, Hu, Yuchen, and Jiang, Zhenling (2024). "Estimating Treatment Effects under Recommender Interference: A Structured Neural Networks Approach". *arXiv preprint arXiv:2406.14380*.

Zhang, Luna, Ursu, Raluca, Honka, Elisabeth, and Yao, Oliver (2023). "Product discovery and consumer search routes: Evidence from a mobile app". *Available at SSRN*, 4444774.

Web Appendix to "Optimizing Multi-Stage Personalization in the Customer Journey"

A Randomization, Manipulation, and Interference Checks

A.1 Randomization Check

For the personalized discovery-stage experiment, I compare user activity in the 60 days prior to the experiment. I conduct a t-test to check the balance between the treatment and control conditions. As shown in Table A.1 Panel (A), I find no systematic differences between the two conditions in discovery-stage clicks, query-stage clicks, and total orders.

For the personalized query-stage experiment, I analyze user behavior over the 30 days preceding the experiment, including whether the user was active, the number of queries they submitted, and their total spending. As shown in Table A.1 Panel (B), there are no systematic differences between the treatment and control conditions.

Table A.1: Randomization Check

Panel (A): Personalized Discovery-Stage Experiment					
	Non-Personalized		Personalized		
	Mean	SE	Mean	SE	p-value
Discovery-Stage Clicks	2.107	0.008	2.107	0.008	0.986
Query-Stage Clicks	27.784	0.059	27.787	0.059	0.968
Total Orders	0.008	0	0.008	0	0.88
Panel (B): Personalized Query-Stage Experiment					
	Non-Personalized		Personalized		
	Mean	SE	Mean	SE	p-value
Whether Active	0.799	0.001	0.801	0.001	0.188
Queries	1.907	0.004	1.914	0.004	0.246
Spending	0.775	0.0037	0.773	0.004	0.65

Notes. Panel (A) presents the summary statistics of user activities in the past 60 days before the personalized discovery-stage experiment. Panel (B) presents the summary statistics of user activities in the past 30 days before the personalized query-stage experiment. Both queries and spending use $\log(x+1)$ transformation.

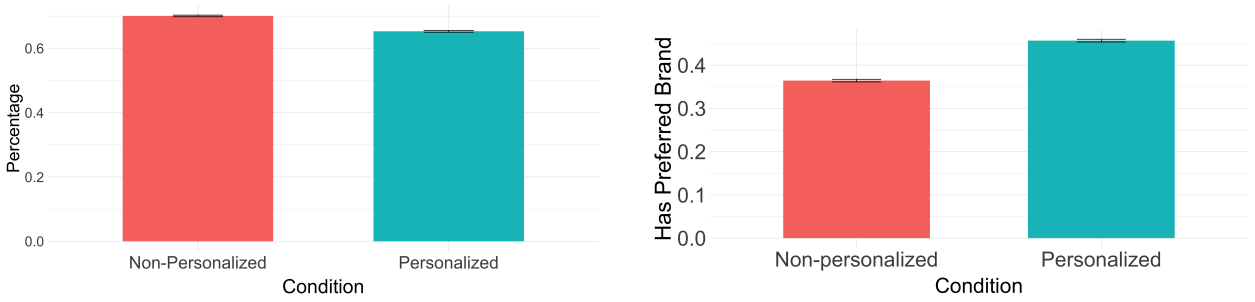
A.2 Manipulation Check

Recall that in the personalized discovery-stage experiment, customers assigned to the non-personalized condition are shown products selected primarily on popularity. As popularity is proxied by histori-

cal performance, these product discoveries should be disproportionately drawn from category–brand pairs with high past revenue. To assess this, I compute, among the top 12 recommended items (i.e., those visible on the initial screen), the share that belong to top-selling category–brand combinations.[40] Figure A.1a reports the share of items that fall within the top 10% of brands by revenue within each category. As expected, the non-personalized condition includes a higher proportion of high-revenue category–brand items than the personalized condition.

In the personalized query-stage experiment, I examine whether the ranking reflects customers' inferred preferences. Section 4.2 presents results for price as an illustrative case; here I focus on brand preferences inferred from users' past transactions. Figure A.1b reports the mean of an indicator equal to one if at least one of the top 12 query results (i.e., the initial screen) matches the user's inferred preferred brand. Customers in the personalized condition are more likely to be shown items from their preferred brands than customers in the non-personalized condition. This pattern is consistent with the ranking incorporating inferred brand tastes.

Figure A.1: Manipulation Checks



(a) Personalizing Discovery Stage

(b) Personalizing Query Stage

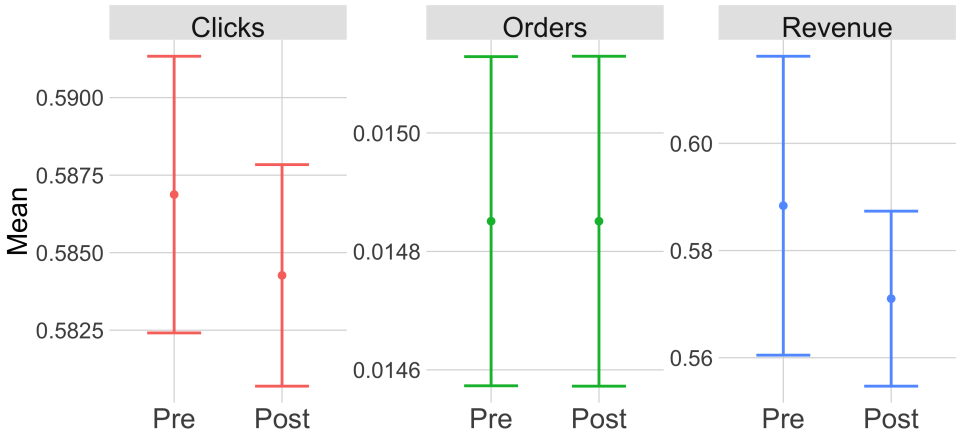
Notes. Panel (a) plots the average share of items among the top 12 discovery products that belong to the top 10% brands by revenue within a category. Panel (b) plots the average of an indicator equal to one if at least one of the top 12 query results matches the user's inferred preferred brand. The analysis uses the initial screen (first 12 items) of the first query session for each user during the experiment. Error bars represent 95% confidence intervals.

[40] Because the experiment was conducted over two years ago at the time of this study, not all discovery-stage recommendations are available due to data storage limitations. Discovery items were recorded only when users generated an interaction event (e.g., clicking a listing or switching tabs), which does not necessarily imply a click on a specific item. I therefore restrict this manipulation check to sessions in which discovery items were recorded. The main treatment-effect estimates using this restricted sample remain consistent with the full-sample results reported in Section 4.1.2.

A.3 Interference Discussion

Personalized Discovery-Stage Experiment SUTVA is violated (i.e., there is "interference") when the treatment assignment of one user influences the outcomes of other users in the experiment. To check potential interference, I focus on users in the control condition and compare their activity before and after the experiment. If there is no interference, control users should exhibit consistent patterns in clicks, orders, and revenue across both periods, because they do not experience any change. Specifically, I compare each control user's behavior in the discovery stage prior to the experiment with their first session following the experiment's launch. Figure A.2 shows no significant differences in these outcomes, thereby mitigating concerns about interference.

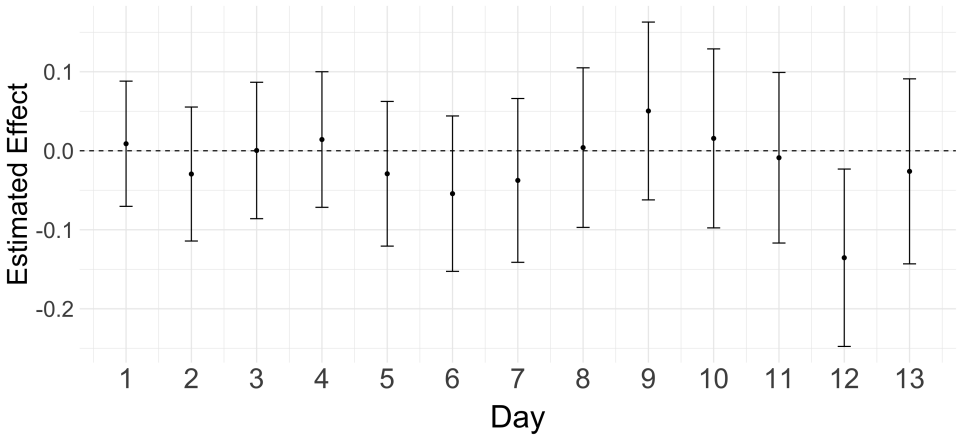
Figure A.2: Comparisons of Users in Control Condition Pre- and Post-Experiment



Notes. This figure compares discovery-stage outcomes for users in the control condition before versus after the experiment, including the number of clicks, the number of orders, and total revenue. The pre-period captures each user's last activity within the two weeks leading up to the experiment, and the post-period corresponds to their first discovery engagement afterward. The analysis includes only users who see product discoveries in both periods. The error bar represents the 95% confidence interval.

In addition, I examine the treatment effects of personalization over time by analyzing results separately by day. If interference were present, I would expect treatment effects to vary across days, as the likelihood of interference increases with more users being assigned to the experiment (e.g., Fong et al., 2026). However, as shown in Figure A.3, the treatment effects remain stable across days, further alleviating concerns about interference.

Figure A.3: Treatment Effects of Personalized Discovery Stage by Day



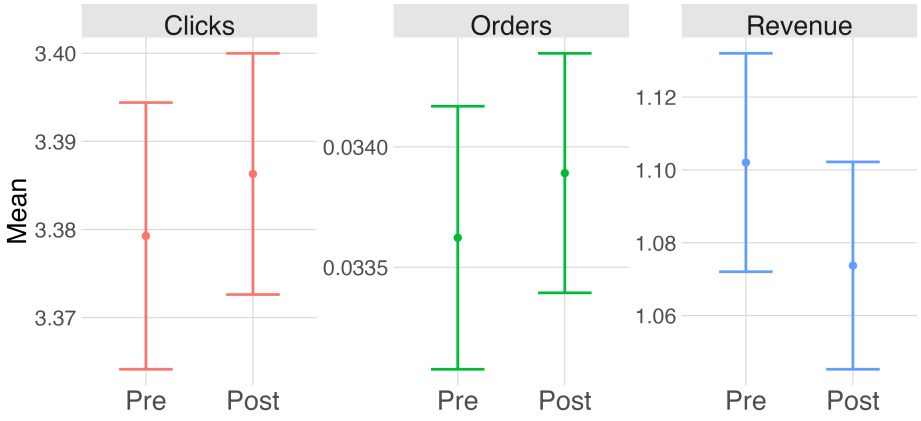
Notes. This figure plots the the estimation results of β^t in the following equation: $\log(Revenue + 1) = \beta_0 + \beta_1 D_t + \beta_2 Personalized_t + \sum_t \beta^t D_t \times Personalized_i + \epsilon_{it}$, where D_t indicates the day. The omitted day is the first day (Day = 0). Standard errors are in parentheses, clustered at the individual level. The error bar represents the 95% confidence interval.

Personalized Query-Stage Experiment First, in this experiment, only a small share of users were assigned to each experimental condition: 10% to the treatment condition and 10% to the control condition. This low saturation rate reduces the likelihood of substantial interference bias (e.g., Fong, 2024).

Second, following the approach above, I test for interference by examining users who are consistently exposed to the status quo ranking algorithm both before and after the experiment. Specifically, I focus on the 70% of users assigned to the status quo ranking in both periods and compare their clicks, orders, and revenue before versus after the experiment. Figure A.4 shows that these outcomes remain stable across periods, providing evidence against interference.

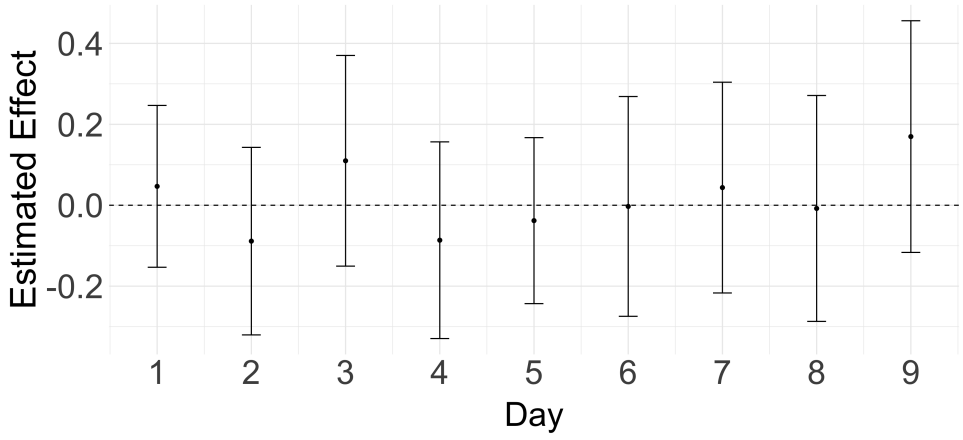
Third, similarly, I estimate the treatment effects of personalized query-stage ranking over days. Figure A.5 plots these estimates for each day during the experiment. None of the effects differ significantly from those on the first day, further minimizing concerns about interference.

Figure A.4: Comparisons of Users in Status Quo Condition Pre- and Post-Experiment



Notes. This figure presents the number of clicks, orders, and total revenue at the query stage for users in the status quo condition. The pre-period reflects each user's last activity within the 12 days preceding the experiment (matching the duration of the personalized query-stage experiment), and the post-period captures their first session. The analysis is restricted to users who conduct a search query in both periods. The error bar represents the 95% confidence interval.

Figure A.5: Treatment Effects of Personalized Query Stage by Day



Notes. This figure plots the the estimation results of β^t in the following equation: $\log(Revenue + 1) = \beta_0 + \beta_1 D_t + \beta_2 Personalized_t + \sum_t \beta^t D_t \times Personalized_i + \epsilon_{it}$, where D_t indicates the day. The omitted day is the first day (Day = 0). Standard errors are in parentheses, clustered at the individual level. The error bar represents the 95% confidence interval.

B Additional Evidence and Discussion

B.1 Spillovers from Discovery Stage to Query Stage

I use the experimental condition in the personalized discovery-stage experiment as an instrumental variable to examine the relationship between the discovery and query stages. The experimental condition exogenously influences outcomes in the discovery stage and is uncorrelated with unobserved factors that affect outcomes in the query stage. Furthermore, it impacts the query stage only indirectly, through its effect on the discovery stage.

I estimate the following two-stage equations:

$$Y_i^D = \kappa_0 + \kappa_1 Personalized_i + \xi_i, \tag{B.1}$$

$$Y_i^Q = \varphi_0 + \varphi_1 \hat{Y}_i^D + \epsilon_i, \tag{B.2}$$

where Y_i^D and Y_i^Q denote customer i's outcomes in the discovery and query stages, respectively, and $\hat{Y}_i^D$ is the predicted value from Equation (B.1). I examine three dependent variables: the number of clicks, orders, and logged revenue. Estimation results are presented in Table B.1. I find that an increase in discovery-stage outcomes (clicks, orders, and revenue) leads to a decline in the corresponding outcomes in the query stage. This pattern suggests a negative spillover effect from the discovery stage to the query stage.

Table B.1: Relationship between Discovery Stage and Query Stage

	Query Clicks (1)	Query Orders (2)	Log(Query Rev) (3)
Constant	4.504***	0.0250***	0.0736***
	(0.0738)	(0.0024)	(0.0074)
Discovery Clicks	-3.120***		
	(0.1723)		
Discovery Orders		-0.7868***	
		(0.2444)	
Log(Discovery Rev)			-0.8287***
			(0.2387)
1-stage F-stats	841.6	43.9	38.3
Observations	3,050,146	3,050,146	3,050,146

Notes. This table shows the estimation results of Equations (B.1) and (B.2). Standard errors are in parentheses, clustered at the individual level. * $p < 0.1$, ** $p < 0.5$, *** $p < 0.01$.

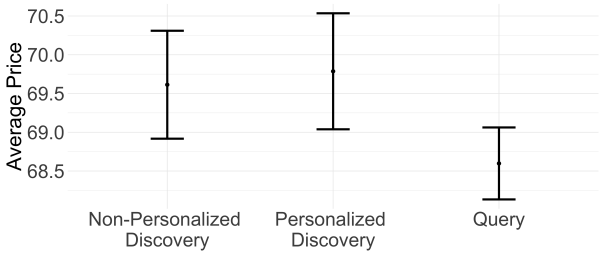
B.2 Discovery versus Query Stages

B.2.1 Price and Item Condition

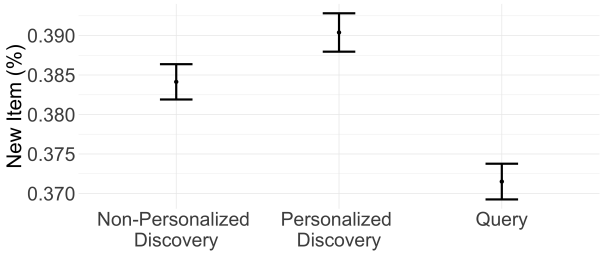
I present the average price and the proportion of new items at each stage. Figure B.1a shows that average prices are similar under personalized and non-personalized discovery stage, and are higher than the prices shown in the query stage. Figure B.1b shows that items shown in the personalized discovery stage are more likely to be new, suggesting that customers tend to prefer new products. In addition, the proportion of new items is higher in the discovery stage than in the query stage.

Figure B.1: Price and Item Condition Across Stages

(a) Average Price



(b) Proportion of New Items



B.2.2 Product Variety

I examine the variety of products presented to customers. I measure product variety using two standard metrics in the literature (e.g., Chen et al., 2024; Holtz et al., 2020): Shannon entropy and the Herfindahl-Hirschman Index (HHI).

Shannon entropy quantifies the dispersion of a distribution. For a discrete random variable Y with support $\{y_1, \ldots, y_n\}$ and associated probabilities $P(y_i)$, entropy is defined as:

$$H(Y) = -\sum_{i=1}^{n} P(y_i) \log_2 P(y_i), \tag{B.3}$$

where the logarithm is base 2. Entropy is maximized under a uniform distribution, which corresponds to maximal variety, and it decreases as the distribution becomes more concentrated.

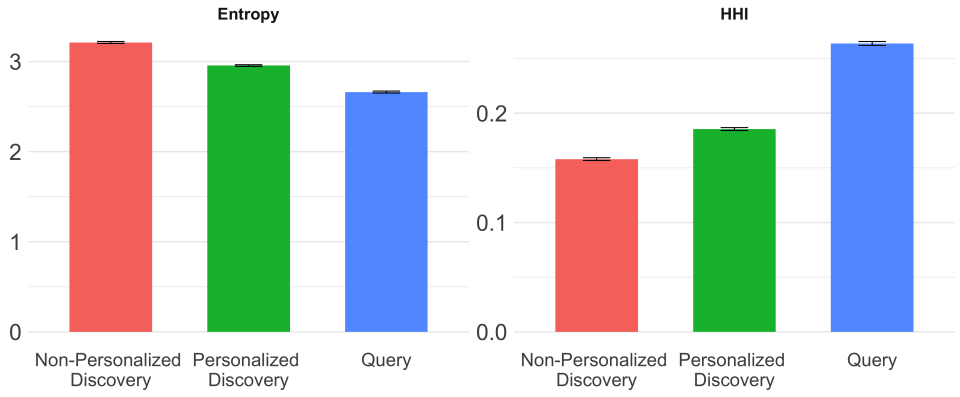
HHI captures concentration by summing the squared probabilities:

$$HHI(Y) = \sum_{i=1}^{n} \left(P(y_i)\right)^2. \tag{B.4}$$

Higher HHI values indicate greater dominance by a cluster and thus lower product variety.

Using Equations (B.3) and (B.4), I compute Shannon entropy and HHI at the session level, focusing on the top 12 items shown on the initial screen for both stages. I define each brand-physical condition combination as a cluster. Figure B.2 presents these two metrics for the personalized discovery stage, the non-personalized discovery stage, and the query stage. As shown, personalizing discovery stage reduces product variety, as indicated by lower entropy and higher HHI. Even with personalization, the discovery stage still exhibits greater variety than the query stage, suggesting that it exposes users to a broader range of products.

Figure B.2: Shannon Entropy and HHI



Notes. This figure presents the Shannon entropy and HHI of the top 12 products, using data from the personalized discovery-stage experiment. Higher entropy and lower HHI indicate greater variety.

B.3 Query Analysis

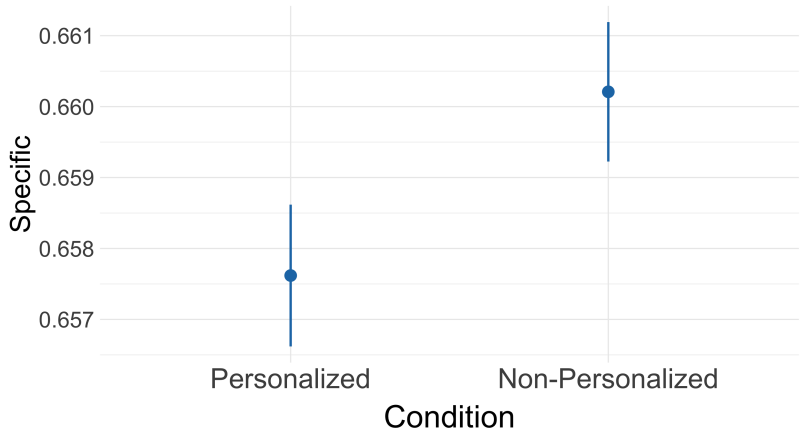
B.3.1 Inferring Demand States via Query Specificity

Yuan et al. (2025) identify two distinct customer states when visiting a platform: demand fulfillment and demand formation. Customers in the fulfillment state have well-defined needs and tend to compensate for less relevant discovery-stage recommendations by submitting more specific queries. In contrast, customers in the demand formation state rely more heavily on product discovery, and reduced relevance at the discovery stage leads them to submit more generic queries. These contrasting patterns offer a lens for inferring whether a customer is in a demand formation or fulfillment state. I adopt this framework to identify customer states.

I examine how personalization at the discovery stage influences the specificity of users' search queries. I follow the rule-based approach outlined in Yuan et al. (2025) to measure query specificity. Each query is decomposed into two types of components: generic words and decoration words. Generic words refer to terms that describe a product's basic type or category (e.g., *jeans*, *headphones*), while decoration words include descriptive modifiers such as brand (*Lululemon*, *LEGO*), color (*yellow*, *red*), physical condition (*new*), material (*cotton*, *leather*), or size (*plus size*). I compile a comprehensive dictionary of decoration terms derived from the product attributes available on the platform. A query is classified as *generic* if it contains only generic words, and as *specific* if it includes at least one decoration word. For example, the query *"black UGG boots"* contains a generic term (*boots*) along with decoration words (*UGG* as the brand and *black* as the color), and is categorized as specific.

In Figure B.3, I present query specificity for users in the personalized and non-personalized conditions. The figure shows that users in the non-personalized condition submit more specific queries. This pattern suggests that many customers arrive with well-defined demand and compensate for less personalized discovery-stage recommendations by refining their query inputs.

Figure B.3: The Impact of Personalizing Discovery Stage on Search Query Specificity

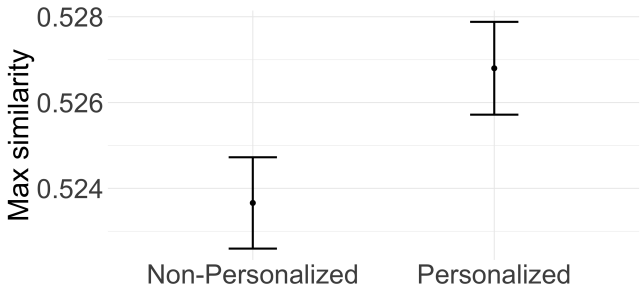


Notes. This figure presents the query specificity in the personalized and non-personalized conditions from the personalizing discovery-stage experiment. The error bar represents the 95% confidence interval.

B.3.2 Correlation between Discovery Products and Query

I examine how closely the products shown in the discovery stage align with the query a customer submits later in the same session. To quantify this alignment, I compute a semantic similarity score between each discovery-stage item's title and the session's query using Sentence-BERT embeddings and cosine similarity. Figure B.4 shows that the maximum similarity is higher under the personalized condition than the non-personalized condition. This pattern suggests that a broader product assortment does not increase customer interest or trigger spontaneous query initiation.

Figure B.4: Correlation between Discovery Products and Query by Condition



Notes. This figure presents the correlation between discovery-stage products and the subsequent query in the personalized and non-personalized conditions using data from the discovery-stage personalization experiment. The error bar represents the 95% confidence interval.

B.4 Stratification by Discovery-Stage Purchase

To examine whether and how users continue into the query stage differently depending on whether the discovery stage offers a good match, I stratify users by whether they purchase at the discovery stage and compare outcomes between the personalized and non-personalized conditions within each stratum. Table B.2 reports effects on query initiation and total revenue across both stages. Among users who purchase at the discovery stage, personalization reduces the likelihood of initiating a query, with no corresponding change in total revenue. Among users who do not purchase at the discovery stage, personalization also reduces query initiation and lowers total revenue. These results suggest that when discovery-stage personalization fails to provide a good match, customers are more likely to exit rather than continue the journey.

Table B.2: Personalizing Discovery Stage on Query and Total Revenue by Stratification

	Non-Personalized	Personalized	Change (%)	t-statistic	p-value
Purchase at Discovery Stage					
Whether Query	0.321	0.308	-4.189	-2.442	0.015
Total Revenue	43.578	43.809	0.526	0.375	0.708
No Purchase at Discovery Stage					
Whether Query	0.580	0.556	-4.130	42.046	< 0.001
Total Revenue	0.568	0.544	-4.134	2.370	0.018

Notes. This table reports the effects of personalizing the discovery stage on query initiation and total revenue, stratified by whether the customer makes a purchase at the discovery stage. I report the mean values of outcomes for the control condition and the treatment condition. The difference reflects the mean difference between the treatment and control conditions. The final two columns report t-statistics and p-values from two-tailed tests for differences in means.

C Ranking Model at the Query Stage

In the re-ranking stage, I employ LambdaMART (Burges, 2010), a widely used pairwise learning-to-rank algorithm that combines the LambdaRank framework with Multiple Additive Regression Trees (MART). LambdaMART leverages gradient-boosted decision trees and uses a cost function derived from LambdaRank to optimize ranking performance. As one of the most effective learning-to-rank algorithms, LambdaMART has been widely adopted in both academic research and industry applications. In marketing literature, Yoganarasimhan (2020) finds that LambdaMART outperforms alternative ranking models in terms of improvements in Normalized Discounted Cumulative Gain (NDCG) metrics. In industry, LambdaMART powers Microsoft's Bing search engine.[41]

Training Data The training data include item-level features, contextual features, and user-level features (for the personalized model). Item-level features capture static attributes (e.g., category, brand) and recent interaction metrics over the past 30 days (e.g., cumulative likes). Contextual features capture market conditions at the time a query is submitted, such as real-time inventory for relevant products. User-level features include behavioral variables (e.g., item views in the past 30 days) and demographic information (e.g., inferred age). The training sample comprises all query-stage activity on the platform over a two-week period.

LambdaMART predicts item relevance via a ranking score. I construct this score from user interactions (e.g., purchases, likes), placing greater weight on higher-intent actions such as purchases. The weights are calibrated using historical user activity data.

Training Process The model is trained by tuning several hyperparameters to balance flexibility and overfitting. These include regularization terms (which control model complexity), the learning rate (which determines the step size during optimization), and the number of leaves per tree (which affects the model's ability to capture interaction effects). I also tune the maximum tree depth, the minimum number of observations per leaf, the number of estimators, the bagging fraction, and the feature-subsampling ratios. The model is trained to optimize NDCG over the top 100 positions. For evaluation, I split the data into 80% for training and 20% for testing.

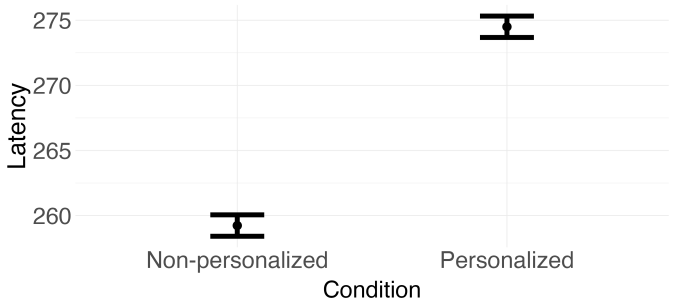
[41] Microsoft Research, "RankNet: A Ranking Retrospective," `https://www.microsoft.com/en-us/research/blog/ranknet-a-ranking-retrospective/`, accessed March 2026.

D Latency

The Effect of Personalization on Latency As shown in Figure D.1, in the personalized query-stage experiment, average latency is higher in the personalized condition (274 milliseconds) compared to the non-personalized condition (259 milliseconds). Based on the discussion with the company, this increase is primarily driven by the additional time required to retrieve and compute user-specific features.[42]

Since the conclusion of this experiment, the company has significantly expanded its infrastructure investments to improve system performance. These latency results are thus specific to my experimental period. However, they also highlight another cost of personalization—the continual infrastructure investments required to maintain responsiveness at scale.

Figure D.1: The Impact of Personalized Query Stage on Latency



Notes. This figure compares latency under personalized and non-personalized conditions from the personalized query-stage ranking experiment. The error bar represents the 95% confidence interval.

The Effects of Latency on Customers From the customer's perspective, increased latency can lead to longer loading times, potentially degrading the overall shopping experience. To quantify the relationship between backend latency and observed loading time, I estimate the following regression:

$$Y_k = \iota_0 + \iota_1 latency_k + \epsilon_k, \tag{D.1}$$

where $latency_k$ denotes the backend latency (in milliseconds, log-transformed) for session k, and Y_k represents the outcome of interest. This analysis uses data from the personalized query-stage

[42] As the personalized discovery-stage experiment was conducted more than two years prior to my data access, backend latency logs from that period are no longer available.

experiment. I first consider Y_k as the customer-observed loading time for session k. Since clickstream data is recorded at the second level, observed loading times are rounded to the nearest full second.[43] 37.8% of sessions have an observed loading time of 0 seconds, and 55.8% register a loading time of 1 second.[44] The estimated coefficient on latency, $\hat{\iota_1} = 0.20$ (standard error = 0.05), indicates a statistically significant and positive relationship: longer backend latency is associated with longer observed loading times.

Next, I investigate the impact of latency on consumer engagement by estimating Equation (D.1) with three outcome variables: the number of clicks, the number of orders, and logged revenue per session. In Table D.1, a 1% increase in latency is associated with a 4.6% decrease in clicks, a 0.4% decrease in orders, and a 0.1% decrease in revenue. These findings demonstrate the economic significance of minimizing latency in personalization systems.

Table D.1: Effects of Search Query Latency

	Clicks (1)	Orders (2)	Log(Rev) (3)
Constant	4.101***	0.0703***	0.2149***
	(0.1420)	(0.0091)	(0.0267)
Log(Latency)	-0.0464*	-0.0038**	-0.0130***
	(0.0258)	(0.0016)	(0.0048)
Observations	101,419	101,419	101,419
R^2	0.00004	0.0001	0.0001

Notes. This table reports the estimated effects of backend search latency (log-transformed) on consumer engagement in the query stage. Standard errors are in parentheses, clustered at the individual level. * $p < 0.1$, ** $p < 0.5$, *** $p < 0.01$.

[43] The latency dataset is not directly linked to individual user sessions, so I match latency to customer sessions based on timestamp and exclude ambiguous cases (e.g., concurrent queries by different users at the same timestamp).
[44] Sessions with loading times above 3 seconds (fewer than 1%) are excluded as outliers.

E Robustness Checks of Alternative Algorithm Designs

E.1 Alternative Discovery-Stage Model

I evaluate an alternative discovery-stage recommendation model because the "quality" of recommendations may influence consumer decisions. To do so, I draw on a separate field experiment run by the company that tested a more advanced personalized recommendation algorithm on the discovery page. Relative to the baseline model, this approach represents a substantial upgrade along several dimensions. It uses a transformer-based architecture that incorporates visual product representations (e.g., images) and a richer set of user engagement signals. It also introduces a forecasted-serendipity component that predicts future user actions and tailors recommendations with greater contextual relevance. Consistent with these design changes, the new model improves standard ranking metrics, including recall, precision, and NDCG.

The experiment evaluating this enhanced algorithm was conducted over a two-week period in Spring 2023. During the experiment, 50% of users were randomly assigned to the status quo personalized discovery-stage algorithm (the treatment condition in Section 4.1), while the remaining 50% of users received the more advanced algorithm at the discovery stage.

I report the experimental results in Table E.1. I find that customers exposed to a more advanced algorithm click on 45.5% more items, place 17.4% more orders, and generate 16.2% more revenue within the discovery stage compared to those in the status quo condition. However, they are 1.6% less likely to initiate a query, click on 2% fewer items, place 3.7% fewer orders, and generate 2.8% less revenue in the query stage. The total orders and revenue across two stages remain unchanged.

E.2 Alternative Ranking Model

The effectiveness of personalized ranking also depends in part on the "quality" of the underlying ranking algorithm. For instance, if the baseline non-personalized algorithm is already highly optimized, the marginal benefit of personalization may be limited. In contrast, greater gains may arise when the baseline algorithm has more room for improvement.

I use the status quo ranking model as the base algorithm and develop a personalized counterpart that additionally incorporates user-specific features. Recall that in the personalized search experiment, among the remaining 80% of users not included in the experimental conditions, 70%

Table E.1: Treatment Effects of More Advanced Personalized Discovery Stage

Variable	Personalized	More Personalized	Change (%)	t-statistic	p-value
Discovery Clicks	0.216	0.247	14.245	19.799	< 0.001
Discovery Orders	0.005	0.006	20.132	9.904	< 0.001
Discovery Revenue	0.230	0.284	23.614	5.858	< 0.001
I(Initiate a Query)	0.600	0.592	-1.322	-12.595	< 0.001
Query Clicks	3.427	3.366	-1.796	-8.096	< 0.001
Query Orders	0.038	0.037	-3.823	-5.156	< 0.001
Query Revenue	1.691	1.651	-2.323	-1.721	0.085
Total Clicks	3.643	3.612	-0.846	-3.986	< 0.001
Total Orders	0.043	0.043	-1.056	-1.502	0.133
Total Revenue	1.921	1.936	0.783	0.607	0.544

Notes. This table presents the treatment effects of a more advanced personalization algorithm at the discovery stage. I report the mean values of outcomes under both the control condition (baseline personalization) and the treatment condition (personalization using a more advanced algorithm). The reported difference reflects the mean outcome in the treatment condition minus that in the control condition. The final two columns display the corresponding t-statistics and p-values from two-tailed tests for differences in means.

were randomly assigned to a non-personalized status quo ranking model and 10% were assigned to a personalized version of the same model that integrates user-specific features.

Table E.2 reports the treatment effects of personalization under the status quo ranking algorithm. Personalization leads to statistically significant improvements across all query-stage outcomes: clicks increase by 1.7%, orders by 6.9%, and revenue by 7.3%. These results convey two main takeaways. First, the choice of ranking algorithm itself influences user engagement and platform revenue. For example, the LambdaMART and status quo models produce different revenue outcomes. Second, holding the algorithm constant, incorporating user-level data can enhance overall revenue.

Table E.2: The Effects of Personalized Query Ranking (Alternative Model)

Variable	Non-Personalized	Personalized	Change (%)	t-statistic	p-value
Search Clicks	2.464	2.506	1.670	3.209	0.001
Search Orders	0.025	0.027	6.912	3.799	< 0.001
Search Revenue	0.866	0.929	7.334	2.220	0.026
Rec Clicks	0.821	0.824	0.392	0.565	0.572
Rec Orders	0.002	0.002	-2.873	-0.447	0.655
Rec Revenue	0.045	0.050	11.195	0.775	0.438
Total Clicks	3.285	3.329	1.350	3.182	0.001
Total Orders	0.027	0.029	6.324	3.601	< 0.001
Total Revenue	0.911	0.980	7.526	2.337	0.019

Notes. This table presents the treatment effects of personalized query rankings based on the status quo algorithm. I report the mean values of outcomes for the control condition (non-personalized status quo model) and the treatment condition (personalized status quo model). The difference reflects the mean difference between the treatment and control conditions. The final two columns report t-statistics and p-values from two-tailed tests for differences in means.

F Belief Estimation

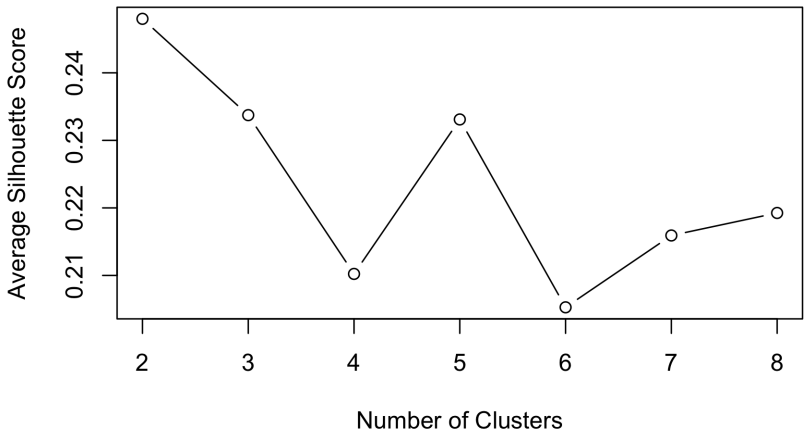
F.1 Conducting a Query

Beliefs about Query Stage. To improve the precision of the estimated empirical distribution, I group similar queries using natural language processing techniques. Specifically, I first embed every query string into a TF-IDF vector space over unigrams and bigrams. I then run k-means clustering on those embeddings. Next, I choose the number of clusters by maximizing the average silhouette score under a cosine-distance metric. The silhouette score measures how similar an object is to its own cluster compared to other clusters. For example, "PS5" and "PlayStation 5" belong to the same query group. These procedures ultimately give me a total of 39 query groups.

I then empirically estimate the distribution of product attributes for each query group across sessions. For customers in the non-personalized condition, expectations are formed based on the empirical distribution of products observed across all non-personalized users. For those in the personalized condition, I segment them based on their historical activities, including preferences (e.g., price level, brand), activity levels (e.g., item clicks in the past 30 days), purchase histories, and inferred demographic age. Customers' expectations are assumed to follow the empirical distribution of products seen by customers within their respective segments. To group customers by segments, I again apply k-means clustering to group individuals in the personalized condition based on observed user attributes and use the silhouette scores to assess the quality of the clustering. I iterate the number of clusters from 2 to 8 and find that the silhouette score is maximized when the segment is 2 (Figure F.1). Based on the inferred segments, customers in one group are generally more active, with greater past engagement (e.g., item clicks, queries, and purchases), while those in the other group are less active. For example, the median number of purchased products since account creation is 7 for the less active segment and 49 for the more active segment. I would expect the optimal number of clusters to be larger when considering all customers on the platform.

Finally, to reduce computational burden, I follow Koulayev (2014) and Ghose et al. (2019) and use a bootstrap sampling procedure. Specifically, in each draw I sample 48 products from the empirical distribution, conditional on the query and the personalization condition (and, in the

Figure F.1: Silhouette Score of Clusters



personalized condition, conditional on the segment as well).[45]

Beliefs based on Discovery Stage Recall that when the discovery-stage products do not match the customer's intent, the customer evaluates whether to query based on the expected payoff from the query-stage environment. When the discovery-stage products match the customer's intent, I treat the discovery stage as an additional signal about the latent payoff u_i^q. I define intent-match if at least three products in the discovery stage match the query submitted by the customer.[46] When intent matches, customers may expect the query stage to display products similar to those shown in the discovery stage. To capture this, I construct an empirical distribution of discovery-stage products that approximates what a customer could plausibly encounter at the query stage. Specifically, I pool products shown in the discovery stage across sessions and stratify them by position. Within each position, I further restrict attention to products with similar attributes, yielding a position-specific empirical distribution of comparable discovery-stage items. I then apply a bootstrap procedure analogous to the one described above: in each draw, I construct a synthetic discovery by independently sampling one product for each position from the corresponding position-specific empirical distribution.

[45] The top 48 products account for 74% of clicks and 80% of purchases.
[46] I also try definitions of intent-match based on one or two products, and the results are similar.

F.2 Inspecting a Product

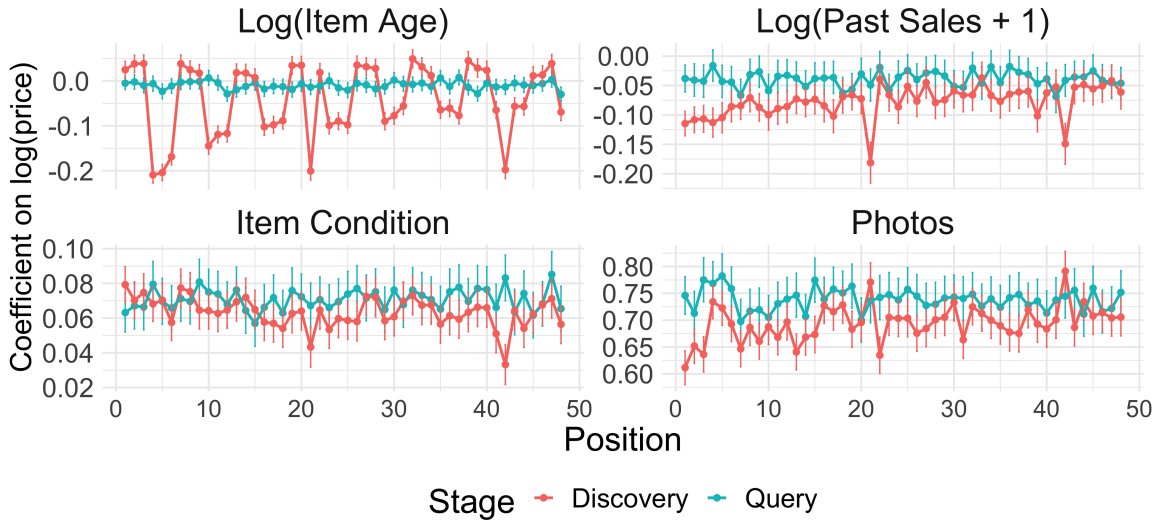
I estimate the following equation to examine the relationship between a product's hidden attributes and its visible attributes:

$$Z_j = \phi_0 + \phi_1 \log(price_j) + category_j + brand_j + \eta_j, \tag{F.1}$$

where Z_j are the hidden attributes, including the number of photos, item physical condition (measured on a scale from 1 to 5, where 1 indicates poor condition and 5 indicates new), item age and its seller's past sales. The visible attributes include price, category, and brand. The error term η_j is assumed to have a mean of zero.

Figure F.2 illustrates the estimation results by plotting the estimated (logged) price coefficients. The results suggest that higher-priced items tend to include more photos, be in better physical condition, and be listed by less experienced sellers (i.e., those with fewer past sales). In addition, the coefficients exhibit variation across both stages and positions.

Figure F.2: Beliefs about Hidden Attributes



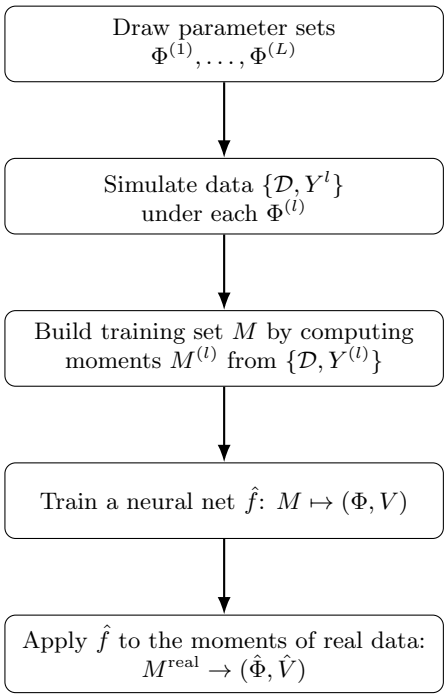
Notes. This figure displays the estimated price coefficients from Equation (F.1) for the discovery stage and query stage respectively. Error bars represent 95% confidence intervals.

G Search Model Estimation

G.1 Model Estimation using Neural Networks

I estimate the structural search model using the NNE approach proposed by Wei and Jiang (2025), which leverages a neural network to learn a mapping from datasets to the structural parameters. Figure G.1 illustrates the steps to train a neural network to recover the parameter vector.

Figure G.1: Neural Network Training Illustration



The process can be described as follows.

1. **Draw Parameter Sets:** Generate L independent draws of model parameters $\mathbf{\Phi}^{(1)}, \ldots, \mathbf{\Phi}^{(L)}$ uniformly from a predefined parameter space. In my application, I set $L = 20{,}000$.

2. **Simulate Data:** For each draw $\mathbf{\Phi}^{(l)}$, simulate consumer behavior using the structural search model to generate the data $\{\mathcal{D}, Y^{(l)}\}$. The dataset $\mathcal{D}$ includes item attributes $(\boldsymbol{X}, \boldsymbol{Z})$ and cost variables (Pos, $Personalized$). The simulated outcomes $Y^{(l)}$ include inspection, query submission, and purchase.

3. **Compute Moments:** From each simulated dataset $\{\mathcal{D}, Y^{(l)}\}$, compute a vector of data moments $M^{(l)}$ that summarize the key data patterns. I construct the following data moments

at the option-level (including items and query), session-level, and individual-level: means and variances of $Y^{(l)}$ and cross-covariances between $\mathcal{D}$ and $Y^{(l)}$. As suggested by Wei and Jiang (2025), NNE is robust to redundant moments, because the neural net can learn from the training set which moments are informative about the parameters.

4. **Train a Neural Network:** Collect all moment-parameter pairs $(M^{(l)}, \mathbf{\Phi}^{(l)})_{l=1}^{L}$ into a training set, where M is an input to the neural net and $\mathbf{\Phi}$ is the output. The neural network is trained to learn the inverse mapping from empirical data moments M to model parameters $\mathbf{\Phi}$ and covariance matrix $\boldsymbol{V}$. Following Wei and Jiang (2025), I use a cross-entropy loss to teach the network to map from M back to $(\mathbf{\Phi}, \boldsymbol{V})$. The neural network architecture is defined using a sequence of layers. The input layer normalizes the input data using symmetric rescaling. This is followed by a fully connected layer with 64 hidden nodes and a ReLU activation function. The output layer is another fully connected layer, with the number of outputs matching the dimensionality of the output data. The initial learning rate is set to 0.01 and updated using a piecewise schedule, decreasing every 1,000 iterations. The network is trained using the Adam optimizer over a maximum of 500 epochs with a mini-batch size of 500 samples. I use a 90/10 split for training and testing, and model validation is conducted every 500 iterations. As demonstrated by Wei and Jiang (2025), the accuracy of NNE is not sensitive to the specific architecture or hyperparameter choices of the neural network. I experiment with different hyperparameter settings and find very consistent predictive results.

5. **Apply to Real Data:** Compute the moments M^{real} on actual data $\{\mathcal{D}, Y^{\text{real}}\}$. Feed M^{real} into the trained network to obtain both the point estimate $\hat{\mathbf{\Phi}}$ and the statistical accuracy for the point estimate $\hat{\boldsymbol{V}}$.

G.2 Monte Carlo Simulation

I generate a dataset consisting of 3,000 customers and each customer has three sessions. In each session, customers follow a two-stage sequential search process: first a discovery stage, then a query stage. Each stage displays 24 products. Customers submit one of the two queries available. Customers are randomly assigned, with equal probability, to either a personalized query result or a non-personalized query result and are categorized into one of two segments (Segment 1 or Segment

2). For simplicity, product attributes and rankings are randomly generated.

Each product has three visible attributes: price, category and brand, and four hidden attributes: number of photos, product age, item condition, and sellers' past sales. Logged prices are drawn from a normal distribution: $\log(\text{Price}) \sim \mathcal{N}(3.5, 1)$. Category is a dummy variable indicating whether the item belongs to the same category as the customer's purchase intent. The probability that an item is in the relevant category is 0.3 in the discovery stage and 0.9 in the query stage. There are five brands, each with the following probabilities of being chosen in the discovery stage: 0.05, 0.15, 0.15, 0.20, and 0.25. In the query stage, the corresponding probabilities are 0.10, 0.02, 0.10, 0.05, and 0.08. The remaining probability mass in each stage corresponds to unbranded items. The number of photos is drawn from a discrete distribution over $\{1, 2, 3, 4, 5\}$ with probabilities $P = (0.10, 0.05, 0.25, 0.40, 0.30)$, respectively. Item condition is also sampled from a discrete distribution over $\{1, 2, 3, 4, 5\}$. Logged product age follows $\mathcal{N}^+(4, 1)$ and sellers past sales (logged) follow $\mathcal{N}^+(0, 1)$; both are normal distributions truncated below at zero.

Consumer utility depends on both visible and hidden product attributes. I follow the procedures described in Section 5.2.2 to estimate customers' beliefs about product utility conditional on visible attributes and beliefs about the utility of conducting a query. Customers also face two types of costs discussed in Section 5.2.3: (1) an inspection cost that varies with a product's position in the list and (2) a query cost that depends on the personalization condition.

Customers begin their decision process in the discovery stage, choosing whether to inspect an item, conduct a search query, or exit (i.e., select the outside option). If they conduct a query, they exit the discovery stage and proceed to the query stage, where they view a new set of 24 products related to the query they submit and decide which to inspect. Ultimately, customers choose to purchase the most preferred product among all inspected products across both stages or opt for the outside option.

Table G.1 reports the results of the estimation. Column (1) lists the true parameter values, and Column (2) presents the corresponding estimates from the model. Overall, the estimated parameters align closely with the true values.

Table G.1: Monte Carlo Simulation Results

Variables	(1) True	(2) Estimated
Preferences		
Mean Preferences		
Constant	-3.0	-3.098 (0.066)
Price	-0.4	-0.443 (0.033)
Same Category	0.8	0.828 (0.020)
Brand 1	0.4	0.412 (0.038)
Brand 2	0.4	0.448 (0.048)
Brand 3	0.5	0.552 (0.057)
Brand 4	0.3	0.349 (0.089)
Brand 5	0.1	0.141 (0.036)
N Photos	0.3	0.292 (0.090)
Item Age	-0.2	-0.238 (0.011)
Physical Condition	0.3	0.293 (0.019)
Seller Past Sales	0.3	0.286 (0.010)
Inspection Base (Discovery)	-3.0	-3.110 (0.055)
Position	1.0	0.964 (0.064)
Inspection Base (Query)	-4.0	-3.841 (0.256)
Query Base	-1.5	-1.487 (0.027)
Personalization	0.8	0.879 (0.126)
Preferences Heterogeneity (SD)		
Price	0.2	0.200 (0.058)
N Photos	0.1	0.094 (0.062)
Item Age	0.1	0.100 (0.057)
Physical Condition	0.1	0.094 (0.058)
Seller Past Sales	0.1	0.114 (0.065)
Costs		
Inspection Base (Discovery)	-3.0	-3.260 (0.281)
Position	1.0	0.964 (0.137)
Inspection Base (Query)	-4.0	-3.941 (0.256)
Query Base	-1.5	-1.587 (0.244)
Personalization	0.8	0.829 (0.019)

Notes. The estimation is based on 3,000 customers, each with 3 sessions (9,000 sessions in total). Standard errors are in parentheses.

G.3 Model Fit

To evaluate model fit, I compare predicted and observed customer outcomes using parameter-free moments: the average number of inspections per session, the percentage of sessions with a search query, and the percentage of sessions with a purchase. Table G.2 shows that the model captures these moments well.

Table G.2: Model Fit

Moments	Predicted	Observed
Session level		
Inspections	1.614	1.672
Query Percentage	0.847	0.851
Orders	0.031	0.031
Individual level		
Inspections	4.192	4.326
Orders	0.067	0.070

H Utility-Based Ranking Models

At each stage, the platform ranks candidate items for customer i using a stage-specific match index,

$$\nu_{ij}^{g} \;=\; \omega^{g} u_{ij} \;+\; (1-\omega^{g})\bar{u}_{j}, \qquad g \in \{D, Q\}, \tag{H.1}$$

where $g = D$ denotes the discovery-stage product list and $g = Q$ denotes the query-stage product list. Individual-level utility is $u_{ij} = \boldsymbol{X}_j'\boldsymbol{\alpha}_i + \boldsymbol{Z}_j'\boldsymbol{\beta}_i$, and $\bar{u}_j = \boldsymbol{X}_j'\bar{\boldsymbol{\alpha}} + \boldsymbol{Z}_j'\bar{\boldsymbol{\beta}}$ is the utility implied by mean preference coefficients across customers. Equivalently, $\bar{u}_j$ can be interpreted as a baseline score that captures demand on average and thus proxies for popularity. The parameter $\omega^g \in [0, 1]$ governs the degree of personalization at stage g: larger values place more weight on individual tastes, while smaller values place more weight on average tastes. In particular, $\omega^g = 1$ corresponds to a fully personalized ranking, $\omega^g = 0$ to a non-personalized ranking, and intermediate values to partial personalization. In the simulations, I implement non-personalized, partially personalized, and fully personalized variants by setting $\omega^g \in \{0, 0.5, 1\}$ at the relevant stage.

At the discovery stage, the non-personalized condition selects products from the cohort of items shown to customers who arrive around the same time, which approximates the market's "popular" products at the moment of arrival. In contrast, the personalized condition selects products from a candidate set constructed using customers with similar historical activity. In both conditions, the products are selected and ranked according to Equation (H.1). At the query stage, the platform first retrieves up to 100 products relevant to the user's query. These retrieved products are then re-ranked according to Equation (H.1).

I Additional Tables and Figures

Table I.1: User-Level Treatment Effects of Personalization at Discovery Stage

Variable	Non-Personalized	Personalized	Change (%)	t-statistic	p-value
Discovery Clicks	0.421	0.490	16.545	52.351	< 0.001
Discovery Orders	0.011	0.012	6.044	7.914	< 0.001
Discovery Revenue	0.537	0.565	5.044	3.758	< 0.001
Query Clicks	3.100	2.956	-4.632	-27.764	< 0.001
Query Orders	0.0224	0.0216	-3.665	-6.674	< 0.001
Query Revenue	0.789	0.750	-5.024	-5.315	< 0.001
Total Clicks	3.517	3.443	-2.115	-13.802	< 0.001
Total Orders	0.033	0.033	-0.322	-0.687	0.492
Total Revenue	1.306	1.294	-0.962	-1.159	0.246

Notes. The table shows the treatment effects of personalized discovery stage on outcomes at the user level. I take the mean outcomes across sessions in the experiment for each user. The difference is measured using the outcomes in the personalized condition minus the outcomes in the non-personalized condition. The last two columns report t-statistics and p-values from the corresponding two-tailed tests for differences in means.

Table I.2: User-Level Treatment Effects of Personalization at Query Stage

Variable	Non-Personalized	Personalized	Change (%)	t-statistic	p-value
Discovery Clicks	0.791	0.793	0.255	0.320	0.749
Discovery Orders	0.002	0.002	0.963	0.163	0.871
Discovery Revenue	0.042	0.042	-0.582	-0.058	0.954
Query Clicks	2.242	2.303	2.708	5.132	< 0.001
Query Orders	0.023	0.025	9.222	4.775	< 0.001
Query Revenue	0.812	0.873	7.582	2.264	0.024
Total Clicks	3.032	3.094	2.048	4.595	< 0.001
Total Orders	0.025	0.027	8.680	4.674	< 0.001
Total Revenue	0.854	0.915	7.148	2.212	0.027

Notes. The table shows the treatment effects of personalized query stage on outcomes at the user level. I take the mean outcomes across sessions in the experiment for each user. The difference is measured using the outcomes in the personalized condition minus the outcomes in the non-personalized condition. The last two columns report t-statistics and p-values from the corresponding two-tailed tests for differences in means.

Table I.3: Treatment Effects of Personalized Query Stage on Discovery-Stage Outcomes and Total Outcomes

Variable	Non-Personalized	Personalized	Change (%)	t-statistic	p-value
Discovery Clicks	0.846	0.844	-0.300	-0.241	0.81
Discovery Orders	0.002	0.002	-1.271	-0.140	0.888
Discovery Revenue	0.045	0.045	0.968	0.060	0.952
Total Clicks	3.086	3.175	2.874	5.173	< 0.001
Total Orders	0.027	0.030	9.549	4.144	< 0.001
Total Revenue	0.883	0.982	11.231	3.098	0.002

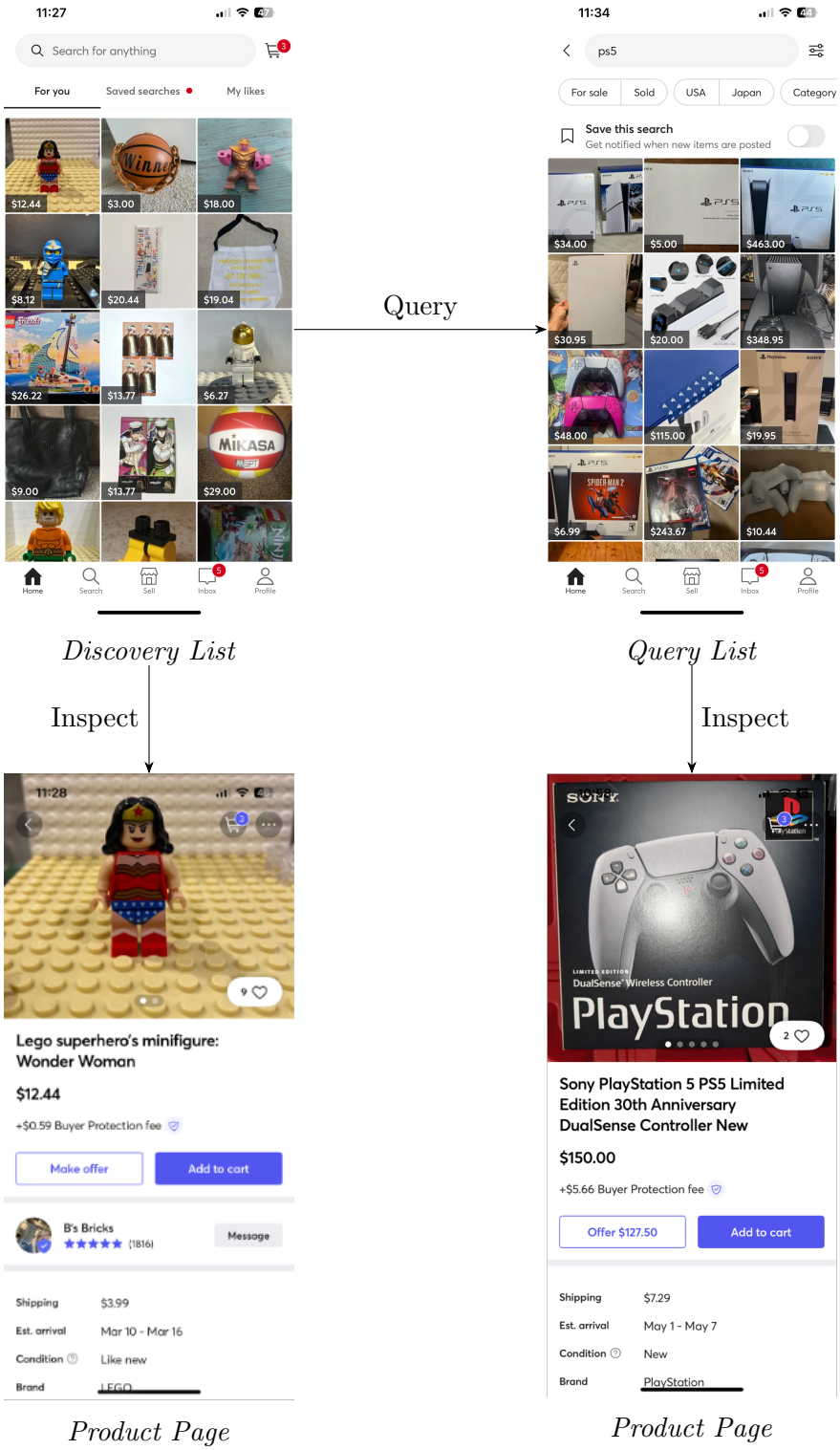
Notes. This table shows the treatment effects of the personalized query stage. I report the mean of outcomes in both personalized and non-personalized conditions. The difference is measured using the outcomes in the personalized condition minus the outcomes in the non-personalized condition. The last two columns report t-statistics and p-values from the corresponding two-tailed tests for differences in means.

Table I.4: Effects of Refinement on Customer Outcomes

	Orders	
	(1)	(2)
Constant	0.0459***	0.0477***
	(0.0008)	(0.0009)
Sorted	0.0443***	
	(0.0071)	
Personalized	0.0031***	0.0029**
	(0.0011)	(0.0012)
Sorted × Personalized	0.0028	
	(0.0093)	
Filtered		-0.0024
		(0.0024)
Filtered × Personalized		0.0018
		(0.0033)
Observations	187,818	187,818
R^2	0.0011	0.00005

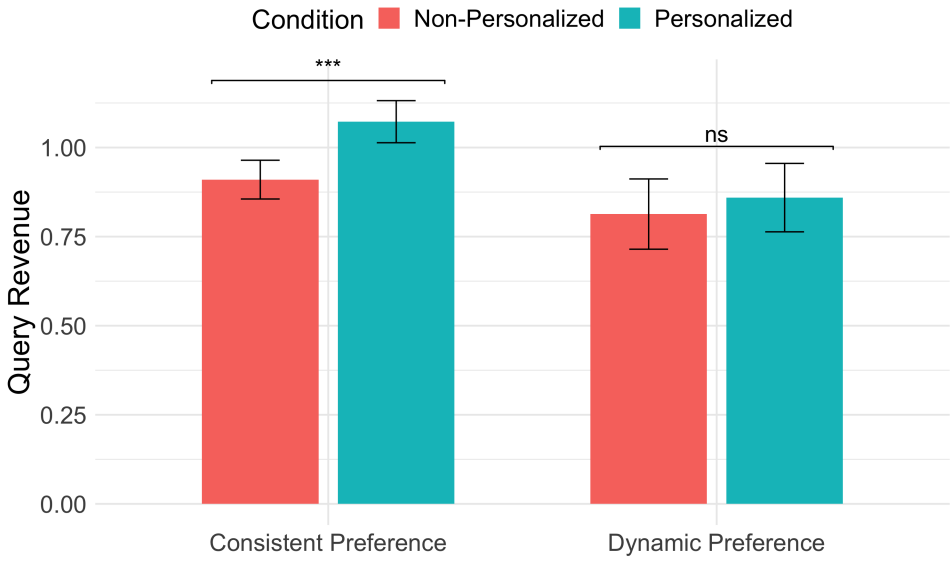
Notes. This table shows the effects of applying sorting and filters on the customer's purchase decisions. Standard errors are in parentheses, clustered at the individual level. * $p < 0.1$, ** $p < 0.5$, *** $p < 0.01$.

Figure I.1: An Illustration of Platform Layout



Notes. This figure shows an example of product display on Mercari in the discovery and query stages. In both lists, customers see each product's image and price. They can inspect a product by clicking on it, which takes them to the product page.

Figure I.2: Heterogeneous Effects of Personalizing Query Stage

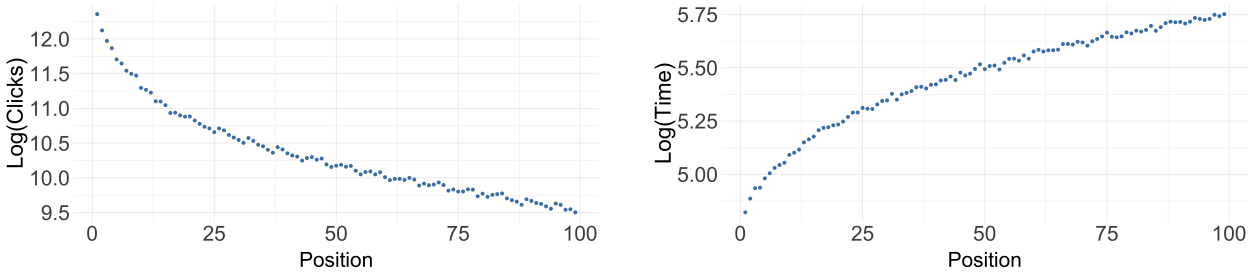


Notes. This figure plots the average revenue at the query stage for customers in the personalized and non-personalized conditions, separately for those with consistent and dynamic preferences. The sample is restricted to customers who clicked more than one item in the 30 days prior to the personalized query stage experiment. Error bars represent 95% confidence intervals.

Figure I.3: Clicks and Time by Item Position

(a) Clicks

(b) Time (seconds)



Notes. Panel (a) presents the logged number of clicks by query-result position; panel (b) presents the logged average time (in seconds) from the start of the session to the click.