Two-Sided Prioritized Ranking: A Coherency-Preserving Design for Marketplace Experiments

(Click here for the latest version)

Job Market Paper of Zahra Khanalizadeh[1]

Based on joint work with

Mahyar Habibi[2] and Negar Ziaeian[3]

May 2025

Abstract

In online two-sided marketplaces, maintaining coherency in observable item attributes, such as prices, is crucial for providing users with a consistent platform experience. For certain interventions, notably those involving price adjustments, it is vital that all users see identical realizations of treatment to prevent confusion or perception of unfairness. However, this constraint complicates causal estimation of treatment effects due to interference between marketplace units. To address this challenge, we propose a novel experimental design called Two-Sided Prioritized Ranking (TSPR), which leverages the marketplace's recommender system as an experimental instrument. TSPR strategically prioritizes items based on their treatment status while ensuring that all users have coherent exposure to treatments. Through simulations using search impression data

[1]Ph.D. Candidate in Economics, University of Washington, USA. Email: zkhnl@uw.edu. I am deeply grateful to my advisors Alan Griffith and Jason Kerwin, as well as Melissa Knox, for their invaluable guidance and support. We thank Avi Goldfarb, Sadegh Shirani, Ludovica Gazze, Manuel Bagues, and Mohammad H. Seyedsalehi for helpful discussions and suggestions. We also appreciate feedback from participants at the 2024 Conference on Digital Experimentation (CODE@MIT) and the 26th ACM Conference on Economics and Computation (EC'25).

[2]Bocconi University, Italy; Lyft, Canada. Email: mahyar.habibi@phd.unibocconi.it.

[3]University of Warwick, United Kingdom. Email: negar.ziaeian-ghasemzadeh@warwick.ac.uk.

from an online travel agency, we demonstrate that our TSPR method accurately estimates the Total Average Treatment Effect (TATE), while traditional baseline item-side estimators, which also satisfy the coherency constraint, substantially overestimate this effect.

Keywords: design of experiments, online platforms, interference, recommender systems, position bias.

1 Introduction

Online platforms, such as e-commerce sites and online marketplaces, frequently conduct randomized controlled experiments (e.g., A/B tests) to optimize user experience, boost engagement, and drive sales. These experiments help mitigate risks associated with changes and innovations while providing rapid product feedback (Kohavi et al., 2020; Bojinov and Gupta, 2022; Xia et al., 2019; Xu et al., 2018; Kohavi et al., 2009).

In the context of price interventions, maintaining coherent prices across users is both a practical necessity and a legal constraint in many jurisdictions. Showing different prices for the same item to different users is tightly regulated, particularly under European competition law, and is considered a red-line violation by many platforms. The importance of this coherency requirement is underscored by historical missteps: in 2000, Amazon's experiment with varying DVD prices led to swift consumer backlash on online forums, ultimately forcing the company to issue public apologies and refunds. Since then, industry best practices have emphasized strict price uniformity, with major platforms such as ride-sharing services pledging that identical trips receive identical fares unless a clearly disclosed discount applies. Informing users that a price mismatch is part of an experiment not only risks reputational damage but also compromises internal validity by altering user behavior. While workarounds like coupon codes or targeted promotions exist, they introduce confounding incentives that blur the interpretation of a pure price effect. Similarly, designs that restrict access to certain items during experimentation, either through removal or selective display, can distort user behavior, degrade trust, and potentially harm platform revenue. For these reasons, our ex-

perimental design is explicitly constructed to maintain both price coherency and full catalog access: all users see the same price for any given item, and all items remain discoverable and accessible to every user.

Standard experimental designs for estimating unbiased treatment effects rely on the Stable Unit Treatment Value Assumption (SUTVA), which states that the treatment assigned to one unit does not influence the outcomes of other units (Rubin, 1974; Imbens and Rubin, 2015). In online marketplaces, this assumption is often violated due to the interconnected nature of users and items. For instance, in item-side experiments, modifying features of treated items—such as offering discounts—can influence demand for non-treated items due to substitution or complementary effects. Such interference, spillover, or network effects have been observed in various platforms, including ridesharing (Chamandy (2016)) and online pricing experiments (Choi and Mela (2019)). Failing to account for interference in randomized experiments can introduce substantial bias, leading to the overestimation or underestimation of the intervention's true impact (Blake and Coey, 2014; Fradkin, 2019).

Under interference, treatment effects depend on how interventions are distributed across units, making the Total Average Treatment Effect (TATE) a central measure in such settings. TATE quantifies the impact of treating all units compared to treating none (Manski, 2013; Munro et al., 2024), and it has been widely adopted in the literature as it captures the full extent of both direct effects and spillovers across units.

We propose the Two-Sided Prioritized Ranking (TSPR) experimental design, tailored for item-side price interventions in two-sided marketplaces, where the outcomes of interest, such as clicks, conversions, or bookings, are observed on the user side. TSPR strategically reorders items within a recommender system's ranked listing. Our approach builds on the well-documented phenomenon of position bias, where items appearing at the top of a list have greater influence on user behavior than those positioned lower (Craswell et al., 2008; Friedberg et al., 2022). Recommender systems, which match items to user queries, are the primary ranking mechanism in online marketplaces. Our design leverages this structure to provide a systematic approach for estimating the TATE of item-side interventions on user-level outcomes. TSPR is specifically tailored to two-sided marketplaces, platforms like Expedia or Airbnb, where users interact with ranked lists of items curated by a central

system.

In the TSPR design, users are randomized into two groups, and items are partitioned into three subsets: the Treated group, which receives the intervention, and two distinct groups that do not, called Untreated and Placebo. For one group of users, the recommender system prioritizes Untreated items at the top of search results, while for the other group, it prioritizes Treated items. Having both Untreated and Placebo groups, with Untreated matching the size of the Treated group, ensures the balanced quality of top-ranked items across user groups. We then estimate TATE by comparing partial outcomes between the two groups of users, where partial outcomes are defined as the cumulative outcome of the prioritized Treated or Untreated items placed at the top of the listings.

To evaluate our methodology, we use an open-source dataset of hotel search impressions from Expedia containing consumer queries, clicks, and booking outcomes. We develop a model of click and booking behavior to generate semi-synthetic data for our Monte Carlo simulations. We simulate a treatment resembling a platform-wide price increase on listed items, which, if applied to all items, reduces the user conversion rate by 0.050. We then compare two estimators across 500 simulation runs. Across simulations, our proposed method estimates TATE with an average of -0.047 and an average bootstrapped standard error of 0.016. In contrast, a naive estimator, which compares the average booking rate between treated and non-treated items, significantly overestimates TATE, producing an average estimated effect of -0.091 and an average standard error of 0.014. We selected this naive estimator as our baseline because it preserves a consistent item view across users—a critical requirement for price experiments. By contrast, methods such as switchback designs or network-based estimators, though valuable elsewhere, violate this consistency and are therefore less suitable for our context.

A key advantage of our design is its ability to preserve a coherent user experience during experimentation. It ensures that no user loses access to any items, regardless of their randomized group assignment, and provides a consistent realization of item treatment across all users. For price interventions, this means all users see the same prices for the same items. A coherent user experience is crucial for online platforms, as it fosters trust, satisfaction, and long-term engagement (Véliz, 2023; Kahneman et al., 1986; Kohavi et al., 2020). However,

many existing methods for estimating TATE in two-sided platforms disrupt the user experience, limiting their practicality for real-world deployment. Switchback testing (Robins, 1986; Sneider and Tang, 2019; Bojinov et al., 2023), for example, alternates treatment assignments over time for the same units, enabling individual-level causal estimation but at the cost of user experience coherency. Frequent treatment fluctuations, such as dynamic pricing changes, not only risk confusing users and distorting engagement patterns but may also introduce carryover effects that compromise internal validity.

Our proposed randomization framework is similar in spirit to two-sided randomization (TSR) methods (Johari et al., 2022; Bajari et al., 2023), which apply independent randomization on both the user and item sides. However, existing TSR designs often do not enforce a coherent user experience: treatment is typically applied only when treated users interact with treated items, which means that different users may be exposed to different versions of the same item (e.g., different prices). In contrast, our two-sided randomization framework ensures a consistent realization of item treatment—every user sees the same version of a given item—while preserving universal access to all items. Despite growing attention to ethical and responsible experimentation in online platforms (e.g., Polonioli et al., 2023; Saint-Jacques et al., 2020), the importance of maintaining user experience coherency within experimental design has received limited attention. By incorporating coherency as a core design principle, our approach is the first to bridge this gap in the causal inference literature for two-sided platforms, while retaining the key statistical benefits of TSR.

Maintaining statistical power is another key strength of our design, as it randomizes at the user level and avoids the limitations imposed by cluster-based approaches. Cluster-based randomization groups related users or items to minimize spillover effects, but suffers from reduced power (Ugander et al., 2013; Eckles et al., 2017; Holtz et al., 2024). Moreover, defining appropriate clusters in dynamic marketplace environments is often infeasible, and poor cluster definitions can lead to severe power loss and unreliable TATE estimates. Even when suitable clusters can be defined, implementing cluster-based randomization can be computationally expensive and operationally complex (Candogan et al., 2023).

Although mitigating interference in ranking experiments has been studied (e.g., Goli et al., 2024; Zhan et al., 2024; Nandy et al., 2021; Ursu, 2018), this prior work typically

focuses on evaluating ranking algorithms themselves. In contrast, we treat the recommender system not as the subject of experimentation but as a tool for implementing the experiment. By integrating recommender systems into experimental design in this way, we bridge the gap between ranking mechanisms and causal inference methods. To the best of our knowledge, this study is the first to propose using recommender systems as an instrument for experimentation in online platforms.

The remainder of this paper is organized as follows: Section 2 formally defines the Two-Sided Prioritized Ranking experimental design and outlines our estimation methodology. Section 3 details the data and simulation setup, followed by the results in Section 4. Finally, Section 5 concludes the paper.

2 Methodology

2.1 Two-Sided Prioritized Ranking (TSPR) Experimentation Setup

We model a two-sided platform as a matching mechanism between a set of queries $q \in Q$, representing user inputs, and a set of items $i \in I$, representing available options. The platform uses a recommender system to compute relevance scores $r_{q,i} \in \mathbb{R}$ for each query-item pair based on the attributes of the query q and the item i, such as user preferences and item features. Once query q is submitted, the platform ranks the available items in descending order of their relevance scores $r_{q,i}$ and presents the ranked list to the user. When a user views the listed items, their interactions determine the outcomes $y_{q,i}$ for each displayed item. For simplicity, we assume all items initially have outcome values of 0, and post-interaction, $y_{q,i}$ takes non-negative real values, representing outcomes such as clicks, bookings, or revenue. Additionally, since we assume each user submits exactly one query, we use the terms "user" and "query" interchangeably.

In the aforementioned setting, standard A/B testing with randomized item-level intervention assignment fails to produce unbiased treatment effect estimates due to the interference between items in the same query, which violates SUTVA. Furthermore, the proposed experimental design must satisfy two key constraints: preserving universal user access to all items

throughout the experiment and maintaining coherent item treatment status across users.

Definition 1 (Coherency)**.** A user experience is said to be *coherent* if all users retain access to the same set of items and view a consistent realization of each item's treatment status, regardless of their group assignment.

Due to item-side interference, the effect of a binary treatment $T \in \{0, 1\}$ on item-query outcomes $y_{q,i}$ depends on the distribution of interventions across units. This motivates our focus on the Total Average Treatment Effect (TATE), which captures both direct effects and spillovers by measuring the difference in expected outcomes between full treatment and no treatment conditions. Given our interest in TATE, we focus on query-level outcomes $Y_q = \sum_i y_{q,i}$, which aggregate individual item-query outcomes across all items displayed in response to query q. For notational simplicity, we omit the query subscript q and denote query-level outcomes as Y. Formally, we denote TATE by θ and define it as:

$$\theta = \mathbb{E}\left[Y | \forall i \in \mathcal{I} : i \in \mathcal{I}^1\right] - \mathbb{E}\left[Y | \forall i \in \mathcal{I} : i \in \mathcal{I}^0\right] \tag{1}$$

where $\mathcal{I}^1 = \{i \in I \mid T_i = 1\}$ and $\mathcal{I}^0 = \{i \in I \mid T_i = 0\}$ denote the sets of treated and non-treated items, respectively. Since each item can only be assigned to one treatment condition ($T_i = 1$ or $T_i = 0$) at a time, only one of the two terms on the right-hand side of the equation is observable at any given point.

The proposed method rests on a few implicit assumptions. First, we assume that items at the top of the listings are significantly more influential on users' behavior (Craswell et al. (2008)), and the influence of items gradually vanishes as we move further down the list. Consequently, the effective exposure of a user to the treatment depends on the extent to which treated items appear near the top of the ranked list, as these items receive a disproportionate share of the user's attention. By strategically changing the ordering of items, we alter the *effective exposure* of a user to the treatment. Second, our method requires the average number of relevant items per query to be sufficiently large to ensure that the repositioning scheme is effective in maximizing the exposure to treated items for one group of queries while minimizing it for the other. Third, we assume that user-side interference is negligible (e.g., slack supply) and the primary source of interference is the interdependence among outcomes

of the items displayed under the same listing.

Our proposed experimental design for estimating TATE is summarized in Table 1, with Figure 1 illustrating the two-sided randomization scheme and group-specific listing priorities for query results.

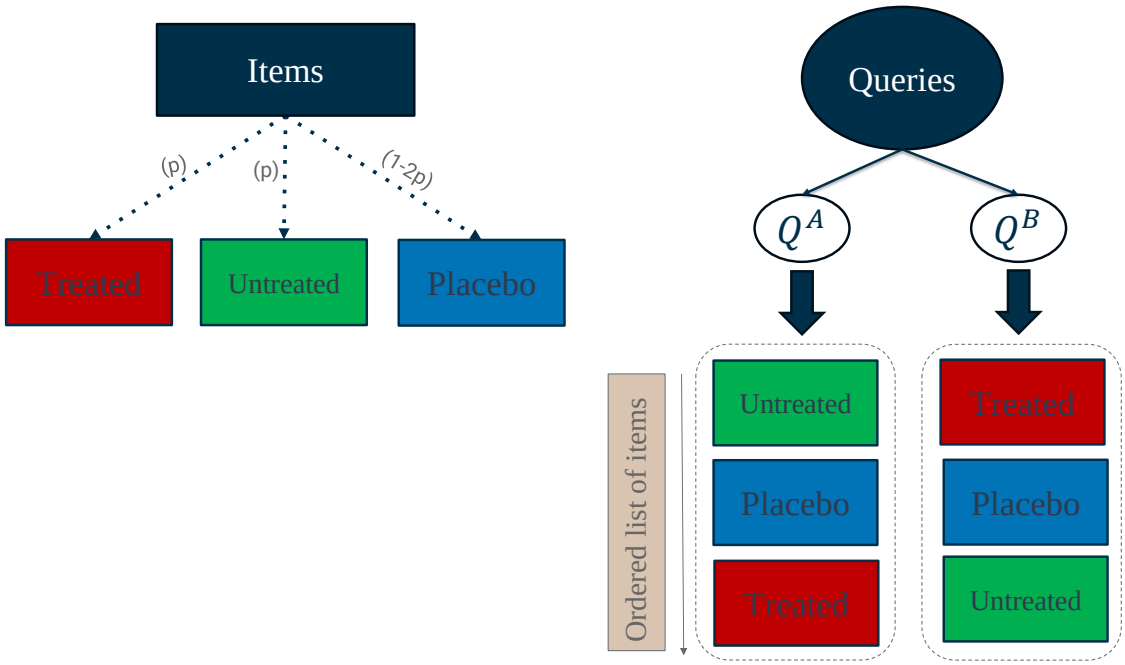
Table 1: Two-Sided Prioritized Ranking (TSPR) Experimental Design

Experiment Setup

1. Set the probability of receiving treatment for an item $p < 0.5$, and minimum relevance threshold $\underline{r}$.
2. Randomize items into Treated, Untreated, and Placebo subsets with probabilities p, p, and $1 - 2p$, respectively. Apply the treatment only to the Treated group.
3. For each incoming query q:
   - 3.1. Randomly assign q to Q^A or Q^B and set the item priorities as follows:
     - If $q \in Q^A$: 1-Untreated, 2-Placebo, and 3-Treated.
     - If $q \in Q^B$: 1-Treated, 2-Placebo, and 3-Untreated.
   - 3.2. Filter the set of relevant items with $r_{q,i} > \underline{r}$.
   - 3.3. Rank items primarily by priority (ascending) and secondarily by relevance score (descending).

As outlined in Table 1, after specifying the global parameters p and $\underline{r}$, we begin by partitioning items into three subsets: Treated, Untreated, and Placebo, with probabilities p, p, and $1 - 2p$, respectively. The intervention is only applied to the items in the Treated subset. Incorporating the Placebo subset serves a crucial purpose in maintaining experimental balance. Without a Placebo subset, when $p < 0.5$, the Untreated subset would be larger than the Treated subset, creating an asymmetric effect in step 3 of our design. Specifically, for queries in Q^A for which we prioritize non-treated items, the larger Untreated pool would yield top-ranked items of higher average quality compared to the top-ranked items from the smaller Treated pool shown to Q^B. This imbalance would cause the recommender system's modification to impact Q^A and Q^B differently, confounding our ability to isolate the intervention's effect. The Placebo subset ensures that the Treated and Untreated are

Figure 1: Two-Sided Prioritized Ranking (TSPR) Experimental Design



Notes: The figure illustrates the TSPR experiment setup. Items are partitioned into three groups, and queries are divided into two subsets. The relevant items for each query are first ordered based on their group-specific priority and then by their relevance score.

roughly equal in size, making the expected match quality per rank comparable between the two groups.

In the next step, the stream of incoming queries are randomized into Q^A or Q^B with equal probability. Then, depending on the randomized allocation, item priorities are assigned such that queries in Q^A will face items in the following order: Untreated, Placebo, and Treated, while the queries in Q^B, will be given items in the order of Treated, Placebo, and Untreated. Before returning the listings, items are filtered to ensure that only items of sufficient relevance are displayed to users. Modifying the recommender system may potentially cause poor-quality matches to be included in the listing, degrade user experience, and create large distortions in user behavior. To address this concern, we introduce a filtering parameter, denoted $\underline{r}$, and only items with a relevance score $r_{q,i} > \underline{r}$ are included in the final listing. The choice of $\underline{r}$ involves a trade-off and should be chosen carefully: setting it too low risks including irrelevant items that could deteriorate user experience, while setting it too high might result in queries receiving an insufficient number of items to display.

2.2 TATE Estimation

Building on the observations from the experimental method described above, we propose an estimator for the Total Average Treatment Effect (TATE) that does not require specifying an exposure mapping or imposing strong assumptions about substitution patterns among items.

The experimental design produces data with two important characteristics. First, because the recommender system prioritizes items rather than exclusively displaying those from one subset, queries in Q^B may see some non-treated items, while queries in Q^A may encounter some treated items. Second, the number of Treated (or Untreated) items appearing at the top of listings varies across queries based on item availability and relevance. These characteristics necessitate an assumption about how the intervention's expected impact on query-level outcomes relates to the number of treated items placed at the top of listings.

We define *partial outcome*, denoted by $Y^l = \sum_{i=1}^{l} y^i$, as the cumulative outcome of the first l listed items. Given the assumption that individual item-level outcomes are non-negative ($y_i \geq 0$), the expected value of the partial outcome, $\mathbb{E}[Y^l]$, is non-decreasing in l, which reflects the fact that including additional items in the calculation of outcome can only increase or maintain the total observed outcome. Similarly, we define the partial treatment effect, θ^l, as the expected effect on partial outcome, Y^l, from applying the intervention only to the first l listed items. That is,

$$\theta^l = \mathbb{E}[Y^l | \forall i \leq l : i \in \mathcal{I}^1 \wedge \forall i > l : i \in \mathcal{I}^0] - \mathbb{E}[Y^l | \forall i : i \in \mathcal{I}^0]. \tag{2}$$

Assumption 1. For all values of l, the partial treatment effect, θ^l, is a fraction of the total treatment effect, θ, by a factor of the ratio of the expected partial outcome to the expected total outcome in the absence of treatment. Formally,

$$\theta^l = \frac{\mathbb{E}[Y^l | \forall i : i \in \mathcal{I}^0]}{\mathbb{E}[Y | \forall i : i \in \mathcal{I}^0]} \theta, \quad \forall l. \tag{3}$$

The assumption above allows us to aggregate the outcomes of queries faced with different numbers of Treated or Untreated units at the top of their corresponding listings. Due to the

declining item influence down the listing position, as l increases, the marginal contribution to the expected partial outcome $\mathbb{E}[Y^l]$ decreases. Consequently, the marginal effect of treating an additional item at rank l gradually approaches zero as we move further down the listing, reflecting the diminishing impact of lower-ranked items on the query-level outcome.

In the TSPR design, the first term in the RHS of Equation 2 corresponds to queries in Q^B where the Treated block contains l items. The second term represents the expected partial outcome Y^l when no items in the list are treated. Under TSPR, this quantity is only observed when there are no relevant Treated units for a query in Q^A, a scenario that may be rare in practice, making direct estimation difficult. Given our assumptions that each query has a relatively large number of available items and that lower-listed items have negligible impact on user behavior, we propose the following approximation:

$$\mathbb{E}[Y^l|\forall i : i \in \mathcal{I}^0] \approx \mathbb{E}[Y^l|\forall i \leq l : i \in \mathcal{I}^0] \tag{4}$$

where l denotes the number of Untreated items. This term can be represented by a query in Q^A with a block of l Untreated items. The introduction of Placebo items in our design further supports the validity of this approximation.

Using Equations 1, 2, 3, 4, and assuming for all l, $\mathbb{E}[Y^l|\forall i \leq l : i \in \mathcal{I}^0] > 0$, we have:

$$\begin{aligned} \theta \approx\ & \mathbb{E}[Y|\forall i : i \in \mathcal{I}^0] \\ & \times \left(\frac{\mathbb{E}[Y^l|\forall i \leq l : i \in \mathcal{I}^1 \wedge \forall i > l : i \in \mathcal{I}^0] - \mathbb{E}[Y^l|\forall i \leq l : i \in \mathcal{I}^0]}{\mathbb{E}[Y^l|\forall i \leq l : i \in \mathcal{I}^0]} \right) \end{aligned} \tag{5}$$

for any choice of l.

Equation 5 establishes a connection between the observable differences in partial outcomes from experimental data and the TATE. The first term on the RHS of the equation can be estimated by running a *pre-experiment* phase in which the modified recommender system is deployed with the same composition of Treated, Placebo, and Untreated items, *before* applying any treatment to the items in the Treated group. The second term on the RHS of the equation can be estimated *during the experiment* using the relative difference in partial outcomes between queries in Q^A and Q^B that have exactly l Untreated and Treated

items at the top of their listings.

To provide a single estimate of TATE, we need to combine the estimates resulted from all observed values of l having at least one correspondent query in Q^A and Q^B. Here, we use frequency weighting and estimate TATE as,

$$\hat{\theta} = \sum_{l=1}^{L} w_l \, \bar{Y}_0 \left(\frac{\frac{1}{|Q_l^B|} \sum_{Q_l^B} Y_B^l - \frac{1}{|Q_l^A|} \sum_{Q_l^A} Y_A^l}{\frac{1}{|Q_l^A|} \sum_{Q_l^A} Y_A^l} \right); w_l = \frac{|Q_l^B| + |Q_l^A|}{|Q^B| + |Q^A|} \tag{6}$$

where Q_l^B (Q_l^A) denote queries in Q^B (Q^A) for which the number of Treated (Untreated) items was equal to l, and $\bar{Y}_0$ is the expected outcome estimated using pre-experiment phase with the modified recommender system. Standard errors can be obtained via bootstrapping.

3 Data and Simulation

To illustrate our methodology, we use an open-source dataset of hotel search impressions from Expedia (Adam et al., 2013), capturing consumer queries and their corresponding search behaviors—specifically, clicks and booking outcomes—over an eight-month period spanning 2012 and 2013. The dataset encompasses nearly 10 million observations derived from approximately 400,000 unique search impressions. Each search impression represents the result of a consumer query, providing a list of hotels along with their observable characteristics.

Consumers interact with the platform in three stages. First, consumers initiate queries by specifying trip details (destination, travel dates, booking window, etc.). Second, they receive a ranked list of hotel results through an experimental setup: two-thirds of users see listings ranked by the platform's original recommender system, while the remaining one-third encounter randomly sorted results. This experimental variation in ranking mechanisms allows us to model how item positions influence click and booking behavior. Finally, users engage by clicking on hotels to view details and may either complete a booking or leave without purchasing.

To evaluate our experimental design, we implement a series of Monte Carlo simulations that replicate consumer interactions in an online two-sided marketplace, incorporating query-driven item ranking, click behavior, and booking decisions. We assume that the platform

Table 2: Summary Statistics of Search Impressions

	Mean	Median	Min	Max
Randomized Ranking (Yes=1)	0.30	0	0	1
Total Hotels per Impression	24.56	29	4	33
Clicks per Impression	1.11	1	1	30
Bookings per Impression	0.69	1	0	1

maintains a pool of available items, denoted as N, and displays a subset n_q in response to each query.

To model user interactions, we assume that each item displayed to a consumer has a net (hidden) utility, denoted as v. The relevance score r, which represents the recommender system's match score between a consumer's query and an item, is modeled as $r = v + \epsilon$ with ϵ following a normal distribution $N(0, \sigma^2)$. We assume that the original ranking system is decreasing in r. However, for randomly ranked search impressions, the sorting order is determined randomly.

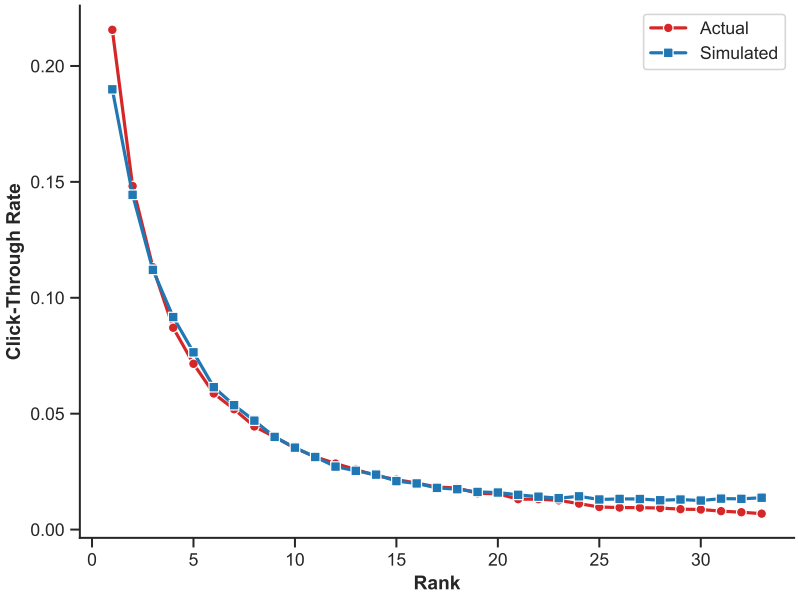
Click probabilities are modeled as a logistic function of the raw and quadratic rank values, hidden utilities, and prior user clicks on lower-ranked options. Booking decisions are modeled as a logit choice among clicked items, depending solely on net utility v. To ensure the simulation aligns with real-world behavior, hyperparameters σ_e and n_q are selected to match simulated conversion rates with observed data. This is achieved through an iterative process, where click and booking parameters are first estimated using the data-generating process, followed by user action simulations. The simulated conversion rates are then compared with empirical rates, and hyperparameters are adjusted to minimize discrepancies. Figure 2 shows that the simulated click-through rate closely matches the observed data, demonstrating the convergence of the simulation to real-world behavior.

Table 2 presents summary statistics at the search impression level, highlighting key patterns in click and booking behaviors across random and relevance-based rankings.

4 Results

We conduct counterfactual simulations for 20,000 queries using our estimated models of click and booking behavior. First, to establish a simulated ground truth for TATE, we simulate

Figure 2: Click-Through Rate by Item Rank



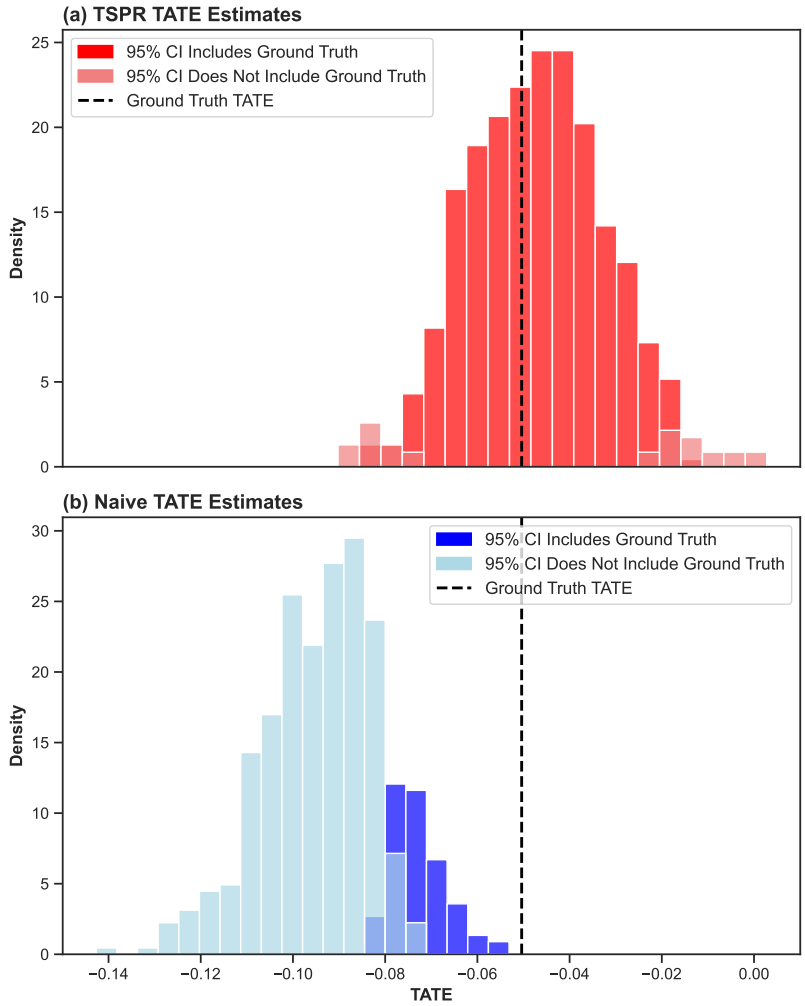
Notes: This figure presents the actual click-through rate (CTR) and the simulated CTR as a function of item position in the query results from a hold-out sample not used in the estimation of the click and booking models.

the marketplace under two extreme scenarios: one where no items receive treatment and another where all items are treated. The treatment is implemented as a constant reduction in users' hidden utility from booking an item, resembling the effect of a platform-wide price or markup increase, which translates to a 0.05 decrease in the conversion rate. In these simulations, we maintain the recommender system without any modifications.

We then implement our Two-Sided Prioritized Ranking (TSPR) experimental design to estimate TATE in a setting where treatment is applied to 25% ($p = 0.25$) of the items. Following our methodology, we randomly assign each query to either group A or B with equal probability. For one group, the recommender system is modified to prioritize Treated items in the ranking, while for the other group, it prioritizes Untreated items. The remaining items are positioned according to the experimental design outlined in Table 1, maintaining access to all items while creating the necessary variation in exposure to treatment.

Additionally, to form a baseline for comparison, we simulate an item-side randomized experiment, which is an extension of the Horvitz-Thompson estimator Horvitz and Thompson (1952) to two-sided platforms and commonly used as a baseline in the literature (e.g., Johari

Figure 3: TATE Estimates



Notes: The figure presents Total Average Treatment Effect (TATE) estimates from 500 simulated experiments. The vertical dashed line represents the ground truth TATE, corresponding to a 0.05 reduction in booking rate. Panel (a) plots the histogram of TATE estimates from the TSPR method with 25% treatment coverage ($p = 0.25$), yielding an average estimate of -0.047 (average bootstrapped SE: 0.016). Panel (b) plots the histogram of TATE estimates from the naive estimator under the same conditions, yielding an average estimate of -0.091 (average bootstrapped SE: 0.014). TATE estimates with 95% confidence intervals that contain the ground truth are shown in bold color.

et al., 2022; Bajari et al., 2023). In this setup, items are randomly assigned to the treatment group (T) with probability p or the control group (C) with probability $1-p$. However, no randomization at the query level takes place. TATE is estimated as the difference in means of total outcomes per query, adjusted by the inverse of the inclusion probability of treated and control items:

$$\hat{\theta}_{IS} = \frac{\sum_{i \in T} \sum_{q \in Q} y_{q,i}}{p|Q|} - \frac{\sum_{i \in C} \sum_{q \in Q} y_{q,i}}{(1-p)|Q|} \tag{7}$$

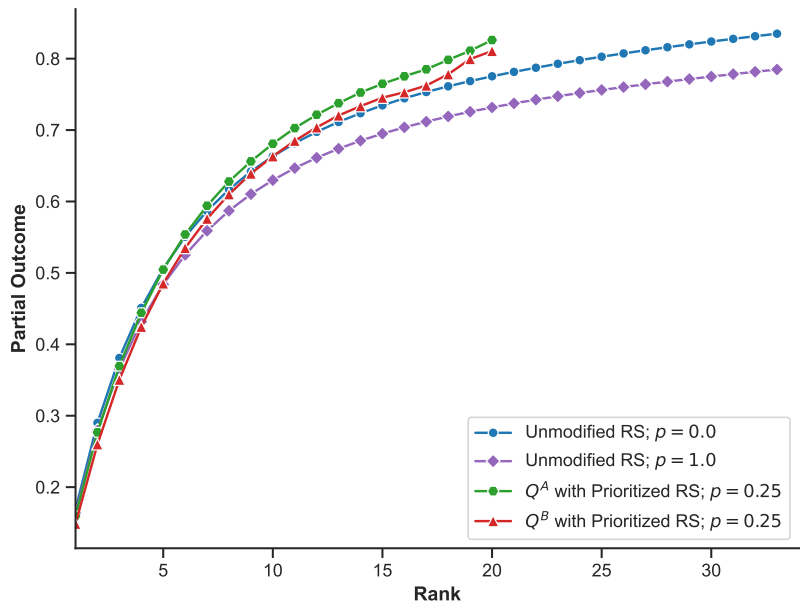
where Q denotes the set of all queries, and $y_{q,i}$ represents the outcome for item i in query q .

Figure 3 presents the distribution of the TATE estimates from the TSPR setup and contrasts it with the distribution of $\hat{\theta}_{IS}$ across 500 runs. TSPR estimates TATE with an average of -0.047 (average bootstrapped SE: 0.016). In contrast, the baseline estimator significantly overestimates the effect with an average of -0.091 (average bootstrapped SE: 0.014), roughly double the true value, despite using a similar treatment group size. This overestimation occurs because the baseline approach fails to account for interference between treated and non-treated items within the same listing. The proportion of the TATE estimates with 95% confidence intervals that contain the ground truth TATE (-0.05) is considerably larger for TSPR compared to the baseline estimator.

The relevance threshold parameter $\underline{r}$ serves a dual methodological and practical purpose: maintaining partial outcomes close to the original recommender system while preserving user experience through quality control of top-positioned items. In practice, platforms can determine an appropriate value for $\underline{r}$ using historical data or pre-intervention experiments with the modified recommender system, allowing them to balance the trade-off between listing quality and estimates' accuracy. In our simulations, we set $\underline{r} = 1.7$ to keep partial outcomes under the modified recommender system close to the baseline case, as illustrated in Figure 4. The figure demonstrates that marginal contributions to partial outcomes are substantial for small values of l but rapidly diminish with increasing position. This supports our approximation (4) suggesting minimal contributions from items down the list.

Figure 5 examines the robustness of our estimates to the choice of relevance threshold $\underline{r}$. The TATE estimates remain relatively stable under moderate changes in $\underline{r}$, suggesting that our methodology is robust to the specific choice of this parameter. However, as illustrated,

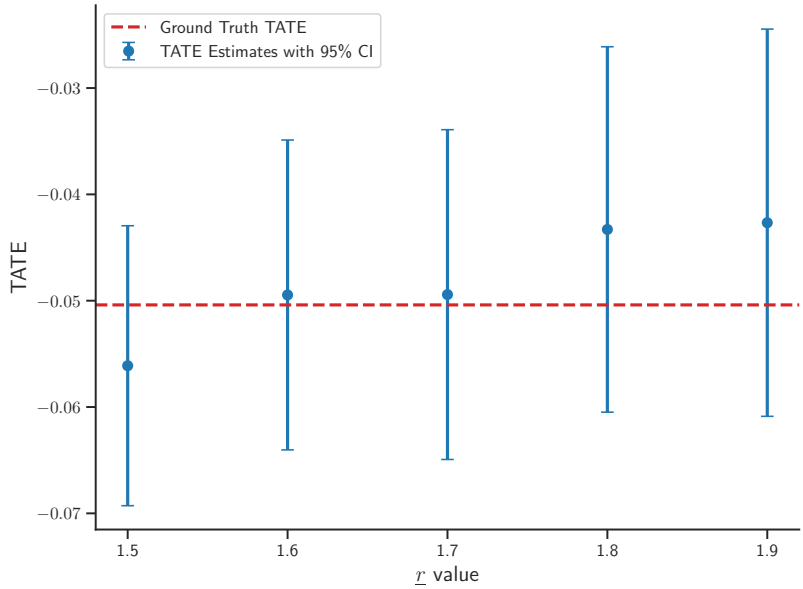
Figure 4: Partial Outcomes Across Ranks



Notes: The figure plots the partial outcomes Y^l for rank l, in four scenarios across 100 simulations. The first two scenarios are under the unmodified recommender system with no treatment ($p = 0.0$) and full treatment ($p = 1.0$). The other two scenarios illustrate the partial outcomes for Q^A and Q^B in the simulated experiments when the probability of assignment to both the Treated and Untreated group is $p = 0.25$.

higher values of $\underline{r}$ restrict the number of Treated or Untreated items that qualify for top positions and reduce the range of block sizes available for estimation in Equation 5, leading to larger standard errors in our estimates.

Figure 5: Sensitivity Analysis of the TATE Estimates to $\underline{r}$



Notes: The figure illustrates the TATE estimates and their 95% bootstrapped confidence intervals for different choices of $\underline{r}$ from 100 simulated experiments in the TSPR setup when the probability of assignment to both the Treated and Untreated group is $p = 0.25$.

5 Conclusion

This paper introduces a novel experimental design for two-sided marketplaces that leverages recommender systems to estimate the Total Average Treatment Effect (TATE) while addressing interference and maintaining a coherent user experience. By reordering items in query listings, our Two-Sided Prioritized Ranking (TSPR) design minimizes bias from network spillovers and ensures equal access to items for all users. Using a semi-synthetic dataset of hotel search impressions, we demonstrate that our design provides reliable TATE estimates while a standard item-side estimator significantly overestimates TATE.

The TSPR method is particularly designed for platforms relying on ranking algorithms, such as e-commerce sites and online marketplaces. Our design enables such platforms to

estimate treatment effects while maintaining a coherent user experience and preserving platform functionality. It addresses concerns regarding unintended consequences of randomized experiments, such as user disengagement or inequitable access to items.

Our paper opens promising avenues for future research. First, although we evaluate our method using semi-synthetic data, validating it with real-world data from diverse platforms would enhance its generalizability and provide deeper insights into its practical implementation. Second, our design focuses on item-side interventions and assumes negligible cross-user interference. Extending it to user-side interventions or settings with significant cross-user interference remains an open challenge and would greatly expand its applicability.

References

Adam, Hamner, B., Friedman, D. and SSA_Expedia (2013), 'Personalize expedia hotel searches - icdm 2013', https://kaggle.com/competitions/expedia-personalized-sort. Kaggle.

Bajari, P., Burdick, B., Imbens, G. W., Masoero, L., McQueen, J., Richardson, T. S. and Rosen, I. M. (2023), 'Experimental design in marketplaces', *Statistical Science* **38**(3), 458–476.

Blake, T. and Coey, D. (2014), Why marketplace experimentation is harder than it seems: the role of test-control interference, *in* 'Proceedings of the Fifteenth ACM Conference on Economics and Computation (EC '14)', Association for Computing Machinery, New York, NY, USA, pp. 567–582.

Bojinov, I. and Gupta, S. (2022), 'Online experimentation: Benefits, operational and methodological challenges, and scaling guide', *Harvard Data Science Review* **4**(3).

Bojinov, I., Simchi-Levi, D. and Zhao, J. (2023), 'Design and analysis of switchback experiments', *Management Science* **69**(7), 3759–3777.

Candogan, O., Chen, C. and Niazadeh, R. (2023), 'Correlated cluster-based randomized experiments: Robust variance minimization', *Management Science* **70**(6), 4069–4086.

Chamandy, N. (2016), 'Experimentation in a ridesharing marketplace—lyft engineering', https://eng.lyft.com/experimentation-in-a-ridesharing-marketplace-b39db027a66e. Accessed October 1, 2022.

Choi, H. and Mela, C. F. (2019), 'Monetizing online marketplaces', *Marketing Science* **38**(6), 948–972.

Craswell, N., Zoeter, O., Taylor, M. and Ramsey, B. (2008), An experimental comparison of click position-bias models, *in* 'WSDM'08 - Proceedings of the 2008 International Conference on Web Search and Data Mining', pp. 87–94.

Eckles, D., Karrer, B. and Ugander, J. (2017), 'Design and analysis of experiments in networks: Reducing bias from interference', *Journal of Causal Inference* **5**(1), 20150021.

Fradkin, A. (2019), 'A simulation approach to designing digital matching platforms', *Boston University Questrom School of Business Research Paper* . Forthcoming.

Friedberg, R., Rajkumar, K., Mao, J., Yao, Q., Yu, Y. and Liu, M. (2022), 'Causal estimation of position bias in recommender systems using marketplace instruments', *arXiv preprint arXiv:2205.06363* .

Goli, A., Lambrecht, A. and Yoganarasimhan, H. (2024), 'A bias correction approach for interference in ranking experiments', *Marketing Science* **43**(3), 590–614.

Holtz, D., Lobel, F., Lobel, R., Liskovich, I. and Aral, S. (2024), 'Reducing interference bias in online marketplace experiments using cluster randomization: Evidence from a pricing meta-experiment on airbnb', *Management Science* .

Horvitz, D. G. and Thompson, D. J. (1952), 'A generalization of sampling without replacement from a finite universe', *Journal of the American statistical Association* **47**(260), 663–685.

Imbens, G. W. and Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press, USA.

Johari, R., Li, H., Liskovich, I. and Weintraub, G. Y. (2022), 'Experimental design in two-sided platforms: An analysis of bias', *Management Science* **68**(10), 7069–7089.

Kahneman, D., Knetsch, J. L. and Thaler, R. (1986), 'Fairness as a constraint on profit seeking: Entitlements in the market', *The American Economic Review* **76**(4), 728–741. **URL:** *http://www.jstor.org/stable/1806070*

Kohavi, R., Crook, T., Longbotham, R., Frasca, B., Henne, R., Lavista Ferres, J. and Melamed, T. (2009), Online experimentation at microsoft, *in* 'Third Workshop on Data Mining Case Studies and Practice Prize'.

Kohavi, R., Tang, D. and Xu, Y. (2020), *Trustworthy online controlled experiments: A practical guide to a/b testing*, Cambridge University Press.

Manski, C. F. (2013), 'Identification of treatment response with social interactions', *The Econometrics Journal* **16**(1), S1–S23.
URL: *http://www.jstor.org/stable/23364965*

Munro, E., Kuang, X. and Wager, S. (2024), 'Treatment effects in market equilibrium'.
URL: *https://arxiv.org/abs/2109.11647*

Nandy, P., Venugopalan, D., Lo, C. and Chatterjee, S. (2021), 'A/b testing for recommender systems in a two-sided marketplace', *Advances in Neural Information Processing Systems* **34**, 6466–6477.

Polonioli, A., Ghioni, R., Greco, C., Juneja, P., Tagliabue, J., Watson, D. and Floridi, L. (2023), 'The ethics of online controlled experiments (a/b testing)', *Minds and Machines* **33**(4), 667–693.

Robins, J. M. (1986), 'A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect', *Mathematical Modelling* **7**(9-12), 1393–1512.

Rubin, D. B. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies', *Journal of Educational Psychology* **66**(5), 688–701.

Saint-Jacques, G., Sepehri, A., Li, N. and Perisic, I. (2020), 'Fairness through experimentation: Inequality in a/b testing as an approach to responsible design', *arXiv preprint arXiv:2002.05819* .

Sneider, C. and Tang, Y. (2019), 'Experiment rigor for switchback experiment analysis', https://doordash.engineering/2019/02/20/experiment-rigor-for-switchback-experiment-analysis/.

Ugander, J., Karrer, B., Backstrom, L. and Kleinberg, J. (2013), 'Graph cluster randomization: network exposure to multiple universes'.
URL: *https://arxiv.org/abs/1305.6979*

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

Véliz, C., ed. (2023), *Oxford Handbook of Digital Ethics*, Oxford Handbooks, Oxford University Press. Online edition published on Oxford Academic, 10 Nov. 2021. Accessed 11 Feb. 2025.

Xia, T., Bhardwaj, S., Dmitriev, P. and Fabijan, A. (2019), Safe velocity: a practical guide to software deployment at scale using controlled rollout, *in* '2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)', IEEE, pp. 11–20.

Xu, Y., Duan, W. and Huang, S. (2018), Sqr: Balancing speed, quality and risk in online experiments, *in* 'Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining', pp. 895–904.

Zhan, R., Han, S., Hu, Y. and Jiang, Z. (2024), 'Estimating treatment effects under recommender interference: A structured neural networks approach', *arXiv preprint arXiv:2406.14380* .