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]

November 16, 2025

Abstract

In online two-sided marketplaces, users must see a coherent platform environment in which all items remain accessible and item attributes, such as prices or features, are displayed consistently across users. Many interventions, especially those involving price changes, require such coherency to avoid confusion or perceptions of unfair treatment. These constraints complicate causal inference because users interact with shared items, and treatment applied to one item can spill over onto others, creating interference bias in standard A/B tests. We propose a new experimental design, Two-Sided Prioritized Ranking (TSPR), which uses the marketplace's ranking algorithm as an experimental instrument. TSPR assigns treatment through prioritized exposure while preserving coherency and

[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.

limiting spillovers. Our goal is to estimate the lift, defined as the proportional change in total outcomes between counterfactual worlds in which all items are treated or untreated. Using Monte Carlo simulations based on large-scale search impression data from an online travel agency, we show that TSPR produces lift estimates that are substantially less biased than those from coherent item-side A/B tests. We also compare TSPR to cluster-randomized designs, which can reduce interference when clusters isolate exposure but require detailed knowledge of the underlying network. TSPR provides a practical and scalable alternative for platforms that must preserve coherency while obtaining reliable estimates of treatment effects.

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

1 Introduction

Online platforms such as e-commerce sites and online marketplaces rely heavily on randomized controlled experiments to guide product decisions. These experiments help platforms evaluate changes safely, improve user experience, and increase engagement and sales, while providing timely and credible feedback on new features (Kohavi et al., 2020; Bojinov and Gupta, 2022; Xia et al., 2019; Xu et al., 2018; Kohavi et al., 2009).

A central challenge for price experiments is the need to maintain coherent prices across users. Showing different prices for the same item to different users is often infeasible in practice and is legally restricted in many jurisdictions. Overt price variation is tightly regulated under European competition law and is viewed by many platforms as a red-line violation. Historical failures reinforce the importance of coherency. In 2000, Amazon experimented with varying DVD prices, which triggered immediate consumer backlash on online forums and forced the company to issue public apologies and refunds. Since then, major platforms have emphasized strict price uniformity. For example, ride-sharing services commit to charging identical fares for identical trips unless a clearly disclosed discount applies. Disclosing

that a price difference is part of an experiment not only risks reputational costs but also undermines internal validity by altering user behavior. Workarounds such as coupon codes or targeted promotions introduce their own confounding incentives and no longer identify a clean price effect.

Maintaining full catalog access is another important requirement for coherent experimentation. Designs that remove items from search results or display different item sets to different users can distort normal browsing patterns, erode trust, and harm platform revenue. For these reasons, our experimental design is constructed to preserve both price coherency and full catalog access. All users see the same price for any given item, and all items remain discoverable and available to every user throughout the experiment.

Standard experimental designs for estimating unbiased treatment effects rely on the Stable Unit Treatment Value Assumption (SUTVA), which requires 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 frequently violated because users and items interact through shared rankings and limited attention. In item-side experiments, for example, modifying features of treated items, such as offering discounts, can shift demand toward or away from non-treated items through substitution or complementarity. These forms of interference, spillover, or network dependence have been documented in several settings, including ridesharing platforms (Chamandy, 2016) and online pricing experiments (Choi and Mela, 2019). When interference is ignored, estimates from randomized experiments can be substantially biased, which leads to overestimation or underestimation of the true effect of the intervention (Blake and Coey, 2014; Fradkin, 2019).

Under interference, the impact of an intervention depends on how treatment is distributed across units, which makes the total change in outcomes between fully treated and fully untreated worlds a natural target of inference. We refer to this quantity as the lift. Lift measures the proportional change in aggregate outcomes when all items receive treatment compared to when none do. It captures the full effect of an intervention, including both its direct impact and all spillovers across items and users (Manski, 2013; Munro et al., 2024). This makes lift an appropriate summary measure in marketplaces where interference is pervasive.

We propose the Two-Sided Prioritized Ranking (TSPR) experimental design, which is tailored for item-side price interventions in two-sided marketplaces where the outcomes of interest, such as clicks or bookings, are observed on the user side. Recommender systems are the main mechanism that match items to user queries and determine the ranking that shapes user behavior. TSPR takes this structure as given and uses it to estimate the total lift of an item-side intervention on user-level outcomes. The design works by reordering items within the recommender system's ranked lists in a way that assigns treatment through priority shifts rather than by hiding items or altering the set of available options. TSPR is designed for platforms such as Expedia or Airbnb where users interact with ranked lists curated by a central system. Our approach builds on the well-documented phenomenon of position bias, where items displayed near the top of a list receive more attention and have greater influence on user choices than those ranked lower (Craswell et al., 2008; Friedberg et al., 2022).

Position bias is a systematic pattern in which users interact more frequently with items that appear higher in a ranked list, regardless of their true relevance. This behavior is well documented in search and recommendation settings, where users tend to click on early results largely because of their position rather than their intrinsic quality (Craswell et al., 2008). Empirical evidence shows a steep decline in click probability as an item moves down the ranking (Friedberg et al., 2022). Several behavioral models explain this pattern. The examination hypothesis assumes that users must first examine an item before deciding whether to click on it, while cascade models propose that users inspect items sequentially from top to bottom and stop once they find a satisfactory option (Craswell et al., 2008; Richardson et al., 2007). Joachims et al. (2017) further attribute part of this behavior to trust bias, where users place excessive confidence in the ranking algorithm. These models show that observed clicks combine both position and relevance effects, which leads naive estimators to overstate the performance of higher-ranked items. Our design uses this regularity to create structured variation in exposure while maintaining full coherency in user experience.

In the TSPR design, users are randomized into two groups and items are partitioned into three sets: Treated, Untreated, and Placebo. Treated items receive the intervention, while Untreated and Placebo items do not. One user group is shown rankings that prioritize

Untreated items at the top of the list. The other user group is shown rankings that prioritize Treated items. The Untreated set is sized to match the Treated set so that the overall quality of items placed at the top of search results remains comparable across user groups. This structure induces systematic variation in treatment exposure while preserving full coherency in both the set of items displayed and the prices shown to users.

To evaluate our methodology, we use an open-source dataset of hotel search impressions from Expedia that contains user queries, clicks, and booking outcomes. We estimate a model of click and booking behavior and use it to generate semi-synthetic data for our Monte Carlo simulations. The intervention we study is a platform-wide price increase that enters the utility model as a negative shift in latent value. We assess the performance of TSPR by applying it to both the semi-synthetic data and the real impression graph extracted from the Expedia dataset. In each setting, we simulate user actions under the observed ranking structure and compute lift estimates across repeated simulation runs. We then compare the TSPR estimates (mean lift $= -0.144$) to those produced by a standard item-side A/B test (mean lift $= -0.219$) and to the ground truth lift of -0.125. Across simulations, the naive item-side estimator exhibits substantial bias because it ignores interference between treated and non-treated items that appear together in ranked lists, and this bias becomes more severe when items compete more intensively for user attention. TSPR reduces this bias by inducing structured variation in treatment exposure that aligns with how users interact with ranked results. TSPR achieves more accurate lift estimates, although with higher variance than the naive estimator, and this bias–variance tradeoff appears in both the semi-synthetic environment and the real impression data.

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 that every user observes the same realization of an item's treatment status. For price interventions, this means that all users see the same prices for the same items. A coherent user experience is important for online platforms because it supports user trust, satisfaction, and long-term engagement (Véliz, 2023; Kahneman et al., 1986; Kohavi et al., 2020).

Many existing methods for estimating treatment effects in two-sided platforms disrupt

this coherency, which limits 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. This design allows for individual-level causal identification but produces frequent treatment fluctuations that can confuse users and distort normal engagement patterns. When these fluctuations involve prices or other salient features, they may also create carryover effects that undermine internal validity. For these reasons, methods that violate user experience coherency are difficult to implement for item-side price interventions in large-scale marketplaces.

Our randomization framework is closely related to two-sided randomization (TSR) methods (Johari et al., 2022; Bajari et al., 2023), which apply independent randomization on both the user side and the item side. However, existing TSR designs typically do not enforce a coherent user experience. In standard TSR implementations, treatment is applied only when a treated user interacts with a treated item, which allows different users to see different versions of the same item, including different prices. This violates the coherency requirement that motivates our setting.

In contrast, our two-sided randomization framework ensures a consistent realization of item treatment across all users. Every user sees the same version of a given item and retains access to the full catalog regardless of their own assignment. Although recent work has highlighted the importance of ethical and responsible experimentation in online platforms (e.g., Polonioli et al., 2023; Saint-Jacques et al., 2020), the role of user experience coherency within experimental design has received limited attention. By incorporating coherency as a core design principle, our approach is the first to address this gap in the causal inference literature for two-sided platforms while retaining the key statistical benefits of TSR.

Maintaining statistical power is another strength of our design because it randomizes at the user level and avoids the limitations of cluster-based approaches. Cluster-based randomization groups related users or items to reduce spillover, but it suffers from reduced power due to the smaller number of effective randomization units (Ugander et al., 2013; Eckles et al., 2017; Holtz et al., 2024). In many marketplace settings, defining suitable clusters is difficult because user interactions and item relationships evolve over time. Poorly chosen clusters can lead to substantial power loss and unreliable lift estimates. Even when appropriate clus-

ters can be identified, implementing cluster-based randomization can be computationally expensive and operationally complex (Candogan et al., 2023).

Although prior work has examined how to mitigate interference in ranking experiments (e.g., Goli et al., 2024; Zhan et al., 2024; Nandy et al., 2021; Ursu, 2018), this research typically focuses on evaluating or improving the ranking algorithms themselves. Our perspective is different. We do not treat the recommender system as the object of experimentation but instead use it as the mechanism through which the experiment is implemented. By integrating the recommender system directly into the design of the experiment, we connect ranking mechanisms with causal inference in a way that has not been explored in the literature. 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 the paper is organized as follows. Section 2 formally defines the Two-Sided Prioritized Ranking design and describes the estimation methodology. Section 3 presents the data and simulation setup. Section 4 reports the empirical performance of the design. Section 5 concludes.

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 $q \in Q$, which represent user inputs, and a set of items $i \in I$, which represent the 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 attributes of the query and the item, such as user preferences and item features. When a user submits query q, the platform ranks all available items in descending order of $r_{q,i}$ and displays the ordered list to the user. After viewing the list, the user may interact with some of the displayed items, and these interactions generate outcomes $y_{q,i}$. For simplicity, we assume that all items begin with outcome value zero and that $y_{q,i}$ takes non-negative real values after user interaction, representing clicks, bookings, or revenue. Because each user submits exactly one query in our setting, we use the terms

"user" and "query" interchangeably.

In this environment, standard item-level A/B testing fails to produce unbiased estimates of treatment effects because items shown together in the same query can affect each other's outcomes. This violates the Stable Unit Treatment Value Assumption (SUTVA) due to interference between items. Any experimental design for this setting must also satisfy two operational constraints. First, users must retain access to the full catalog of items during the experiment. Second, all users must observe a coherent realization of item treatment status, meaning that every user sees the same version of each item throughout the experiment. These constraints rule out many existing designs and motivate the structure of our Two-Sided Prioritized Ranking approach.

Definition 1 (**Coherency**)**.** A user experience is *coherent* if all users retain access to the same set of items and if every user observes the same treatment status for any given item, independent of their randomized group assignment.

Due to item-side interference, the effect of a binary treatment $T_i \in \{0, 1\}$ on item–query outcomes $y_{q,i}$ depends on how treatments are distributed across items. This motivates our focus on the *global lift* (Φ), which captures both direct effects and spillovers by comparing expected outcomes under full treatment and full control. Because our estimand is defined at the query level, we aggregate item outcomes within each query and work with $Y_q = \sum_i y_{q,i}$. For notational simplicity we omit the query index and write Y.

We define global lift as

$$\Phi = \frac{\mathbb{E}[\, Y \mid \forall i \in I : \, T_i = 1 \,]}{\mathbb{E}[\, Y \mid \forall i \in I : \, T_i = 0 \,]}, \tag{1}$$

where I is the set of all items. The numerator corresponds to the expected query-level outcome when all items are treated, and the denominator corresponds to the expected outcome when all items are untreated. Since in practice each item can only be in one treatment state at a time, only one of these two quantities is observed, which makes Φ fundamentally a counterfactual estimand.

The proposed method rests on several assumptions that we now discuss. First, we assume that items at the top of the listing exert a disproportionate influence on user behavior (Craswell et al., 2008), and that this influence declines rapidly with rank. Effective exposure

to the treatment therefore depends on the extent to which treated items appear near the top of the ranked list, since these positions receive most of the user's attention. By strategically altering the ordering of items, we manipulate users' effective exposure to treated versus untreated items.

Second, the method requires that each query contains a sufficiently large set of relevant items. This ensures that the repositioning scheme can meaningfully increase the exposure of one group of queries to treated items while decreasing it for the other group.

Third, we assume that user-side interference is negligible. This corresponds to a slack-supply environment in which inventory or availability constraints are not binding over the experiment horizon. Under slack supply, one user's actions do not affect item availability for others, and interference arises entirely *within queries*, across items displayed in the same ranked list. In our model, within-query interference operates through two mechanisms: (i) limited attention to early ranks and (ii) unit-demand substitution, since booking one item reduces the probability that other items in the same query are chosen.

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

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. Only items in the Treated subset receive the intervention. The inclusion of a Placebo subset is essential for maintaining balance in the experiment.

In marketplace experiments, the probability of assignment to either treatment or control is typically well below 0.5, often on the order of a few percent, in order to limit opportunity costs and to mitigate potential negative effects on user experience if the new feature performs worse than the existing one (Ha-Thuc et al., 2020). Without a Placebo subset, the Untreated subset would be substantially larger than the Treated subset. This would create an asymmetric effect in step 3 of our design. In particular, for queries in Q^A, where non-treated items are prioritized, the larger Untreated pool would produce top-ranked items of higher average quality than the top-ranked items drawn from the smaller Treated pool shown to Q^B. Such an imbalance would cause the recommender system modification to affect the two query

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:

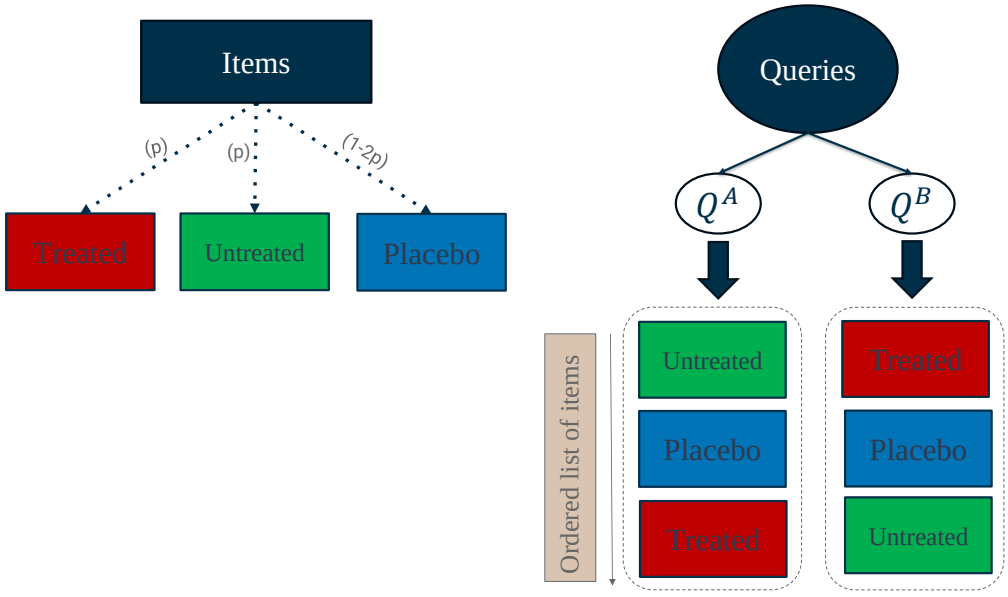
   3.1. Randomly assign q to Q^A or Q^B and set the item priorities as follows:

   - 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).

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.

groups differently, confounding the estimation of the intervention's effect.

The Placebo subset prevents this imbalance by ensuring that the Treated and Untreated subsets are of comparable size. As a result, the expected match quality of top-ranked items is similar across the two user groups, which allows the variation induced by the prioritization scheme to isolate the treatment effect rather than reflect differences in pool size or quality.

In the next step, incoming queries are randomized into Q^A or Q^B with equal probability. Item priorities are then assigned so that queries in Q^A receive items in the order Untreated, Placebo, Treated, while queries in Q^B receive items in the reverse order: Treated, Placebo, Untreated. This prioritization induces systematic differences in exposure to treated items across the two query groups.

Before the ranked listings are returned to users, we apply a relevance filter to ensure that only sufficiently relevant items appear in the final display. Modifying the recommender system can introduce low-quality matches that would not have been shown under the baseline ranking, which may degrade user experience and create large behavioral distortions. To address this concern, we introduce a filtering parameter $\underline{r}$ and retain only items with relevance scores $r_{q,i} > \underline{r}$. The choice of $\underline{r}$ involves a trade-off: values that are too low risk including irrelevant items, while values that are too high may leave some queries with too few items to display. In practice, $\underline{r}$ can be selected to balance listing quality with the effectiveness of the prioritization scheme.

2.2 Theoretical Setup and Estimation Framework

This section introduces the analytical framework used throughout the paper. We begin by outlining the setup and notation, then define the estimand of interest that captures treatment effects under varying ranking and attention conditions. We next formalize the identifying assumptions required for consistent estimation and describe the estimator that operationalizes these ideas in practice.

The parameter Φ captures the *relative (multiplicative) effect* of treatment, defined as the

proportional lift in total outcomes under full treatment compared to no treatment:

$$\Phi = \frac{\mathbb{E}!\left[Y_{q,T}^{\infty}\right]}{\mathbb{E}!\left[Y_{q,U}^{\infty}\right]} - 1. \tag{2}$$

Thus, Φ represents the percentage change in expected total outcomes when all items are treated.

A TSPR experiment is characterized by the set of treated queries Q^B, the set of control queries Q^A, the set of items $\mathcal{I}$, the treatment intensity p, the treatment type T, and the randomization and re-ranking scheme implemented according to Algorithm 1.

Definition 2 (**Partial Outcome**)**.** The *partial outcome*, denoted $Y_q^l = \sum_{i=1}^{l} y_q^i$, is the cumulative outcome for query q over the first l listed items. Since item-level outcomes are non-negative ($y_{q,i} \geq 0$), $\mathbb{E}[Y_q^l]$ is non-decreasing in l.

For notational simplicity, we drop the query index q and refer to query-level outcomes as Y. All expectations are taken over queries within a given experimental arm.

Definition 3 (**Attention Function**)**.** The attention function is defined as $\mathbb{E}[Y^l] = F(l)\mathbb{E}[Y]$, where $F : \mathbb{N} \to (0, 1)$ is increasing and concave, and $F(\infty) = 1$.

Assumption 1 (**Attention and Treatment Separability**)**.** If all items are treated, the treatment affects the level but not the shape of the attention function.

Assumption 1 implies that if treatment were rolled out to all items, the expected partial outcome would satisfy

$$\mathbb{E}[Y^l \mid \textit{full treatment}] = (1 + \Phi) \cdot F(l) \cdot \mathbb{E}[Y \mid \textit{no treatment}]. \tag{3}$$

In a TSPR experiment, the platform perturbs its baseline relevance ranking to generate variation in exposure. Such changes alter how attention is distributed across the list. Because the baseline recommender maximizes outcomes by favoring highly relevant items near the top, these perturbations generally lower total and partial outcomes. We denote the distorted attention function by $D(F(l))$.

Assumption 2 (**Multiplicative Distortion**). Distortion is multiplicative: $D(F(l)) = d(l)F(l)$, where $d(l) \in (0, 1)$ captures attenuation in attention at depth l.

Assumption 3 (**Symmetric Distortion**). The distortion function $d(l)$ is identical across experimental arms. Thus, the re-ranking distortion introduced by TSPR affects groups A and B equally in expectation.

Consequently, under the distorted ranking,

$$\mathbb{E}[Y^l \mid TSPR] = d(l)F(l)\mathbb{E}[Y \mid Original\ Rec.\ Sys.]. \tag{4}$$

Equation (3) characterizes partial outcomes under full treatment. Under TSPR, however, treatment is applied to only a small subset of items, but the re-ranking scheme uses position bias to maximize exposure to treated items for queries in Q^B and minimize exposure for queries in Q^A. Partial treatment and ranking distortion therefore require a more general formulation.

We introduce two functions, $\tau(\cdot)$ and $\nu(\cdot, \cdot)$, that characterize how treatment exposure interacts with ranking in treatment-dominated (Q^B) and control-dominated (Q^A) listings. The function $\tau(\cdot)$ captures the *scaling of treatment effects* when treated items fill the top positions in group B, reflecting substitution or complementarity across these items. The function $\nu(\cdot, \cdot)$ captures the *contamination effect* for group A, where a small number of treated items may appear in lower ranks and influence expected outcomes.

Invoking Assumptions 1 and 2, for a query in group B with n_b treated items at the top:

$$\mathbb{E}[Y_B^l \mid TSPR] = (1 + \tau(n_b)\Phi)\ d(l)\ F(l)\ \mathbb{E}[Y \mid no\ treatment]. \tag{5}$$

Assumption 4 (**Partial Treatment Effect**). $\tau : \mathbb{N} \to \mathbb{R}^+$ satisfies $\tau(n) \to 1$ as $n \to \infty$. The function may be concave ($\tau(1) < 1$), convex ($\tau(1) > 1$), or constant ($\tau(\cdot) = 1$).

Assumption 4 implies that treating only the top-ranked items preserves the direction of the treatment effect, and its magnitude converges to the full-treatment effect as the treated block grows. Substitutability or complementarity between items determines whether $\tau(1)$ is above or below one.

We assume that items are substitutes *in expectation*: while some queries may contain complements, on average treated items at the top draw demand away from untreated items below. This assumption allows $\tau(1) > 1$ on average.

For a query in group A, with n_u untreated items, n_p placebo items, and n_a treated items appearing later in the list, we model partial outcomes for $l \leq n_u + n_p$ as:

$$\mathbb{E}[Y_A^l \mid TSPR] = (1 + \nu(n_u + n_p, n_a)\Phi)d(l)F(l)\mathbb{E}[Y \mid no; treatment]. \tag{6}$$

Assumption 5 (**Contamination Effect**)**.** The nuisance function $\nu : \mathbb{Z} \times \mathbb{Z} \to [0, 1)$ is decreasing in the number of untreated and placebo items and increasing in the number of treated items. It satisfies $\nu(\cdot, 0) = 0$, and $\nu \to 0$ as exposure to treated items becomes negligible.

We define the partial lift of treatment as the ratio of partial outcomes across the two groups:

$$1 + \phi(l) = \frac{\mathbb{E}[Y_B^l]}{\mathbb{E}[Y_A^l]} = \frac{1 + \tau(n_b)\Phi}{1 + \nu(n_u + n_p, n_a)\Phi}. \tag{7}$$

Under the null of no treatment effect, $\phi(l) = 0$ for all l, which is satisfied in TSPR because distortion affects both groups symmetrically.

Special Case. Under any of the following conditions:

- attention decays sharply with rank,
- the recommendation list is very long, or
- $p \ll 1$, so that the Placebo group is large relative to the Treated group,

the impact of treated items appearing in the lower part of the list on the outcomes of the Untreated items at the top becomes negligible. In these settings, exposure to treated tail items contributes little to the partial outcomes of queries in group A, and the contamination effect is small. As a result, the approximation implied by Assumption 5 holds to a good degree.

2.3 Parametrization

We propose a parametric form for the function $\tau(l)$ that satisfies the properties discussed above. Using the definition of the partial lift in Equation (7), we write

$$\begin{aligned} \phi(l) &= \Phi\,\tau(l) \\ &= \Phi\left(1 + \frac{\gamma}{1+e^{kl}}\right), \qquad k > 0, \end{aligned} \tag{8}$$

where the parameters (γ, k) govern the pattern and strength of interference across items.

Interpretation of parameters.

- γ controls the *direction* and *magnitude* of interference. A value $\gamma > 0$ corresponds to a substitution relationship among items, which amplifies the effect of treating only the top-ranked items. A value $\gamma < 0$ corresponds to complementarity, which attenuates the partial treatment effect.
- When $\gamma \to 0$, we recover the case of *no interference*, so $\tau(l)$ becomes constant in l and partial lift converges directly to Φ.
- The parameter $k > 0$ regulates the *speed* at which $\tau(l)$ converges to 1 as l increases. Large values of k imply that the effect of interference dissipates quickly, so $\phi(l) \approx \Phi$ even for small l.
- The functional form ensures that $\tau(l) \to 1$ as $l \to \infty$, which is consistent with the assumption that partial treatment effects converge to the full-treatment effect when many items are treated.

Estimation. Given data from a TSPR experiment, we compute the empirical partial lift $\phi(l)$ by taking the ratio of partial outcomes for queries in Q^B and Q^A. Specifically, for each l, we compute

$$\widehat{\phi}(l) = \frac{\mathbb{E}[Y_B^l]}{\mathbb{E}[Y_A^l]} - 1,$$

using only queries in group B that have exactly l treated items in the top positions and queries in group A that have exactly l untreated items in the top positions.

With empirical values $\widehat{\phi}(l)$ for a range of l, and under standard regularity conditions (e.g., sufficient support across l), we jointly estimate the parameters (γ, k) and the global lift Φ by fitting the parametric model in Equation (8). This yields a fully parametric estimate of the total lift Φ that incorporates both treatment effects and interference patterns induced by the ranking structure.

2.4 Estimation: Weighted Nonlinear Least Squares

We estimate the total proportional lift Φ and the interference parameters (γ, k) by minimizing a weighted sum of squared deviations between the empirical partial lifts $\widehat{\phi}(l)$ and their model-implied values $\phi^{\text{model}}(l) = \Phi\,\tau(l;\gamma,k)$, where $\tau(l;\gamma,k) = 1 + \frac{\gamma}{1+e^{kl}}$:

$$Q(\Phi,\gamma,k) = \sum_{l=1}^{L} w(l)\left[\widehat{\phi}(l) - \Phi\,\tau(l;\gamma,k)\right]^2, \tag{9}$$

where $w(l)$ denotes a nonnegative precision weight, typically chosen as the inverse of $\mathrm{Var}[\widehat{\phi}(l)]$ or proportional to the number of queries contributing to the l-th partial outcome.

Because the model is linear in Φ but nonlinear in (γ, k), we employ a concentration approach to simplify the optimization. For any given values of (γ, k), the objective in (9) is quadratic in Φ, yielding the closed-form weighted least squares estimator:

$$\widehat{\Phi}(\gamma,k) = \frac{\sum_{l=1}^{L} w(l)\,\tau(l;\gamma,k)\,\widehat{\phi}(l)}{\sum_{l=1}^{L} w(l)\,\tau(l;\gamma,k)^2}. \tag{10}$$

Substituting $\widehat{\Phi}(\gamma,k)$ back into (9) yields a *concentrated* criterion function,

$$Q_c(\gamma,k) = Q\big(\widehat{\Phi}(\gamma,k),\gamma,k\big), \tag{11}$$

which depends only on the nonlinear parameters (γ, k). Minimizing $Q_c(\gamma,k)$ numerically provides $(\widehat{\gamma},\widehat{k})$, and the final estimate of the total lift is obtained by evaluating $\widehat{\Phi} = \widehat{\Phi}(\widehat{\gamma},\widehat{k})$.

The parameters are estimated using the Levenberg–Marquardt algorithm (`scipy.optimize.curve_fit`). Standard errors for the Φ estimates are computed via bootstrap resampling.

2.5 Empirical Support for Assumptions

Figure 3 provides empirical support for Definition 3. Across all experimental conditions, the expected partial outcomes $\mathbb{E}[Y^l]$ increase with rank length l but at a decreasing rate, producing the concave shape characteristic of the cumulative attention function $F(l)$. The marginal contributions to partial outcomes are large for small values of l but diminish quickly as position increases, reflecting the strong position bias documented in the literature. The unmodified recommender system under no treatment ($p = 0$) and full treatment ($p = 1$) exhibits nearly identical attention profiles, indicating that treatment scales outcomes without altering the shape of attention, which is consistent with Assumption 1.

Assumption 3 is reasonable given the short experimental horizon and the slack-supply environment. The limited duration prevents cross-group spillovers such as popularity feedback or relevance drift, while slack supply ensures that user actions in one group do not constrain item availability for the other. Figure 3 also provides empirical support for Assumption 3. The partial-outcome curves for Q^A and Q^B under the TSPR setup ($p = 0.25$) lie slightly below those of the unmodified recommender, but their shapes and magnitudes are almost identical. This pattern shows that the re-ranking distortion attenuates attention in a symmetric fashion across the two experimental arms, consistent with the assumption that the distortion function $d(l)$ is common to both.

3 Data and Simulation

To illustrate our methodology, we use an open-source dataset of hotel search impressions from Expedia (Adam et al., 2013). The data capture consumer queries and their subsequent search behavior, including clicks and booking outcomes, over an eight-month period spanning 2012 and 2013. The dataset contains nearly 10 million observation-level records generated from approximately 400,000 unique search impressions. Each search impression corresponds to a user query and includes the list of hotels returned by the platform 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 maintains a pool of available items, denoted as N, and displays a subset n_q in response to each query.

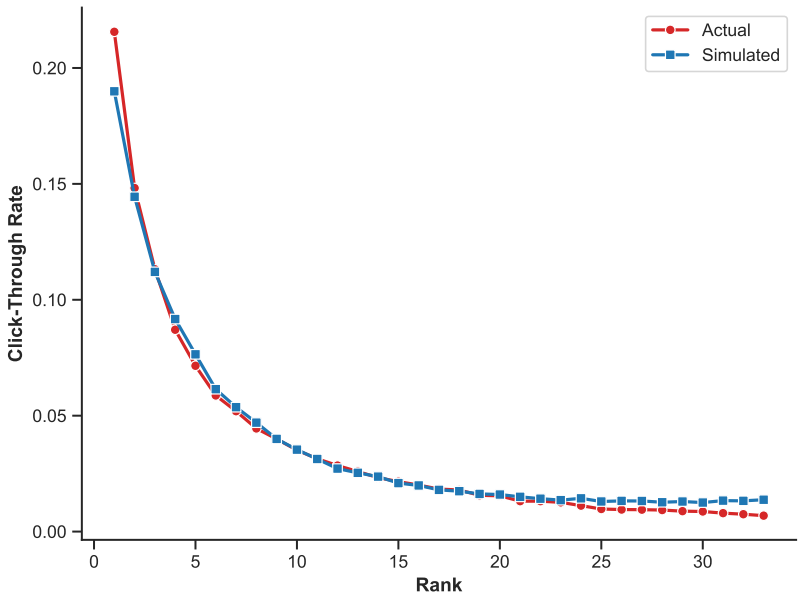
To model user interactions, we assume that each item displayed to a consumer has a net (hidden) utility, denoted as v. The relevance score r, which represents the recommender system's match score between a consumer's query and an item, is modeled as $r = v + \epsilon$ with ϵ following a normal distribution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

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.

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

3.1 Click and Booking Model

Click Model. Click behavior is modeled using a logistic function that incorporates rank-based attention, sequential stopping, and item relevance. For each item j shown at position p to user i, the probability of a click is

$$P(\text{click}_{ij}) \;=\; \text{logit}^{-1}\big(\beta_1 p + \beta_2 p^2 + \beta_3\,\text{prevclicks}_i + \beta_4 v_{ij} + \beta_0\big) \tag{12}$$

where v_{ij} is the latent utility of item j for user i, and prevclicks_i is the number of clicks user i has made on earlier positions in the same query.

- The coefficients β_1 and β_2 capture attention decay across ranks. As the position p

increases, the baseline probability of a click declines, which reflects limited examination of lower-ranked items. This pattern is consistent with the well-documented form of position bias that arises in ranked listings.

- The coefficient β_3 captures a stopping effect. The probability of clicking decreases as more items have already been clicked earlier in the list. This creates a continuation probability that is analogous to the stopping parameter in position-based models with continuation.

- The coefficient β_4 captures item relevance. Higher latent utility v_{ij} increases the likelihood of a click, independent of position or previous clicks.

Simulation proceeds sequentially by position: at each step, the realized click outcome updates prevclicks_i, thereby reducing the likelihood of additional clicks further down the list. This structure embeds both position bias and stopping effects, common ways of modeling position bias in the literature (Craswell et al., 2008; Richardson et al., 2007), parametrically within the logit specification.

Booking Model. Conditional on having clicked at least one item, the user chooses among the clicked set C_i using a multinomial logit model. Each clicked item $j \in C_i$ has a booking utility

$$U_{ij}^{\text{book}} \;=\; \gamma_1 v_{ij} + \gamma_0,$$

where v_{ij} is the latent utility of the item (possibly adjusted for treatment).

The probability of booking item j is

$$P_{ij}^{\text{book}} = \frac{\exp\big(U_{ij}^{\text{book}}\big)}{1 + \sum_{k \in C_i} \exp\big(U_{ik}^{\text{book}}\big)},$$

where the denominator includes an outside option (the "1" term) that allows for the possibility of no booking.

- The coefficient γ_1 measures how strongly latent utility v_{ij} translates into booking likelihood, ensuring that more attractive items are systematically favored.

- The constant γ_0 captures baseline booking propensity, shifting overall booking rates without altering relative preferences across items.
- The inclusion of the outside option (the "1" in the denominator) implements *unit demand*: users may choose not to book at all, and when they do book, they book exactly one item.

One item is then sampled from this distribution (or the outside option), and the booking outcome is recorded. Together, the click and booking models form a sequential choice process: items must first be clicked to enter the choice set C_i, and then the multinomial logit allocates the booking probability among those clicked items.

Treatment. In our simulation framework, treatment enters as a constant shift in the latent utility of an item. This shift affects both click and booking behavior because utility enters each stage of the model. Let T_{ij} denote the treatment indicator for item j shown to user i. Treated items receive a utility shift δ, so that the effective utility becomes $v_{ij} + \delta T_{ij}$.

Click probabilities are therefore

$$P(\text{click}_{ij}) \;=\; \text{logit}^{-1}\big(\beta_1 p + \beta_2 p^2 + \beta_3\,\text{prevclicks}_i + \beta_4\big(v_{ij} + \delta T_{ij}\big) + \beta_0\big)\,. \tag{13}$$

After the user has clicked a set of items C_i, booking proceeds through a multinomial logit model. The booking probability for item $j \in C_i$ is

$$P_{ij}^{\text{book}}(T_{ij}) = \frac{\exp\big(\gamma_1(v_{ij} + \delta T_{ij}) + \gamma_0\big)}{1 + \sum_{k\in C_i}\exp\big(\gamma_1(v_{ik} + \delta T_{ik}) + \gamma_0\big)}.$$

The simulation worlds are generated by applying these estimated click and booking models to the impression data. For each query, we simulate the sequence of user actions by drawing clicks from the click model and then drawing a booking choice from the multinomial logit model conditional on the clicked set. These simulated actions produce the final outcomes used to evaluate each experimental design.

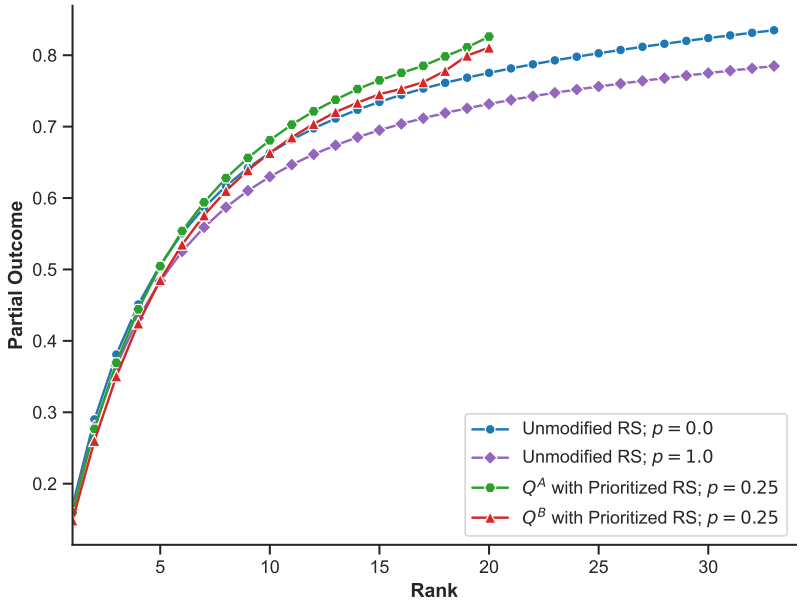
4 Results

We conduct counterfactual simulations for 20,000 queries using the estimated models of click and booking behavior. To establish a simulated ground truth for lift, we simulate the marketplace under two extreme scenarios: one in which no items receive treatment and one in which all items are treated. The treatment enters as a constant reduction in the latent utility of an item, which represents the effect of a platform-wide price or markup increase and implies a 12.5 percent decline in bookings under full treatment. The recommender system is held fixed in both simulations. The resulting proportional change in total bookings serves as the benchmark against which we evaluate the lift estimates produced by each experimental design.

We then implement our Two-Sided Prioritized Ranking (TSPR) experimental design to estimate total lift 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.

The relevance threshold parameter $\underline{r}$ serves both methodological and practical purposes. It keeps partial outcomes close to those produced by the original recommender system and preserves user experience by ensuring that highly relevant items remain in top positions. In practice, platforms can choose an appropriate value for $\underline{r}$ using historical data or small-scale pre-intervention tests with the modified recommender system, which allows them to balance the tradeoff between listing quality and the accuracy of lift estimates. In our synthetic-data simulations, we set $\underline{r} = 1.7$ to keep partial outcomes under the modified system close to the baseline case, as shown in Figure 3. The figure shows that marginal contributions to partial outcomes are large for small values of l but diminish quickly as the position increases. We also find that lift estimates are robust to modest changes in $\underline{r}$, which indicates that the design does not depend on precise calibration of this parameter, although larger adjustments do shift the estimates.

Figure 3: 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 $p = 0.25$.

Figure 4 summarizes the performance of the NLS lift estimator for the TSPR experimental design in the semi-synthetic environment. The distribution of estimated lift is centered near the ground truth value $\Phi = -0.273$, with a mean of -0.253. The absolute bias is therefore about 0.02, which corresponds to a 2 percentage point difference in lift. Sampling variation across the 1,000 Monte Carlo replications is moderate. This experiment isolates the behavior of the estimator in a controlled setting where both the choice sets and the outcome model are generated from the specification in Subsection 3.1.

Figure 5 reports the same estimator and TSPR experimental design applied to the real impression graph `df_single`. In this case the true lift is $\Phi = -0.125$ and the mean estimate is -0.144, which implies an absolute bias of about 0.019, or 1.9 percentage points. The sampling distribution remains concentrated around the true value and is similar in shape to the semi-synthetic case. Taken together, these two figures show that the TSPR design recovers the ground truth lift with small absolute bias, measured in percentage points of lift,

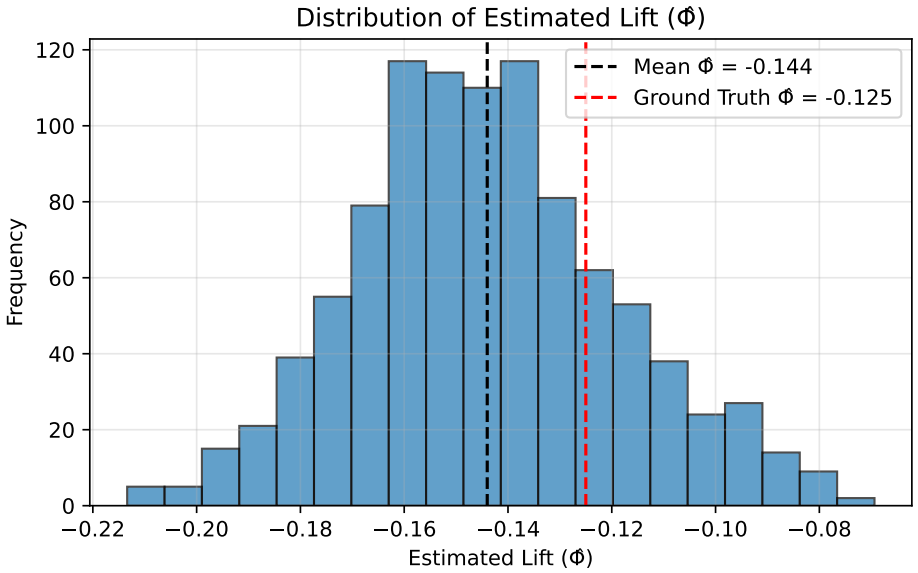


Figure 4: Distribution of estimated lift Φ from TSPR experiments in the semi-synthetic setting. The histogram is based on 1,000 simulation runs based on sampled queries and items, where choice sets are generated synthetically and clicks and bookings follow the model in Subsection 3.1. The dashed vertical line marks the mean estimated lift and the red line marks the ground truth value of Φ.

in both a fully synthetic world and a world based on real search impressions.

4.1 Baseline: Item-side A/B Testing

As a baseline, we consider an item-side randomized experiment that extends the Horvitz-Thompson logic to the two-sided marketplace setting. In this design, randomization occurs only at the level of items. Each item i is independently assigned to treatment with probability p or to control with probability $1-p$. Let $Z_i \in \{0,1\}$ denote the treatment indicator, where $Z_i = 1$ if item i is assigned to treatment and $Z_i = 0$ otherwise. Define the treated and control item sets as

$$T = \{i : Z_i = 1\}, \qquad C = \{i : Z_i = 0\}.$$

Let Q denote the set of all queries, and let $y_{q,i}$ be the observed outcome for item i in query q. In the item-side experiment there is no randomization at the query level. The only source of randomization is the item assignment, which affects the distribution of treated and control items that appear in each query.

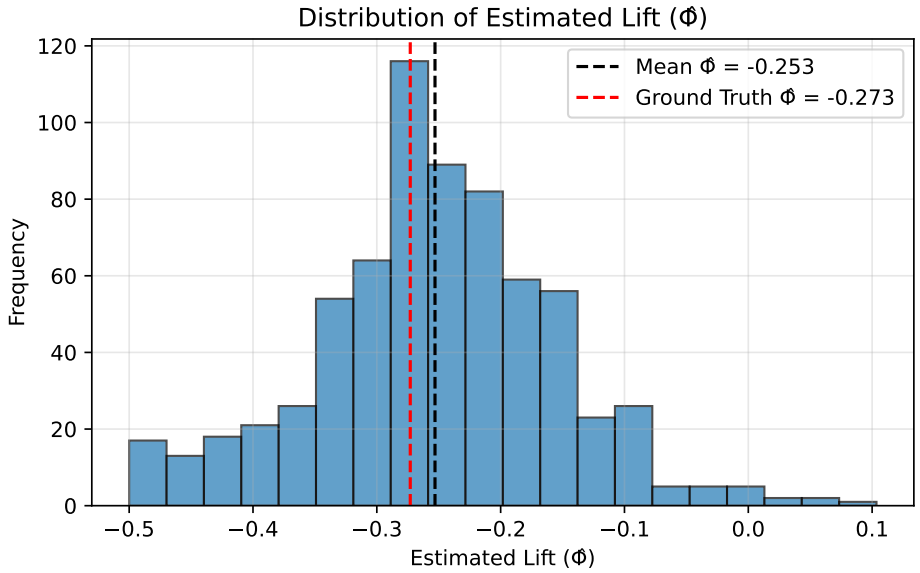


Figure 5: Distribution of estimated lift Φ from TSPR experiments on real search impression data. The histogram is based on 1,000 simulation runs on the `df_single` sample, where the observed impression graph is held fixed and clicks and bookings are simulated according to the model in Subsection 3.1. The dashed vertical line marks the mean estimated lift and the red line marks the ground truth value of Φ.

We define μ_B as the mean total outcome per query under an intervention that treats all items, and μ_A as the mean total outcome per query under an intervention that keeps all items in control. We estimate these quantities using Horvitz-Thompson style estimators based on the item-level randomization:

$$\hat{\mu}_B^{IS} = \frac{1}{|Q|} \sum_{q \in Q} \sum_i \frac{Z_i \, y_{q,i}}{p} = \frac{1}{|Q|} \sum_{q \in Q} \sum_{i \in T} \frac{y_{q,i}}{p}, \tag{14}$$

$$\hat{\mu}_A^{IS} = \frac{1}{|Q|} \sum_{q \in Q} \sum_i \frac{(1 - Z_i) \, y_{q,i}}{1 - p} = \frac{1}{|Q|} \sum_{q \in Q} \sum_{i \in C} \frac{y_{q,i}}{1 - p}. \tag{15}$$

Our estimand of interest is the lift in mean outcome, defined as

$$\phi = \frac{\mu_B}{\mu_A} - 1.$$

The corresponding item-side estimator is the ratio of the Horvitz-Thompson estimators for

μ_B and μ_A:

$$\hat{\phi}_{IS} = \frac{\hat{\mu}_B^{IS}}{\hat{\mu}_A^{IS}} - 1. \tag{16}$$

Since the factor $1/|Q|$ appears in both $\hat{\mu}_B^{IS}$ and $\hat{\mu}_A^{IS}$, it cancels in the ratio. We can therefore write the estimator in the equivalent form

$$\hat{\phi}_{IS} = \frac{\sum_{q \in Q} \sum_{i \in T} \frac{y_{q,i}}{p}}{\sum_{q \in Q} \sum_{i \in C} \frac{y_{q,i}}{1-p}} - 1. \tag{17}$$

This provides a simple item-side baseline that uses only item-level randomization and ignores query-level randomization.

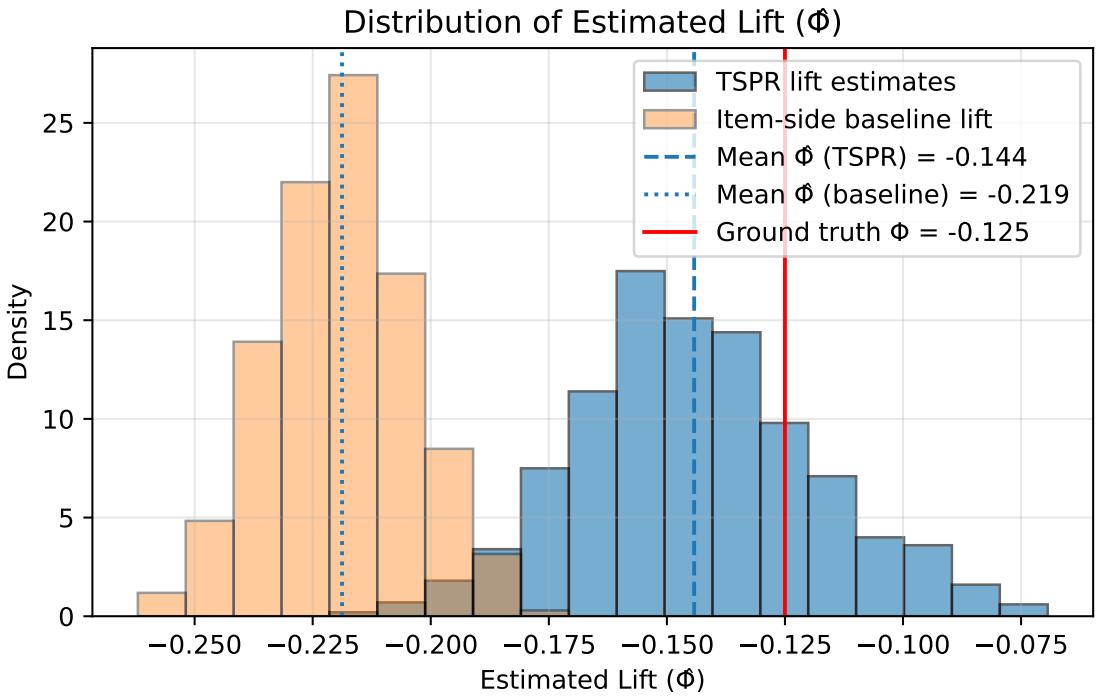


Figure 6: Distribution of lift estimates from 1,000 simulations under the TSPR design and the item-side baseline A/B test. The vertical lines show the mean estimate for each method and the ground truth lift, computed as the proportional change in total outcomes between counterfactual all-treated and all-control simulations.

Figure 6 shows the sampling distributions of lift estimates from 1,000 simulations under the TSPR design and the item-side baseline experiment. The treatment is implemented as a constant reduction in users' hidden utility from booking an item, which resembles

the effect of a platform-wide price or markup increase. The vertical dashed line represents the ground truth lift, corresponding to a booking-rate change of $\phi = -0.125$ computed from counterfactual simulations in which all items are treated or all items are untreated. The blue histogram plots the distribution of ϕ estimates from the TSPR method with 25% treatment coverage ($p = 0.25$), yielding an average estimate of -0.144. These estimates display higher variability but have a mean that lies close to the ground truth. The orange histogram plots the lift estimates from the naive item-side estimator under the same utility change and treatment group size, yielding an average estimate of -0.219. These baseline estimates are more tightly concentrated yet clearly biased downward relative to the ground truth. The figure therefore illustrates that TSPR trades an increase in variance for a substantial reduction in bias compared with the naive item-side design.

The naive item-side estimator exhibits substantial bias because it ignores interference between treated and non-treated items that appear together in ranked lists, and this bias becomes more severe when items compete more intensively for user attention. TSPR reduces this bias by inducing structured variation in treatment exposure that aligns with how users interact with ranked results.

4.2 Baseline: Cluster-Randomized Experiments

As a second baseline, we compare our estimates to lift ratio estimates obtained from cluster-randomized experiments (Holtz et al., 2024). Cluster randomization reduces interference bias because units within a cluster share the same treatment assignment, which limits spillover across treatment arms. Implementing cluster randomization, however, requires detailed knowledge of the underlying network structure and is often costly. When it can be applied correctly, such as in a hotel booking platform that randomizes treatment at the level of geographic clusters (for example, cities), it preserves user experience coherency under our definition. For this reason, cluster-randomized experiments provide a relevant benchmark for evaluating the performance of TSPR.

To construct this baseline in our setting, we use the real search impression data from Expedia described in Section 3. Each observation is a property j that appears in a search query i. The data contain a destination identifier at the search level, `srch_destination_id`,

which we interpret as a geographic cluster such as a city or region. We map each property to the set of destinations in which it appears. In the full dataset, many properties appear in multiple destinations, which creates potential cross-cluster interference.

To obtain a clean cluster structure, we restrict attention to properties that appear in exactly one destination. We denote this restricted sample by `df_single`. For each property in `df_single`, we define its cluster as the unique `srch_destination_id` observed in the impression data. We then treat each destination as a cluster of properties and perform cluster randomization at the destination level. Clusters are independently assigned to treatment or control with probability P_{treat}. All properties in a treated destination inherit the treatment assignment. All properties in a control destination remain in the control group. This design respects geographic segmentation and avoids spillover between treated and control clusters by construction.

The simulation of user behavior uses the observed impression graph and a semi-synthetic outcome model. We keep the realized sets of properties shown in each query and construct a baseline relevance score r_{ij} from observed features of the property and query. This score determines the ranking that users see. Treatment affects latent utility through a property-level treatment effect τ_j that applies to all impressions of property j in treated clusters. Clicks and bookings are then generated from a parametric choice model conditional on the displayed ranking. We estimate lift as the ratio of average booking outcomes in treated clusters to average booking outcomes in control clusters, minus one.

To make the comparison with TSPR design as transparent as possible, we also re-run the TSPR experiment on the same restricted dataset `df_single`. In this exercise, we keep the real impression structure and the same baseline relevance scores r_{ij}. We then apply the original TSPR item-level randomization and ranking scheme on top of these real choice sets. This yields a distribution of TSPR lift estimates that is directly comparable to the distribution of lift estimates from cluster-randomized experiments. Figure 7 reports the empirical distributions of estimated lift under TSPR and under cluster randomization in this common environment and provides the basis for the comparison discussed below.

Across 1,000 simulation runs, both designs produce lift estimates that are centered below zero, consistent with a negative average effect of treatment on bookings. The cluster-

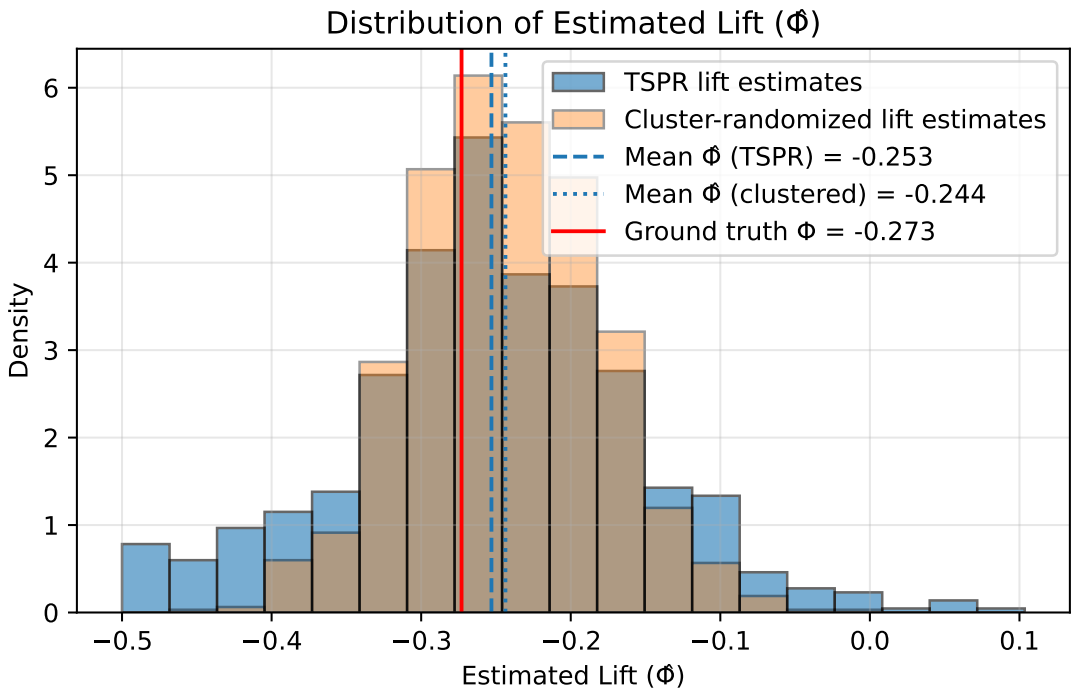


Figure 7: Distribution of estimated lift Φ under TSPR and cluster-randomized designs. Each histogram is based on 1,000 simulation runs using the same impression data and outcome model. Vertical lines mark the mean estimated lift for each design and the true value of Φ.

randomized lift estimator is biased upward relative to the ground truth value $\Phi = -0.273$, with a mean of approximately -0.244. The TSPR-based lift estimates are also biased upward but lie closer to the truth, with a mean of approximately -0.253. The distribution of TSPR estimates is slightly more dispersed than that of the cluster-randomized estimates. This pattern indicates a modest variance cost for TSPR in exchange for a smaller bias relative to a coherent cluster-randomized design that has access to the true geographic clusters.

5 Conclusion

This paper develops a new experimental design for two-sided marketplaces that preserves coherent user experience while allowing low-bias estimation of the lift in mean outcomes induced by an item-side intervention. We define coherency as the requirement that all users retain access to the same set of items and observe a consistent realization of each item's treatment status. The Two-Sided Prioritized Ranking (TSPR) design satisfies this

requirement by implementing treatment through priority shifts within the ranking algorithm rather than through item removal or user-level differentiation. This mechanism ensures that experimentation does not fragment the marketplace or introduce artificial differences in the characteristics of items shown to users, such as displaying different prices or features during the experiment.

Using a semi-synthetic dataset of hotel search impressions, we show that TSPR yields lift estimates that lie substantially closer to the ground truth than those from a naive item-side experiment. TSPR does not eliminate all bias, but it reduces the distortion that arises when items are treated in isolation and produces mean estimates that align more closely with the ground truth lift. The ground truth is defined as the proportional change in total outcomes between counterfactual worlds in which all items are treated or all items are untreated. Across 1,000 simulations in this semi-synthetic setting, TSPR estimates remain centered near this benchmark, whereas the naive item-side design produces estimates that are stable but visibly biased because treated and untreated items appear together in the same listings, which allows the treatment to spill over onto control items and creates interference bias.

We also implement TSPR on real search impression data from a hotel platform. In this setting the impression graph is fixed by the historical data and only the treatment assignment and user responses are simulated. The NLS estimator under TSPR again recovers the ground truth lift with small absolute bias and a sampling distribution that is similar to the semi-synthetic case. This evidence suggests that the statistical performance of TSPR is robust to the choice-set structure induced by real marketplace traffic.

Finally, we compare TSPR to cluster-randomized experiments that randomize treatment at the level of geographic clusters and preserve coherency by assigning all items in a cluster to the same treatment arm. In our simulations on the real impression graph, cluster randomization produces lift estimates with lower variance but a larger bias relative to the ground truth than TSPR. Cluster randomization performs well when clusters coincide with the true pattern of spillovers, but it requires detailed knowledge of the underlying network and can be costly to implement. TSPR offers a practical and scalable alternative in environments where the relevant structure is unknown, only partially observed, or difficult to align with coherent user experience.

Our findings show that ranking-based experimental designs can recover meaningful causal effects in marketplaces where standard A/B tests fail due to interference or violations of coherency. Several directions remain open. External validation using real-world marketplace data would strengthen the empirical grounding of the method. Extending the framework to user-side treatments or environments with strong cross-user interference is another important direction. Developing methods for adaptive or optimal treatment prioritization within ranking algorithms also remains a promising area for future work.

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.

Ha-Thuc, V., Dutta, A., Mao, R., Wood, M. and Liu, Y. (2020), A counterfactual framework for seller-side a/b testing on marketplaces, *in* 'Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval', pp. 2288–2296.

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* .

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

Joachims, T., Granka, L., Pan, B., Hembrooke, H. and Gay, G. (2017), Accurately interpreting clickthrough data as implicit feedback, *in* 'Acm Sigir Forum', Vol. 51, Acm New York, NY, USA, pp. 4–11.

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.

Richardson, M., Dominowska, E. and Ragno, R. (2007), Predicting clicks: estimating the click-through rate for new ads, *in* 'Proceedings of the 16th international conference on World Wide Web', pp. 521–530.

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* .