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

Mahyar Habibi[1], **Zahra Khanalizadeh**[2], Negar Ziaeian[3]

February 15, 2026

Click for the latest version

Abstract

Online marketplaces frequently run pricing experiments in environments where users choose from a list of items. In these settings, items compete for users' limited attention and demand, creating interference among items within a list: Changing prices for any item can affect the demand for others, biasing estimates from item-level A/B tests. Besides, a key consideration in pricing experiments is preserving platform coherency across prices and item availability. This requirement rules out experimental designs such as user-level A/B tests as they violate platform coherency. We propose Two-Sided Prioritized Ranking (TSPR) to estimate the total average treatment effect of price changes in such settings. TSPR exploits position bias in ranked search results to create variation in treatment exposure without compromising coherency. TSPR randomizes both users and items and reorders ranked lists, prioritizing treated items for one group of users and untreated items for the other. All users see the same items at consistent prices, but differ in exposure to treatment as they pay disproportionate attention across ranks. In semi-synthetic simulations based on Expedia hotel search data, TSPR outperforms baseline coherency-preserving experiment designs by reducing estimation bias and providing sufficient statistical power.

Keywords: experimental design, two-sided marketplaces, interference, ranking systems

[1] Lyft, Canada. Email: mahyar.habibi@phd.unibocconi.it

[2] Ph.D. Candidate in Economics, University of Washington, USA. Email: zkhnl@uw.edu. This paper is my Job Market Paper. All errors are my own. I am deeply grateful to my advisors Alan Griffith and Jason Kerwin, as well as Melissa Knox, for their 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).

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

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).

Spillovers, interference, and the estimand. Standard experimental designs rely on the Stable Unit Treatment Value Assumption (SUTVA), which rules out spillovers across units (Rubin, 1974; Imbens and Rubin, 2015). In online marketplaces this assumption is frequently violated. Items compete for users' limited attention and demand within ranked lists, so modifying treated items (for example, via discounts or price increases) can change outcomes for untreated items through substitution or complementarity. Such interference (spillovers, network dependence) has been documented in ridesharing platforms (Chamandy, 2016) and online pricing experiments (Choi and Mela, 2019). When interference is ignored, estimates from randomized experiments can be substantially biased (Blake and Coey, 2014; Fradkin, 2019). For example, if we discount Hotel A in a search result, users who would have booked Hotel B now book A instead, making B appear to perform worse, not because of its true quality, but because of demand spillovers from the treated item.

Coherency constraints in marketplace experiments. Interference alone does not preclude credible estimation: user-level randomization, where all items shown to treated users receive treatment, would avoid mixing treated and untreated items within the same query. However, this approach violates critical operational constraints that practitioners face. We formalize two such constraints. First, **price coherency**: all users must observe the same realized price (or other treatment attribute) for any given item throughout the experiment. Showing different prices to different users raises legal concerns under competition law (European Union, 2008; European Commission, 2025), creates reputational risks when customers discover the practice (WIRED Staff, 2000; Consumer Reports, 2025; Kravitz, 2025), and may alter behavior if users are aware that prices vary experimentally. Second, **full catalog**

access: all users must have access to the complete set of items throughout the experiment. Removing items from users' search results for experimental purposes conflates the treatment effect with the effect of restricting the choice set, fundamentally altering substitution patterns and market structure. These two constraints together rule out many standard experimental designs. User-level randomization violates coherency by showing different prices to different users. Item-level randomization preserves coherency but suffers from substantial bias under interference (Blake and Coey, 2014). While cluster randomization can reduce interference by grouping related units (Ugander et al., 2013; Eckles et al., 2017; Holtz et al., 2024), it requires well-defined clusters that align with spillover patterns, which is often difficult in marketplaces where interactions evolve dynamically, and can be computationally expensive to implement (Candogan et al., 2023). This creates a methodological gap: how can platforms credibly estimate treatment effects while maintaining both price coherency and full catalog access under interference?

Design idea: Two-Sided Prioritized Ranking (TSPR). We propose the Two-Sided Prioritized Ranking (TSPR) experimental design for item-side interventions in two-sided marketplaces where outcomes of interest (clicks, bookings) are observed on the user side. TSPR exploits a feature of modern marketplaces: centralized recommender systems rank items for each user query, and users exhibit strong position bias, allocating disproportionate attention to top-ranked items (Craswell et al., 2008; Friedberg et al., 2022). The key insight is that while we cannot vary treatment status across users or remove items (coherency), we *can* vary users' exposure to treatment by systematically reordering the ranked list.

Specifically, TSPR randomizes users into two groups and reorders each user's ranked list so that treated items are prioritized at the top for one group and untreated items are prioritized for the other. A placebo set of items—neither treated nor control—helps balance the average quality of items promoted to top positions across groups, ensuring that differences in outcomes reflect treatment exposure rather than item quality imbalances. This induces systematic variation in treatment exposure through position bias while preserving coherency: all users retain access to the same underlying item set, and each item's treatment status remains consistent across users.

Contributions. We introduce Two-Sided Prioritized Ranking (TSPR), an experimental design for item-side interventions in ranked-list marketplaces that maintains price coherency and full catalog access while addressing interference. Our design uses position bias—typically viewed as a confound in observational studies—as an instrument to create exogenous variation in treatment exposure. Under plausible conditions, including treatment–attention separability (treatment scales outcomes but does not change how attention is allocated across ranks) and symmetric re-ranking distortion (the ranking perturbation affects both experimental arms equally), we show that TSPR identifies the proportional effect of universal treatment relative to universal control. We provide a tractable nonlinear least squares estimator that exploits partial-outcome contrasts across experimental arms and ranks.

Using an open-source Expedia hotel search dataset, we estimate behavioral models of click and booking decisions and conduct Monte Carlo simulations to evaluate performance. TSPR substantially reduces bias relative to item-level A/B tests while recovering the ground truth treatment effect with modest increases in variance. TSPR also outperforms cluster-randomized designs on bias, even when clusters are cleanly defined, an upper bound on cluster randomization performance that is rarely achieved in practice. These results demonstrate that ranking-based designs can credibly estimate treatment effects in settings where standard methods fail due to interference or operational constraints.

Roadmap. Section 2 motivates the coherency constraint, reviews position bias as a source of identification, and explains why existing experimental designs are insufficient. Section 3 formally defines TSPR and shows how it identifies the treatment effect under interference. Section 4 describes our semi-synthetic simulation framework based on Expedia hotel search data. Section 5 demonstrates that TSPR substantially reduces bias relative to standard approaches. Section 7 concludes.

2 Motivation, Background and Related Work

2.1 Coherency constraints in practice

Many marketplace interventions cannot be tested with a standard user-level A/B test. The core constraint is *coherency*: during the experiment, users must observe the same realized item attributes, such as price and key features, and have access to the same item catalog.

First, coherency requires that users see the same item attributes, including price. This is sometimes a legal or operational requirement, and more often a reputational one. Overt price variation is tightly regulated under European competition law (European Union, 2008; European Commission, 2025), which makes price A/B tests difficult to deploy without raising compliance and reputational concerns. Even when not explicitly illegal, platforms face acute trust and brand risks. Consumer reactions to visible price dispersion are typically strong, and perceived unfairness can dominate any short-run learning benefits (Çakır et al., 2025). Recent reporting describes tests on Instacart, one of the largest online grocery marketplaces in North America, in which shoppers purchasing identical items at the same stores were charged different prices; with variation that, scaled to typical usage over a year, can imply meaningful differences in total spending on the order of $1,200 per year (Kravitz, 2025). In a nationally representative Consumer Reports survey of 2,240 U.S. adults conducted in September 2025, 72 percent of respondents who had used Instacart in the previous year reported that they did not want the company to charge different users different prices for any reason (Consumer Reports, 2025). Following these revelations and the ensuing public scrutiny, Instacart announced in December 2025 that it would end "item price tests," noting that showing different prices for the same item at the same store fell short of customer expectations (Instacart, 2025). These concerns are longstanding. Amazon's 2000 DVD pricing experiment triggered immediate backlash and led to public apologies and refunds (WIRED Staff, 2000). Since then, major platforms have become more cautious about overt price experimentation, though personalized pricing remains common in many settings. 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 can complicate interpretation of price effects.

A second coherency requirement is *full catalog access*. Many platforms cannot remove items from search results or show different item sets across users without degrading the user experience, distorting substitution patterns, and harming revenue. Designs that vary availability conflate the effect of the intervention with the effect of restricting choice sets, and they may induce strategic seller or user responses that do not reflect business-as-usual behavior.

These constraints create a gap between how marketplace experiments are typically analyzed and what platforms can actually deploy. This paper is motivated by that feasibility gap and is a step toward expanding the experimentation toolkit for marketplaces by developing designs that respect these practical coherency constraints and preserve the user experience, rather than assuming that visible price variation or choice-set manipulation are acceptable.

2.2 Position bias in ranked lists

Recommender systems and search engines shape user behavior through ranked lists. A key empirical regularity is position bias: items displayed near the top of a list receive more attention and are more likely to be clicked than those ranked lower, even holding intrinsic relevance fixed (Craswell et al., 2008; Friedberg et al., 2022). Empirical evidence shows a steep decline in click probability as an item moves down the ranking (Friedberg et al., 2022). Behavioral models provide mechanisms for this pattern. The examination hypothesis posits that users must first examine an item before deciding whether to click, while cascade models propose that users inspect items sequentially and may stop after finding a satisfactory option (Craswell et al., 2008; Richardson et al., 2007). Joachims et al. (2017) discuss trust bias, whereby users place excessive confidence in the ranking algorithm. Together, these models imply that observed clicks combine position and relevance effects. TSPR exploits position-driven exposure changes induced by systematic re-ranking to identify the platform-wide effect of an item-side intervention.

2.3 Existing experimental designs in marketplaces

User-level A/B tests. A natural starting point is to randomize users into treatment and control groups and apply the item-side intervention to all items shown to treated users. This design is attractive statistically because it avoids mixing treated and untreated versions *within* a treated user's session. However, for price interventions it typically violates the coherency constraint that motivates our setting: the same item can be displayed at different prices to different users. Such cross-user price dispersion is often infeasible in practice and may create legal, reputational, and internal-validity concerns if users perceive or are told that prices vary due to experimentation.

Item-side A/B tests. A common coherency-preserving baseline is to randomize items into treated and control groups and expose all users to the same realized treatment status for each item. This preserves price coherency, but it can be biased under interference because treated and untreated items appear together in ranked lists and compete for attention and demand.

Cluster randomization. Cluster-based randomization groups related users or items to reduce between-cluster spillovers, and has been studied in network experimentation and online marketplaces (Ugander et al., 2013; Eckles et al., 2017; Holtz et al., 2024). Its performance depends on how well cluster boundaries align with spillover patterns. In many marketplace settings, defining suitable clusters is difficult because user interactions and item relationships evolve over time, and poorly chosen clusters can reduce power and yield unreliable estimates. Implementing cluster-based randomization can also be computationally expensive and operationally complex (Candogan et al., 2023).

Crossover or switchback designs. Switchback testing alternates treatment assignments over time for the same units (Brown Jr, 1980; Robins, 1986; Sneider and Tang, 2019; Bojinov et al., 2023). While switchbacks can support causal identification in time-varying environments, frequent treatment fluctuations can confuse users and distort engagement patterns. For salient interventions such as prices, these fluctuations may also create carryover effects that undermine internal validity.

Two-sided randomization (TSR). Two-sided randomization methods apply randomization on both the user side and the item side (Johari et al., 2022; Bajari et al., 2023). Standard TSR implementations apply treatment only when a treated user interacts with a treated item, which can lead different users to see different versions of the same item (including different prices). This violates the coherency requirement that motivates our setting. In contrast, TSPR enforces a consistent realization of item treatment across all users while still using user-level randomization to create exposure variation.

Related concepts: ranking experimentations. Prior work on interference in ranking experiments often focuses on evaluating or improving ranking algorithms (e.g., Goli et al., 2024; Zhan et al., 2024; Nandy et al., 2021; Ursu, 2018). Our objective is different: we do not treat the recommender system as the object of experimentation, but instead use it as the mechanism through which item-side treatment exposure is shifted while preserving coherency. This differs from approaches that rely on naturally occurring ranking noise as exogenous variation, and from interleaving-style methods that primarily optimize ranking quality rather than deliver coherent item-side interventions.

3 Methodology

3.1 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

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

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. This estimand is in one-to-one correspondence with the total average treatment effect emphasized in the interference literature (Manski, 2013; Munro et al., 2024).

The proposed method rests on several assumptions. 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.[4]

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 experiment intensity parameter p, 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

[4] Extending the design to settings with binding capacity constraints or other forms of user-side interference across queries or users is left to future work.

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

TSPR Experiment Setup

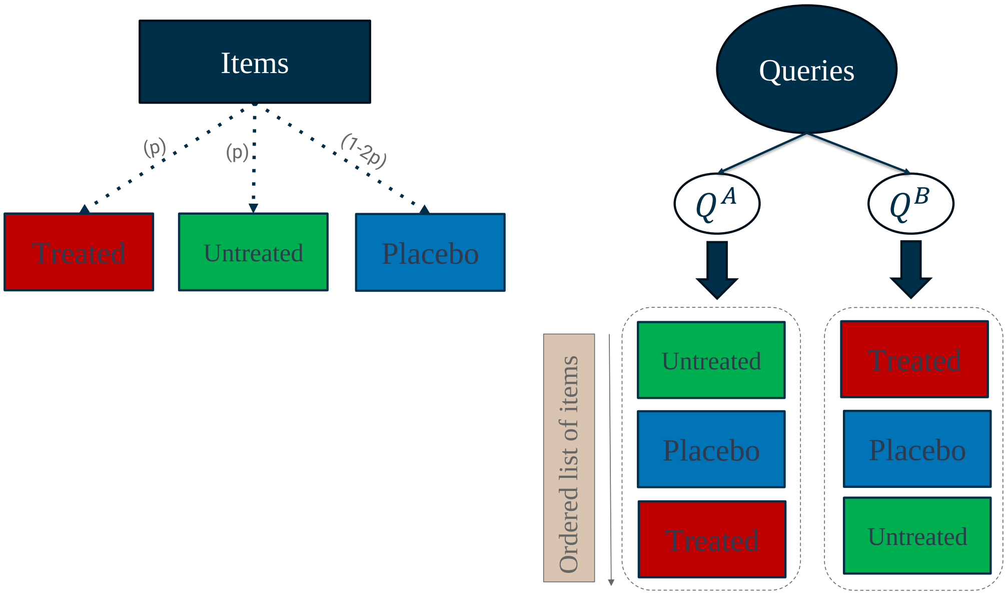
1. Set the probability of receiving treatment for an item $p < 0.5$.
2. Randomize items into Treated, Untreated, and Placebo subsets with probabilities p, p, and $1 - 2p$, respectively. Apply the treatment only to the Treated group.
3. For each incoming query q:

   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. 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.

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 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.

Placebo items also create a buffer between treated and untreated items in the ranked list. This separation sharpens the interpretation of rank depth as treatment exposure in our partial-outcome contrasts: at small depths, outcomes are driven primarily by exposure to the prioritized block rather than by immediate mixing of treated and untreated items. As a result, placebo reduces contamination of the control arm from treated items and improves the signal-to-noise ratio of the within-depth contrasts that identify lift.

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.

3.2 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. It is defined as the proportional lift in the total outcome under universal treatment (all items treated) relative to universal control (all items untreated) (Equation 1). Thus, Φ represents the percentage change in expected total outcomes when all items are treated.

A TSPR experiment is characterized by the set of treatment-prioritized queries Q^B, the set of control-prioritized 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, $F(l)$, is defined so that under a given ranking: $\mathbb{E}[Y^l] = F(l)\mathbb{E}[Y]$, where $F : \mathbb{N} \to (0, 1)$ is increasing and concave, and $F(l) \to 1$ as $l \to \infty$.

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 but still under original recommender system, the expected partial outcome would satisfy

$$\mathbb{E}[Y^l \mid \textit{full treatment}] = (1 + \Phi) \cdot F(l) \cdot \mathbb{E}[Y \mid \textit{no treatment}]. \tag{2}$$

We now characterize how moving from the platform's original ranking to the TSPR ranking experiment (Table 1) changes expected partial outcomes. There are two channels. First, re-ranking can change user attention, meaning how attention is allocated across positions (for example, time spent evaluating items and clicks). Second, it can change outcomes through the treatment itself. Assumptions 2 and 3 formalize the distortion induced by re-ranking and how it affects partial outcomes under TSPR. Assumptions 4 and 5 then describe how the partial treatment exposure queries receive affects partial outcomes, both when treated items are prioritized at the top for group B queries and when treated items are down-ranked for group A queries.

In a TSPR experiment, the platform perturbs its baseline relevance ordering to induce exogenous variation in exposure. Such changes alter how user 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. When items that would appear lower under the platform's baseline ranking are moved into early positions, users may click less, search less deeply, or abandon sooner. The magnitude of this perturbation is governed by the treatment assignment probability p from the experimental design: larger p implies a larger expected deviation from the platform's baseline ordering within early ranks. We capture this channel by allowing the baseline attention function $F(l)$ to be attenuated under TSPR, and denote the distorted attention function by $D(F(l); p)$.

Assumption 2 (Multiplicative Distortion). TSPR re-ranking distortion attenuates attention multiplicatively:

$$D(F(l); p) = d(l; p)\, F(l),$$

where $d(l; p) \in (0, 1]$ is a depth-l attenuation factor that depends on the treatment assignment probability p.

Assumption 3 (Symmetric Distortion). Conditional on the treatment assignment probability p, the re-ranking attenuation is identical across experimental arms. That is, for all depths l,

$$d_A(l; p) = d_B(l; p) \equiv d(l; p).$$

Equivalently, TSPR induces the same expected attention distortion in arms A and B.

Assumption 3 is motivated by the symmetry of the TSPR design. Items are randomly assigned to Treated, Untreated, and Placebo labels, independently of their baseline relevance. As a result, the distributions of baseline relevance among Treated and Untreated items are identical in expectation. The two query arms then apply mirror-image block prioritization rules: arm B promotes the Treated block while arm A promotes the Untreated block, and in both arms items are otherwise the same and only re-ordered within a fixed candidate set. When within-block ordering follows the platform's baseline ranking, the primary source of perturbation is the block swap itself, whose magnitude is governed by the treatment assignment probability p. Under these conditions, the expected quality of the top-l positions, and therefore the induced attenuation in attention, is the same across arms, implying $d_A(l;p) = d_B(l;p)$ for all l.

Equation (2) 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.

Invoking Assumptions 1, 2, and 3, and writing $d(l)$ for the distortion function at a fixed treatment probability p, the expected partial outcome at rank l for a query assigned to group B under TSPR satisfies

$$\mathbb{E}[Y^l \mid \text{Full treatment}, \text{TSPR}] = (1+\Phi)\, d(l)\, F(l)\, \mathbb{E}[Y \mid \text{No treatment}, \text{original ranking}]. \quad (3)$$

Similarly, for a query in group A in TSPR,

$$\mathbb{E}[Y^l \mid \text{No treatment}, \text{TSPR}] = d(l)\, F(l)\, \mathbb{E}[Y \mid \text{No treatment}, \text{original ranking}]. \quad (4)$$

We now 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.

Building on equation 3, for a query in group B with $l \leq n_b$ treated items at the top:

$$\mathbb{E}[Y^l_B \mid \text{TSPR}] = (1+\tau(n_b)\Phi)\ d(l)\ F(l)\ \mathbb{E}[Y \mid \text{No treatment}]. \quad (5)$$

Assumption 4 (**Partial Treatment Effect**)**.** $\tau : \mathbb{N} \to \mathbb{R}_+$ satisfies $\tau(l) \to 1$ as $l \to \infty$. The function may converge from below ($\tau(1) < 1$) with a concave shape in l, from above ($\tau(1) > 1$) with a convex shape in l, or be constant ($\tau(\cdot) = 1$).

Assumption 4 ensures that partial lift has the same sign as the full-treatment effect, and that $\phi(l) = \Phi\, \tau(l)$ converges to Φ as the treated block grows. The sign of $\tau(1) - 1$ is a reduced-form summary of net interference at shallow depths. When items are substitutes, treating a small block at the top *amplifies* the per-item effect: with $\Phi < 0$, for instance, the single treated item at rank 1 loses demand to the many untreated items below it, making $\phi(1)$ more negative than Φ and hence $\tau(1) > 1$. Conversely, if items are complements, where the treatment effect grows with the number of co-treated items, treating a single item in isolation produces a smaller effect than treating the full catalog, yielding $\tau(1) < 1$. As the treated block expands and fewer untreated items remain to absorb substitution (or contribute complementarities), $\tau(l)$ declines toward 1.

In our main application (an Expedia-like marketplace), items are substitutes in expectation: booking one hotel forgoes others in the same query. This substitution channel implies $\tau(1) > 1$, with $\tau(l)$ decreasing toward 1 as l grows.

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:

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

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

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

In practice, the contamination term $\nu(\cdot)$ is small whenever treated items in the tail of group A listings receive little effective attention, either because attention decays sharply with rank, because the list is long, or because p is small so that the placebo buffer separating untreated and treated blocks is wide.

We therefore adopt the approximation $\nu \approx 0$ for the remainder of the analysis, reducing the partial lift to the single-function form $\phi(l) = \Phi\,\tau(l)$. This simplification is what allows us to recover Φ from the depth profile of $\widehat{\phi}(l)$ using either the parametric or nonparametric estimators described below.

3.3 Estimation

Given data from a TSPR experiment, we first construct the empirical partial lift at each depth l. For each block size l, we compare queries from Q^B that have exactly l Treated items in the top positions to queries from Q^A that have exactly l Untreated items in the top positions, computing the partial outcome up to position l in both cases. The empirical partial lift is

$$\widehat{\phi}(l) \;=\; \frac{\widehat{\mathbb{E}}[Y_B^l]}{\widehat{\mathbb{E}}[Y_A^l]} \;-\; 1, \tag{8}$$

where $\widehat{\mathbb{E}}[Y_B^l]$ and $\widehat{\mathbb{E}}[Y_A^l]$ denote the sample means of the partial outcome in the treatment-prioritized and control-prioritized arms, respectively, conditional on block size l.

With empirical values $\widehat{\phi}(l)$ for a range of depths $l = 1, \ldots, L$, and under standard regularity conditions (e.g., sufficient support across l), we recover the global lift Φ using either a parametric or a nonparametric approach. Both methods exploit the relationship $\phi(l) = \Phi\,\tau(l)$, where $\tau(l) \to 1$ as $l \to \infty$, so that $\phi(l)$ converges to the full-treatment effect as depth increases.

3.3.1 Parametric Estimation: Weighted Least Squares

We impose a parsimonious parametric form on the depth function that satisfies the qualitative restrictions derived above; namely, that $\tau(l)$ is smooth, positive, and converges to unity as l grows:

$$\tau(l) \;=\; 1 + \frac{1}{l}. \tag{9}$$

This specification has no free parameters beyond the global lift Φ itself. At $l = 1$, $\tau(1) = 2$: when only a single treated item occupies the top position, the partial lift is twice the full-treatment effect, reflecting the concentrated exposure of the treatment block. As l increases, $\tau(l)$ declines monotonically toward 1, so that $\phi(l) \to \Phi$. While we also explored a generalized specification $\tau(l) = 1+\alpha/l$ where α is estimated as a free parameter, we found that this added flexibility yielded no significant improvement in bias reduction; consequently, we maintain the simpler, parameter-free form for all subsequent results.

Since the model $\phi(l) = \Phi\,\tau(l)$ is linear in Φ for known τ, the estimator admits a closed-form weighted least squares solution. We minimize

$$Q(\Phi) \;=\; \sum_{l=1}^{L} w(l)\left[\widehat{\phi}(l) \;-\; \Phi\,\tau(l)\right]^2, \tag{10}$$

where $w(l)$ is a nonnegative precision weight proportional to the number of queries contributing to the l-th partial outcome. Setting the first-order condition to zero yields

$$\widehat{\Phi} \;=\; \frac{\displaystyle\sum_{l=1}^{L} w(l)\;\tau(l)\;\widehat{\phi}(l)}{\displaystyle\sum_{l=1}^{L} w(l)\;\tau(l)^2}. \tag{11}$$

This is equivalent to dividing each $\widehat{\phi}(l)$ by its model-implied scaling factor $\tau(l)$ and taking a precision-weighted average of the resulting depth-specific estimates of Φ. Standard errors are computed via bootstrap resampling at the query level.

3.3.2 Nonparametric Estimation: Isotonic Regression

As a robustness check that avoids imposing a specific functional form on $\tau(l)$, we estimate Φ using isotonic regression (Barlow, 1972). This approach follows recent work in marketplace analytics that utilizes shape restrictions to correct for rank-based biases (e.g., Goli et al., 2024). While Goli et al. employ isotonic regression to improve recommender system accuracy, we leverage it here to nonparametrically recover the global lift Φ from item-level interventions. The key shape restriction is that $|\phi(l)|$ is monotonically decreasing in l: as more items in the top block receive treatment (or control), the partial lift converges toward the full-treatment effect. When $\Phi < 0$, this implies that $\phi(l)$ is increasing; when $\Phi > 0$, $\phi(l)$ is decreasing.

We fit a weighted isotonic regression of $\widehat{\phi}(l)$ on l, using the same precision weights $w(l)$ as above, subject to the appropriate monotonicity constraint. Block sizes with fewer than a minimum number of queries in either arm are excluded prior to fitting to reduce noise from imprecisely estimated cells.

Let $\widetilde{\phi}(l)$ denote the isotonic-regression fitted values. Because $\phi(l) \to \Phi$ as l grows, we estimate the global lift as a weighted average of the fitted values at large depths:

$$\widehat{\Phi}_{\text{iso}} = \frac{\sum\limits_{l \in \mathcal{L}_{\text{top}}} w(l)\, \widetilde{\phi}(l)}{\sum\limits_{l \in \mathcal{L}_{\text{top}}} w(l)}, \tag{12}$$

where $\mathcal{L}_{\text{top}}$ contains the largest 30% of observed block sizes (with a minimum of two blocks). As with the parametric estimator, standard errors are obtained by bootstrap resampling at the query level.

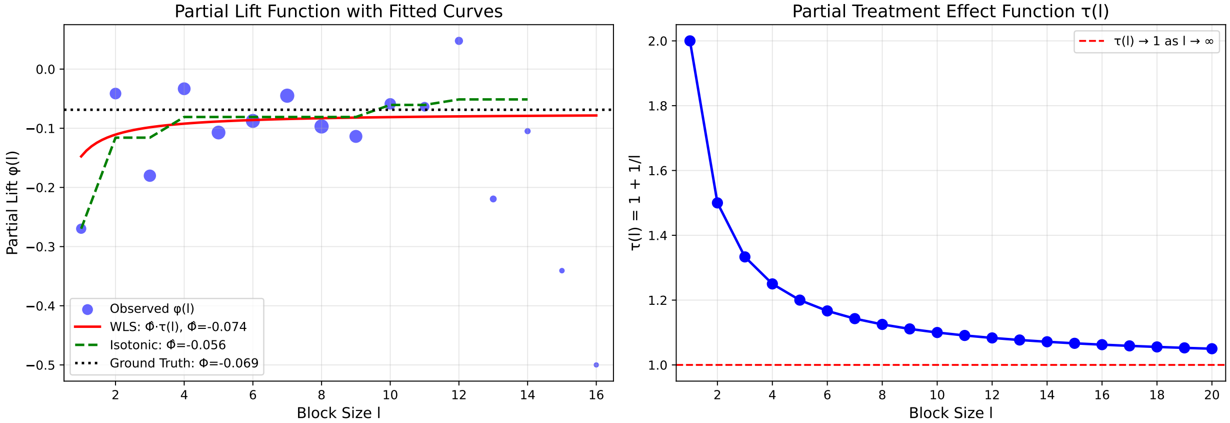


Figure 2: Estimation of Global Lift Φ via TSPR. The left panel shows empirical partial lifts $\hat{\phi}(l)$ (blue dots), with marker size proportional to query-level precision weights $w(l)$. The parametric WLS fit (red) and nonparametric isotonic regression (green dashed) both recover the ground truth Φ (dotted black) as block size increases. The right panel illustrates the structural decay function $\tau(l) = 1 + 1/l$, which provides the scaling factors necessary to map partial lifts back to the global treatment effect.

Figure 2 illustrates the performance of our estimators using a representative simulation.

As shown in the left panel, the empirical partial lifts $\hat{\phi}(l)$ are initially higher in magnitude than the global effect Φ, reflecting the concentrated treatment exposure at low block depths. As the block size l increases and the buffer of treated items grows, the observed lifts converge toward the ground truth. The parametric WLS fit and the isotonic regression both successfully capture this transition, effectively averaging the noisy empirical points according to their query-level precision weights. The right panel highlights the structural decay of $\tau(l)$, which serves as the fundamental scaling mechanism that allows TSPR to recover the full-treatment effect from partial-list interventions.

4 Data and Simulation Setup

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 over an eight-month period. Our training and calibration sample consists of a 20% subset of the cleaned data, comprising nearly 2 million observation-level records across approximately 80,000 unique search impressions.

Consumers interact with the platform in three stages. First, consumers initiate queries by specifying trip details. Second, they receive a ranked list of hotel results. A key feature of this dataset is the experimental variation in ranking: approximately 30% of search impressions were randomly sorted, while the remainder followed the platform's original relevance-based recommender system. This variation allows us to disentangle the causal effect of display position from item relevance. Finally, users engage by clicking on hotels to view details and may subsequently complete a booking.

To evaluate our experimental design, we implement Monte Carlo simulations that replicate this two-sided marketplace. We model user interactions as a function of a latent utility v_{ij}. The platform's relevance score r is modeled as $r = v + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. While the original ranking is sorted decreasing in r, the random ranking allows for unbiased estimation of position effects.

Table 2 presents summary statistics for the search impressions used in our analysis, highlighting the baseline differences in performance between the ranking mechanisms.

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
— Random ranking	0.13	1	0	1
— Original ranking	0.93	1	0	1

4.1 Click Model

Click behavior is modeled using a logistic function incorporating rank-based attention and sequential behavior. For each item j at position p for user i, the probability of a click is:

$$P(\text{click}_{ij}) = \text{logit}^{-1}\Big(\beta_1 p_{ij} + \beta_2 p_{ij}^2 + \beta_3 \text{prevclicks}_i + \beta_4 \mathbf{1}[\text{prevclicks}_i > 0] + \beta_5 v_{ij} + \beta_0\Big) \quad (13)$$

We estimate these parameters in two stages to resolve the endogeneity of position in relevance-sorted results. In the *first stage*, we use only the randomly sorted subset to estimate the position coefficients (β_1, β_2) and click-history effects (β_3, β_4). Because positions are assigned randomly, these estimates represent pure attention effects. In the *second stage*, we fix these coefficients as an offset and estimate the utility coefficient β_5 and intercept β_0 on the full mixed sample. The resulting parameters, shown in Table 3, show a clear initial decline in attention (negative β_1) and a strong negative pressure on subsequent clicks once an initial click has occurred (β_4).

Table 3: Estimated Click Model Parameters

Parameter	Variable	Estimate
β_1	Position (p)	-0.0697
β_2	Position Squared (p^2)	0.0025
β_3	Previous Clicks Count	0.6550
β_4	Has Clicked (Indicator)	-3.6033
β_5	Latent Utility (v)	0.0897
β_0	Intercept	-1.7873

The resulting estimates (Table 3) reveal a nuanced search process. The large negative coefficient for the click indicator $(\beta_4 = -3.60)$ captures the expected "satisficing" effect, where any initial click significantly lowers the marginal probability of further search. However, the positive coefficient for the cumulative click count ($\beta_3 = 0.65$) identifies latent searcher heterogeneity: conditional on not stopping, users with higher click counts exhibit a higher baseline propensity for exhaustive search. This specification ensures the simulation reflects a marketplace populated by both "quick-search" and "high-intensity" consumers. Figure 3 (left) shows that our click model closely fits the click-through-rate based on item position to the real Expedia click impressions data.

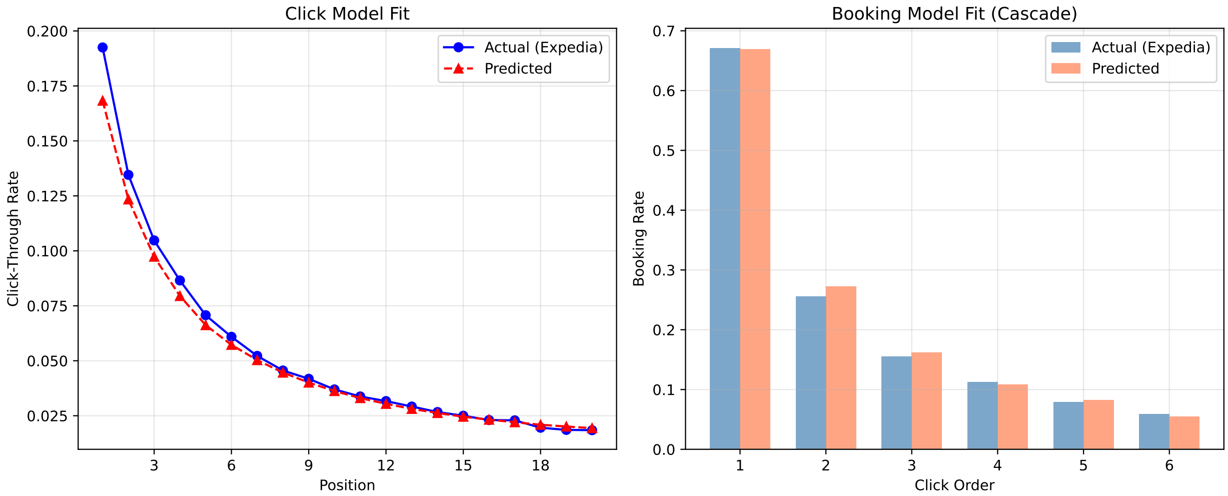


Figure 3: Model fit for click and booking behavior. *Left:* Actual vs. predicted click-through rate by display position. *Right:* Actual vs. predicted booking rate by click order under the cascade booking model. Both panels are computed on a hold-out sample not used in estimation.

4.2 Booking Model

Conditional on the set of clicked items C_i, user i either books one item or takes the outside option (no booking). We model this choice with a multinomial logit that includes a click-order effect to capture cascade behavior in purchasing decisions. The probability of booking item $k \in C_i$ is

$$P(\text{book}_{ik} \mid C_i) \;=\; \frac{\exp\big(\gamma_1\, v_{ik} + \gamma_2\, \text{click\_order}_{ik} + \gamma_0\big)}{1 + \sum_{k' \in C_i} \exp\big(\gamma_1\, v_{ik'} + \gamma_2\, \text{click\_order}_{ik'} + \gamma_0\big)}, \tag{14}$$

where click_order$_{ik}$ records the sequential position in which item k was clicked ($1 = $ first clicked, $2 = $ second, etc.) and the "1" in the denominator represents the outside option. The coefficient γ_2 captures the empirical regularity that earlier-clicked items are more likely to be booked, consistent with directed search in which users click the most promising options first. All parameters are estimated by maximum likelihood on the subsample of clicked items.

Table 4: Estimated Cascade Booking Parameters

Parameter	Variable	Estimate
γ_1	Latent Utility (v)	0.4396
γ_2	Click Order	-0.1774
γ_0	Intercept	0.4271

The negative coefficient for click order (γ_2) in Table 4 confirms a "first-mover" advantage in search, where items discovered earlier in the process are more likely to be converted. A likelihood ratio test strongly rejects the standard multinomial logit (without click order) in favor of the cascade specification ($p < 0.001$; see Appendix A.2), so we adopt the cascade model for all main results. Figure 3 (right) confirms that the cascade model closely matches observed booking rates by click order.

4.3 Treatment and Interference

Treatment is introduced as a constant shift δ in the latent utility: $v^*_{ij} = v_{ij} + \delta T_{ij}$. This shift propagates through both the click and booking stages. Crucially, as shown in Equation 14, an increase in the utility of a treated item increases its own booking probability while simultaneously decreasing the probability for all other items in C_i. This within-query substitution is the primary source of interference we aim to address with the TSPR design.

5 Baselines and Results

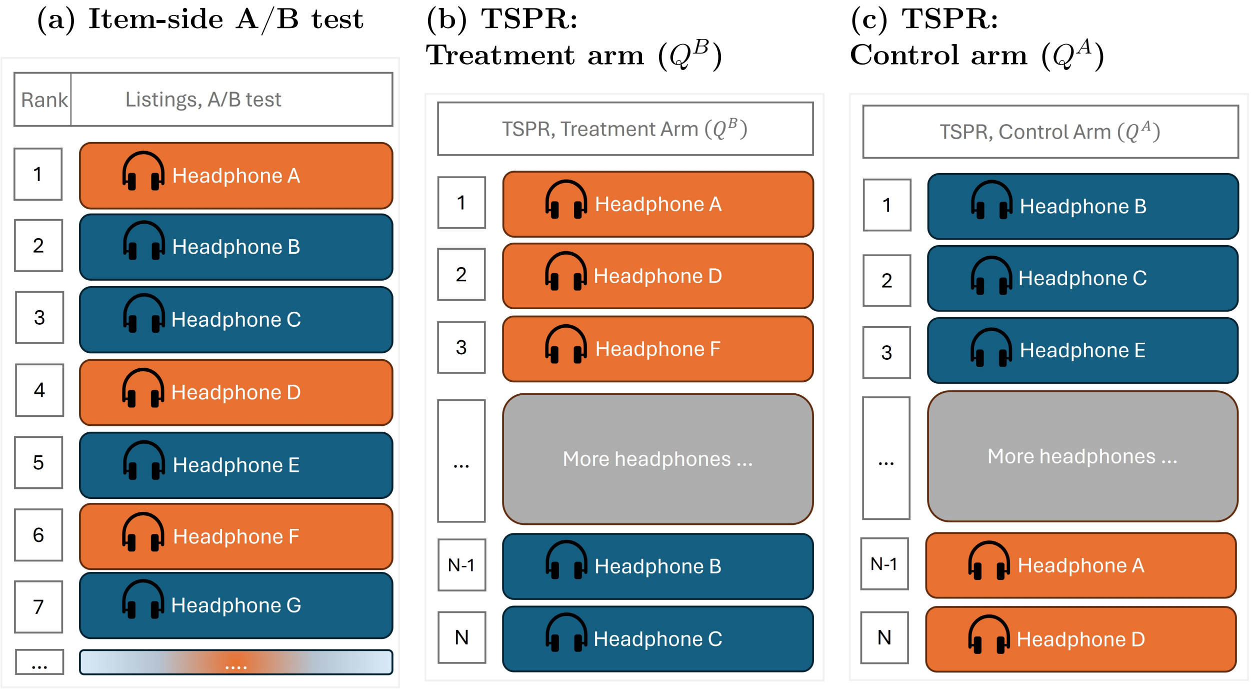
We conduct counterfactual simulations for 20,000 queries using the estimated models of click and booking behavior. To establish a simulated ground truth for lift, we simulate the marketplace under two extreme scenarios: one in which no items receive treatment and one in which all items are treated. The 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 6.9 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, described in Table 1 and estimate the total lift using both the parametric (Section 3.3.1) and non-parametric (Section 3.3.2) approaches. The main results are under treatment probability $p = 0.25$ but we show sensitivity to choice of p in Appendix A.4.

5.1 Performance Baseline: Bernoulli-Randomized A/B Testing

As a baseline, we consider an item-side randomized experiment in which items are Bernoulli randomized at the listing level. Figure 4 contrasts this design with the Two-Sided Prioritized Ranking (TSPR) setup. In the item-side A/B test (panel a), treated and untreated items are randomly interleaved within the same ranked list, so treated items compete directly with untreated items for user attention, generating within-list interference. TSPR instead induces structured variation in treatment exposure through ranking priorities while preserving full catalog access. In the treatment arm (Q^B, panel b), treated items are promoted to the top of the ranking, whereas in the control arm (Q^A, panel c), untreated items are prioritized, with placebo items buffering the two blocks. We simulate both designs and compare the resulting lift estimates.

To estimate the lift in total outcome, we extend the Horvitz and Thompson (1952) logic to the two-sided marketplace setting using item-level randomization. In this baseline design, randomization occurs only at the level of items. Each item i is independently assigned to treatment ($Z_i = 1$) with probability p or control ($Z_i = 0$) with probability $1 - p$, forming the sets $T = \{i : Z_i = 1\}$ and $C = \{i : Z_i = 0\}$. Letting $y_{q,i}$ denote the outcome for item i in query $q \in Q$, we estimate the mean total outcome per query under global treatment (μ_B) and global control (μ_A) using the estimators $\hat{\mu}_B^{IS} = \frac{1}{|Q|} \sum_{q \in Q} \sum_{i \in T} \frac{y_{q,i}}{p}$ and $\hat{\mu}_A^{IS} = \frac{1}{|Q|} \sum_{q \in Q} \sum_{i \in C} \frac{y_{q,i}}{1-p}$. Our estimand of interest is the lift $\Phi_{IS} = \frac{\mu_B}{\mu_A} - 1$. By taking the ratio of



(a) Item-side A/B test

(b) TSPR: Treatment arm (Q^B)

(c) TSPR: Control arm (Q^A)

Figure 4: Illustrative comparison of an item-side A/B test and the Two-Sided Prioritized Ranking (TSPR) design. Orange items indicate treated items, blue items indicate untreated items, and gray items indicate placebo items. Panel (a) shows a Bernoulli-randomized item-side experiment in which treated and untreated items are interleaved throughout the ranking, generating within-list interference. Panels (b) and (c) show examples of the two TSPR query arms: in the treatment arm (Q^B), treated items are prioritized at the top of the ranking, while in the control arm (Q^A), untreated items are prioritized.

our HT estimators, the $1/|Q|$ terms cancel, yielding the item-side lift estimator:

$$\hat{\Phi}_{IS} = \frac{\sum q \in Q \sum_{i \in T} \frac{y_{q,i}}{p}}{\sum_{q \in Q} \sum_{i \in C} \frac{y_{q,i}}{1-p}} - 1 \tag{15}$$

This baseline utilizes only item-level Bernoulli randomization and provides a simple comparison point that ignores query-level randomization.

5.2 Performance Baseline: Cluster-Randomized Experiments

As a second baseline, we compare our estimates to lift ratio estimates obtained from cluster-randomized experiments. Cluster randomization reduces interference bias because units within a cluster share the same treatment assignment, which limits spillover across treatment arms. However, clustering methods often exhibit substantially larger variance, as the clusters are frequently large, which effectively reduces the number of independent units of randomization. Furthermore, implementing cluster randomization requires detailed knowledge of the underlying network structure and is often costly. When it can be applied correctly, 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 4. Each observation consists of a property j appearing in a search query i. Similar to the approach in Holtz et al. (2024), we begin with the co-occurrence of properties across different search queries to construct our clusters. We employ Truncated SVD to find a lower-dimensional dense representation of the co-occurrence matrix and subsequently use k-means to cluster the resulting embeddings.

The dimension of the SVD and the number of clusters k are treated as hyperparameters. We perform a grid search to identify the optimal combination based on the Modularity score. Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. The grid search resulted in selecting 100 dimensions for SVD and 200 clusters for K-means. The resulting modularity score exceeds 0.93, indicating that this method was able to identify highly segregated clusters of properties.

After clustering, we randomly assign clusters to treatment with a probability p_{treat}. For estimation, we compare mean total outcome per query between treated and untreated items, similar to the estimation for Bernoulli-Randomized A/B testing method discussed above.

5.3 Main Results: Bias and Efficiency

The central results of our simulation study are summarized in Figure 5 and Table 5.

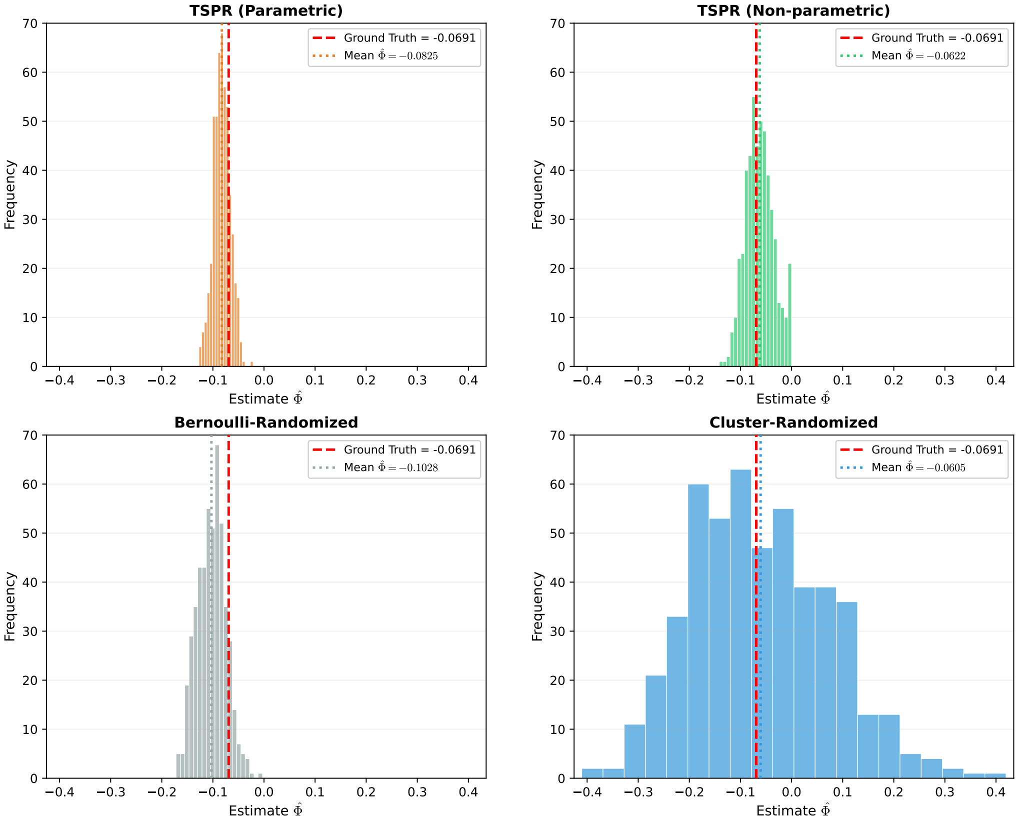


Figure 5: **Comparison of TSPR Estimators against Randomized Baselines.** Histograms showing the distribution of the estimated treatment effect $\hat{\Phi}$ for parametric (WLS) and non-parametric (isotonic) TSPR, compared to Bernoulli-randomized and cluster-randomized designs. The ground truth is indicated by the red dashed line.

Figure 5 compares the performance of the parametric (WLS) and non-parametric (isotonic) estimators under a TSPR experiment to Bernoulli-randomized and cluster-randomized baselines. Across 500 simulation replications, TSPR delivers a substantial improvement over standard marketplace experimental designs, regardless of the estimation method. Relative to the Bernoulli-randomized (naive) baseline, TSPR effectively eliminates the severe interference bias that arises when treated and untreated items compete for attention within the same ranked list. While cluster randomization produces a mean estimate close to the ground truth (low bias), it does so at the cost of much higher variance. Specifically, the cluster-randomized design exhibits standard deviations that are 8.3× and 4.9× those of the TSPR parametric and non-parametric estimates, respectively.

Table 5 quantifies these patterns in terms of bias, dispersion, and overall accuracy. The Bernoulli-randomized design exhibits substantial downward bias (mean $\hat{\Phi} = -0.1028$ vs. $\Phi_{true} = -0.0691$), yielding the largest bias in magnitude (-0.0337) and a correspondingly high RMSE (0.0438). Cluster randomization largely removes bias (bias = 0.0086), but its performance is dominated by extreme variability (empirical SD = 0.1345), producing by far the worst RMSE (0.1348). In contrast, both TSPR estimators achieve a markedly better bias–variance tradeoff. The parametric TSPR estimator attains the lowest RMSE (0.0210) and the smallest empirical SD (0.0162), despite a modest bias (-0.0133). The non-parametric TSPR estimator exhibits a very low bias (bias = 0.0069) while maintaining low variance (SD = 0.0275) and a low RMSE (0.0283). Overall, the table shows that TSPR improves accuracy relative to Bernoulli randomization by sharply reducing interference-driven bias, while avoiding the prohibitive variance costs of clustering.

Table 5: Summary Statistics of Estimator Performance ($\Phi_{true} = -0.0691$)

Estimator	**Mean $\hat{\Phi}$**	**Bias**	**RMSE**	**Empirical SD**
Bernoulli-Randomized	-0.1028	-0.0337	0.0438	0.0279
Cluster-Randomized	-0.0605	0.0086	0.1348	0.1345
TSPR (Parametric)	-0.0825	-0.0133	0.0210	0.0162
TSPR (Non-parametric)	-0.0622	0.0069	0.0283	0.0275

We assess sensitivity to the treatment probability p. While our main results use $p = 0.25$, we stress-test TSPR over $p \in [0.10, 0.45]$, which corresponds to placebo buffer sizes ranging from 80% to 10%. As shown in Appendix A.4, TSPR remains robust across this range, consistently delivering substantially lower bias and variance than the Bernoulli-randomized baseline.

Bootstrap confidence intervals for the TSPR non-parametric estimator achieve 93% empirical coverage at the nominal 95% level, indicating well-calibrated uncertainty quantification (Appendix A.3).

6 Discussion

The Bernoulli-randomized item-side estimator is unbiased only under knife-edge assumptions: each item's outcome must be independent of all other items in the list (no substitution), and user attention must be invariant to rank. In marketplace search, both fail. Users substitute across items (choosing one option crowds out others) and attention is heavily concentrated at the top of the ranking. When these features are present, the naive estimator systematically overstates the treatment effect because treated items displace untreated items from high-attention positions, conflating the causal effect with a reallocation of exposure and demand.

TSPR, in contrast, leverages these same features for identification. Position bias generates meaningful variation in exposure across ranks, and substitution ensures that shifting exposure changes aggregate outcomes across the two query arms. If user behavior were order-invariant, with clicking and booking depending only on the set of utilities $\{v_{qi}\}$ and not their positions, then permuting the ranking would not change outcomes, and TSPR would have no identifying variation. TSPR is therefore most valuable in environments where position bias is strong, a condition that is well documented in search and recommendation settings.

Cluster randomization provides an alternative by assigning groups of related items to the same treatment status, eliminating within-cluster spillovers. When cluster boundaries align with the true interference structure, this approach can achieve low bias, as in our simulation where spectral clustering yielded modularity above 0.93. The efficiency cost, however, is substantial: with roughly 200 clusters instead of 20,000 query-level randomization units in TSPR, the effective sample size falls by about two orders of magnitude, producing roughly 8 times the standard deviation of TSPR's parametric estimator. Moreover, clean clusters are difficult to construct in practice. Substitution patterns are often diffuse and contextdependent, and misspecified boundaries reintroduce the very spillovers clustering is meant to remove. TSPR avoids this fragility, requiring only the ability to modify ranking priorities within the existing recommender system.

7 Conclusion

This paper introduces Two-Sided Prioritized Ranking (TSPR), an experimental design for item-side interventions in online marketplaces that maintains price parity and full catalog access while addressing interference through position-based exposure variation. In simulations calibrated to hotel search data, TSPR substantially reduces bias relative to Bernoulli-randomized A/B tests and achieves an order-of-magnitude reduction in variance compared to cluster randomization, even when clusters are cleanly defined.

TSPR can be implemented by adjusting ranking priorities within an existing recommender system, making it straightforward to deploy on platforms that already support re-ranking logic. The design avoids the operational and data burdens that often accompany marketplace experimentation, such as user-level pricing changes, catalog partitioning, or explicitly modeling the interference network. Although TSPR does not target the global treatment effect without bias, it offers a strongly favorable bias–variance tradeoff in practice: relative to Bernoulli randomization it substantially reduces both bias and variance, and relative to cluster randomization it delivers far higher precision.

TSPR is best suited to settings with strong position bias, slack supply, and short experiment horizons that limit dynamic feedback. The identifying assumptions of treatment-attention separability and symmetric re-ranking distortion are plausible in many marketplace settings, but they can fail if the treatment meaningfully changes user engagement or if high treatment probabilities induce ranking perturbations that users do not tolerate. Practitioners should assess these risks using pre-experiment diagnostics and the sensitivity analyses reported in the appendix.

Natural extensions include adapting TSPR to supply-constrained settings with binding capacity constraints, extending it to continuous or multi-arm treatments, and allowing for user-side interference across queries. More broadly, the results underscore that platform structure, here ranked attention, can provide identifying variation even under interference. We view the coherency constraints that motivate TSPR as realistic operational requirements, and a useful guide for methodological development in platform experimentation.

A Appendix

A.1 Diagnostics for Identifying Assumptions

We provide simulation-based diagnostics for the two main identifying assumptions of the TSPR framework: Assumption 1 (Attention and Treatment Separability) and Assumption 3 (Symmetric Distortion). We refer to experiments in which the TSPR re-ranking mechanism is applied but no treatment effect is imposed ($\tau = 0$) as *dummy experiments.* In such experiments, any systematic difference in outcomes across arms must arise from the design rather than from treatment, providing a direct test of the assumptions.

Separability (Assumption 1). Assumption 1 requires that treatment scales the level of outcomes without altering the shape of the attention function $F(l)$. To assess this, we simulate two counterfactual worlds under the platform's original (unmodified) ranking: one with no treatment ($\tau = 0$) and one with full treatment ($\tau = -0.5$), each averaged over 30 independent draws. Figure 6 plots the normalized cumulative attention function $F(l) = \mathbb{E}[Y^l]/\mathbb{E}[Y]$ as a function of rank l in both scenarios. Both curves are increasing and concave, consistent with Definition 3. The two profiles are nearly indistinguishable: the maximum pointwise absolute difference is $\max_l |F_{\text{treat}}(l) - F_{\text{control}}(l)| = 0.006$, confirming that treatment shifts the level of outcomes without meaningfully altering how attention is allocated across ranks.

Symmetric distortion (Assumption 3). Assumption 3 requires that the TSPR re-ranking induces the same expected attention attenuation in both experimental arms: $d_A(l;p) = d_B(l;p)$ for all ranks l. We present three complementary diagnostics.

(i) Dummy experiment: partial-outcome differences under null treatment. We conduct dummy experiments in which the TSPR re-ranking is applied but no treatment effect is imposed ($\tau = 0$). Under symmetric distortion, $\mathbb{E}[Y_B^l] - \mathbb{E}[Y_A^l] = 0$ for all l. Figure 7 plots these differences for treatment assignment probabilities $p \in \{0.10, 0.20, 0.30, 0.40\}$, each averaged over 50 independent draws. Across all values of p and all ranks, the differences remain small (maximum absolute mean difference < 0.002, relative to baseline partial outcomes of

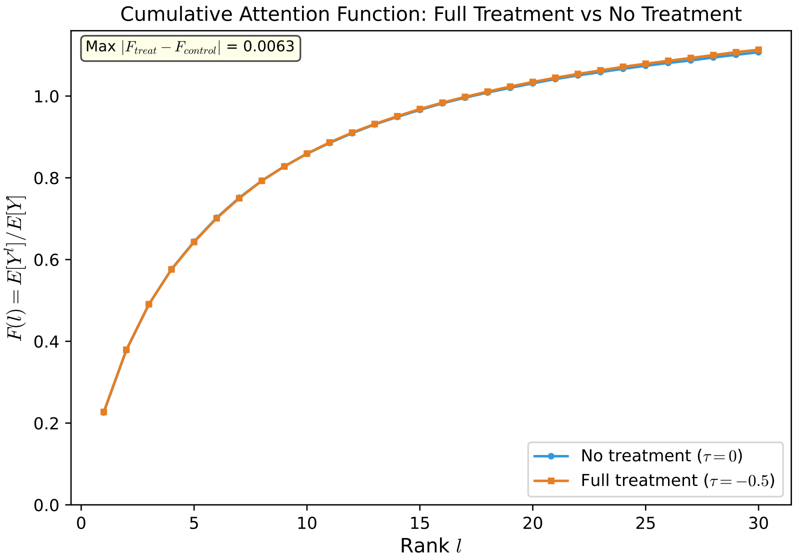


Figure 6: Cumulative attention functions under full-treatment ($\tau = -0.5$) and no-treatment ($\tau = 0$) counterfactuals, each averaged over 30 simulation draws under the original ranking. Both curves are increasing and concave, consistent with Definition 3. The near-complete overlap ($\max |F_{\text{treat}} - F_{\text{control}}| = 0.006$) indicates that treatment scales outcomes without altering the shape of attention, supporting Assumption 1.

order 0.7), fluctuate around zero, and exhibit no systematic pattern. The widening confidence intervals at deeper ranks reflect the cumulative nature of the outcome rather than asymmetric attenuation.

(ii) Placebo outcome balance. Placebo items receive no treatment in either arm and occupy the middle block (priority 2) under TSPR. Any difference in placebo outcomes between arms would therefore indicate asymmetric distortion in the re-ranking mechanism itself. In a dummy experiment with $p = 0.25$, we compare the fraction of queries in which at least one placebo item is booked across arms. The placebo booking rates are 22.5% (Q^A, $n = 9{,}973$) and 23.0% (Q^B, $n = 9{,}996$), with a two-proportion z-test yielding $z = 0.72$ ($p = 0.47$). Figure 8 confirms this balance, showing that rank-specific placebo booking rates are nearly identical across arms throughout the listing.

(iii) Relevance balance by rank. Because items are randomly assigned to treatment groups independently of their baseline relevance, the expected relevance at each display rank should be equal across arms. Figure 9 plots the rank-specific difference in mean pre-treatment relevance, $\mathbb{E}[r \mid Q^B, l] - \mathbb{E}[r \mid Q^A, l]$, averaged over 50 independent simulation draws. The

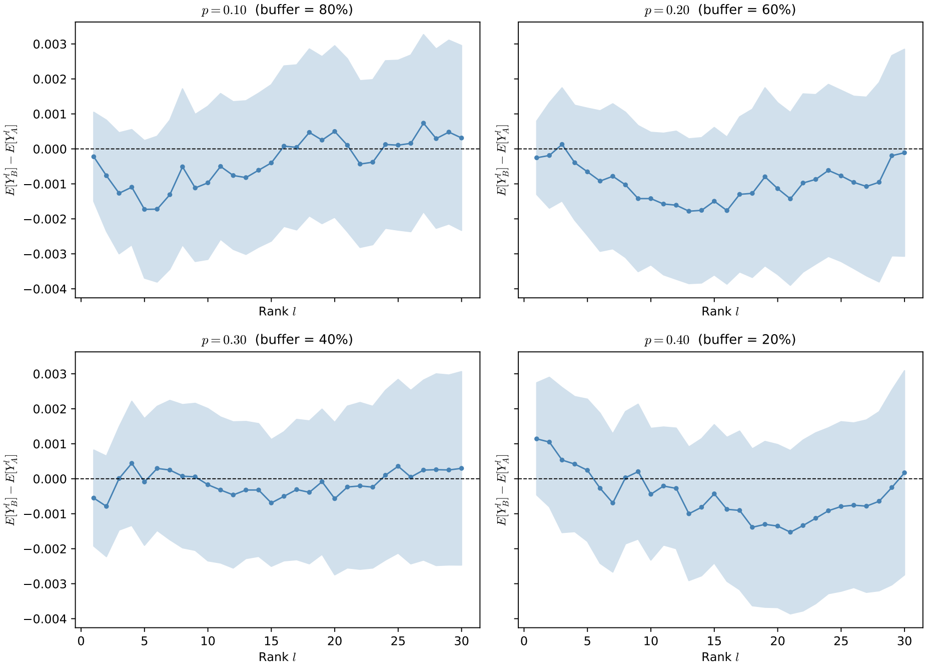


Figure 7: Dummy experiment diagnostic for symmetric distortion. The figure plots $\mathbb{E}[Y^l]_B - \mathbb{E}[Y^l]_A$ by rank l in dummy experiments where treatment labels are assigned but no treatment is applied ($\tau = 0$), averaged over 50 draws. Panels correspond to different treatment assignment probabilities $p \in \{0.10, 0.20, 0.30, 0.40\}$. Shaded regions denote pointwise 95% confidence intervals.

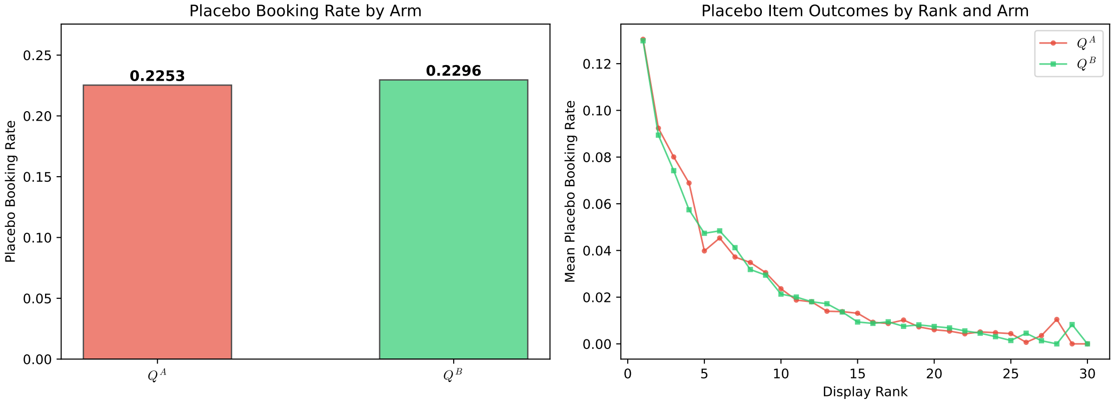


Figure 8: Placebo outcome balance across experimental arms ($p = 0.25$, $\tau = 0$). Left: placebo booking rates by arm. Right: mean placebo booking rate by display rank and arm. Placebo items receive no treatment in either arm; any differences would indicate asymmetric distortion.

differences are tightly centered around zero with no systematic rank-dependent pattern, confirming that the block-swap mechanism does not create quality imbalances between arms.

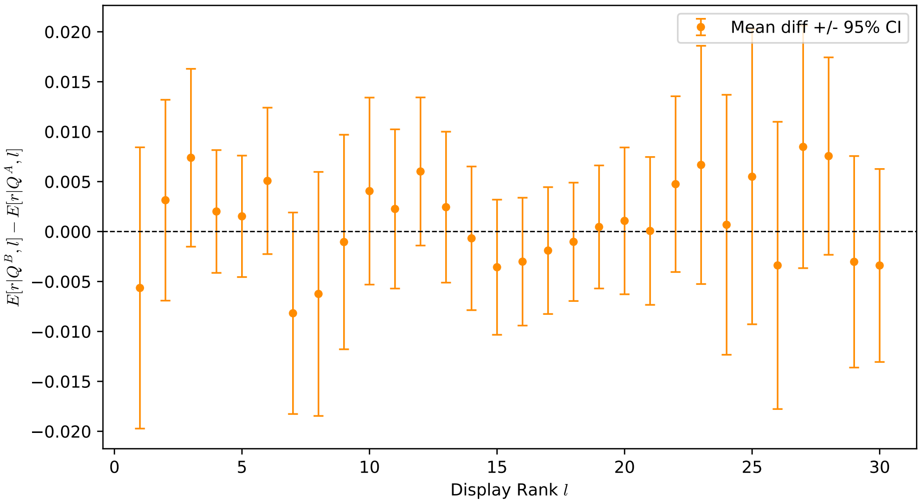


Figure 9: Placebo Balance Check. Rank-specific difference in mean pre-treatment ($\tau = 0$) relevance between arms Q^B and Q^A, in 50 simulations ($p = 0.25$). The horizontal line at zero corresponds to symmetric distortion.

Joint diagnostic: partial outcomes under TSPR versus counterfactuals. As a joint check of both Assumptions 1 and 3, Figure 10 plots mean partial outcomes $\mathbb{E}[Y^l]$ under four scenarios: original ranking with no treatment, original ranking with full treatment, and the two TSPR arms (Q^A and Q^B) with treatment applied at $p = 0.25$. The TSPR arms lie below the original-ranking curves, confirming that re-ranking attenuates outcomes as formalized by the distortion function $d(l; p)$ in Assumption 2.

The gap between the Q^A and Q^B curves reflects the differential treatment exposure that TSPR exploits for identification, not asymmetric distortion: Q^B promotes treated items (which reduce bookings under $\tau < 0$) while Q^A promotes untreated items. Both TSPR curves track the same concave shape as the counterfactuals, consistent with multiplicative rather than shape-altering distortion.

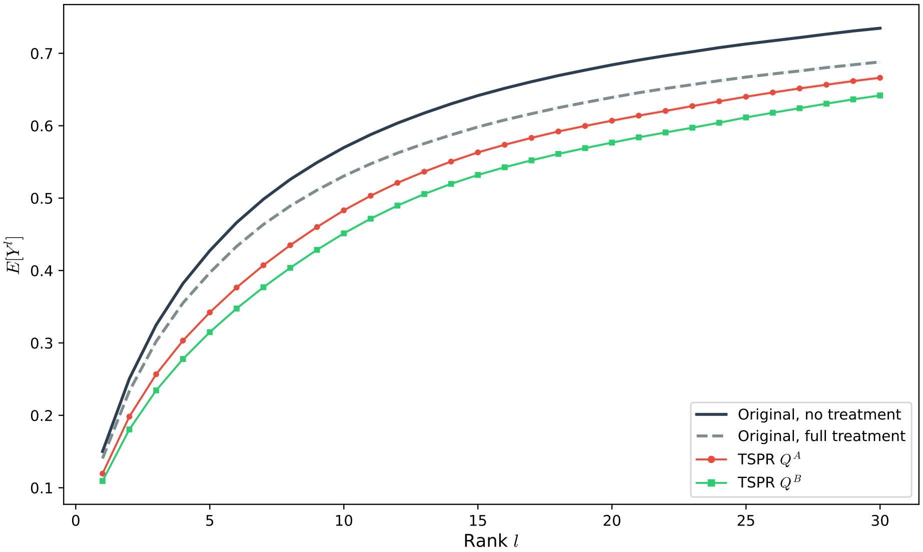


Figure 10: Mean partial outcomes $\mathbb{E}[Y^l]$ by rank under four scenarios, each averaged over 50 simulation draws. The two solid dark lines show the original ranking under no treatment and full treatment. The red and green lines show TSPR arms Q^A and Q^B with treatment applied at $p = 0.25$. The TSPR curves lie below the original-ranking baselines, reflecting re-ranking attenuation. The gap between Q^A and Q^B reflects differential treatment exposure, the identifying variation exploited by TSPR.

A.2 Booking Model Comparison

We compare two specifications of the booking model: a standard multinomial logit (MNL) that conditions only on item utility, and the cascade model defined in Equation (14) that adds a click-order effect. Both models are estimated by maximum likelihood on the same sample of clicked items.

The MNL model is

$$P(\text{book}_{ik} \mid C_i) \;=\; \frac{\exp\big(\gamma_1\, v_{ik} + \gamma_0\big)}{1 + \sum_{k' \in C_i} \exp\big(\gamma_1\, v_{ik'} + \gamma_0\big)},$$

which is nested within the cascade specification by the restriction $\gamma_2 = 0$. Table 6 provides the parameters for a standard Multinomial Logit (MNL) specification without click-order effects. This model was used as a baseline for the likelihood ratio test mentioned in Section 4.2.

Table 6: Standard MNL Booking Parameters (Baseline)

Parameter	Variable	Estimate
γ_1	Latent Utility (v)	0.4472
γ_0	Intercept	0.2376

Table 7 reports the log-likelihoods and the likelihood ratio test.

Table 7: Booking model comparison: MNL vs. Cascade.

Model	Log-likelihood	Parameters
Standard MNL	−49,251.4	2
Cascade (click order)	−49,205.3	3
Likelihood ratio statistic:	$\chi^2 = 92.1$	($p < 0.001$, $df = 1$)

The cascade model achieves a significantly higher log-likelihood with the addition of a single parameter (click order), yielding a likelihood ratio statistic of 92.1 on one degree of freedom ($p < 0.001$). The estimated click-order coefficient is negative ($\hat{\gamma}_2 < 0$), confirming that items clicked earlier are more likely to be booked. We therefore adopt the cascade specification for all main results.

A.3 Coverage Analysis

This appendix provides an assessment of confidence-interval calibration for the two estimators used to estimate the total lift in the TSPR experimental design. For each Monte Carlo run, standard errors are computed via the nonparametric query-level bootstrap with $B = 50$ resamples drawn with replacement. Normal-approximation 95% confidence intervals are constructed as $\hat{\Phi} \pm 1.96\,\widehat{\mathrm{SE}}$.

Decomposing under-coverage. Confidence-interval coverage can fail for two distinct reasons: (i) the bootstrap standard error underestimates the true sampling variability of the estimator (*SE miscalibration*), or (ii) the estimator is biased so that intervals are systematically shifted away from the truth (*bias-induced miss*). We quantify these two channels with the following diagnostics:

- **SE-to-SD ratio** $= \overline{\widehat{\mathrm{SE}}}\ /\ \mathrm{SD}(\hat{\Phi})$: the mean bootstrap SE divided by the empirical standard deviation of the point estimates across Monte Carlo runs. A ratio of 1.0 indicates perfect SE calibration.
- **Bias-to-SD ratio** $= |\mathrm{Bias}|\ /\ \mathrm{SD}(\hat{\Phi})$: the absolute bias normalized by the empirical SD. Values above ≈ 0.5 indicate that bias is a meaningful contributor to under-coverage.

Table 8: Coverage Diagnostics ($\Phi_{\text{true}} = -0.0691$, 500 Monte Carlo runs)

Estimator	**SE/SD**	**|Bias|/SD**	**95% Coverage**	**Primary Driver**
TSPR (Parametric)	0.89	0.82	80.0%	Primarily bias
TSPR (Non-parametric)	0.99	0.25	92.6%	Well calibrated

The parametric TSPR estimator achieves good SE calibration (SE/SD = 0.89). The residual under-coverage (80.0% vs. 95%) is driven almost entirely by the estimator's bias: the parametric form $\tau(l) = 1 + \frac{1}{l}$ does not perfectly match the true shape of the partial treatment-effect function, inducing a systematic shift of ≈ 0.82 empirical standard deviations. This estimator nevertheless achieves the lowest RMSE of all four designs, making it the preferred choice when point-estimation accuracy is the primary objective.

The isotonic regression estimator achieves near-perfect SE calibration (SE/SD = 0.99) and low bias (ratio 0.25), yielding 92.6% coverage, close to the nominal 95% level. This makes it the preferred choice when valid inferential coverage is required, for instance when the goal is to detect whether a treatment effect is statistically significant.

The two TSPR estimators offer complementary strengths: the parametric variant minimizes mean-squared error and is best suited for point estimation, while the non-parametric variant provides well-calibrated confidence intervals for hypothesis testing. Both represent a substantial improvement over conventional designs, which suffer from either severe bias (Bernoulli) or inflated variance with moderate SE miscalibration (cluster randomization).

A.4 Sensitivity to Treatment Probability

We examine the robustness of our estimator to the choice of treatment probability p, which determines the allocation of items across three experimental groups: treated (p), untreated (p), and the placebo buffer ($1 - 2p$). In our main specification, we set $p = 0.25$, yielding equal 25% shares for treated and untreated items with a 50% placebo buffer. We vary p from 0.10 to 0.45, corresponding to buffer sizes ranging from 80% down to 10%.

Figure 11 compares the distribution of lift estimates ($\hat{\Phi}$) for TSPR versus the naive item-side A/B test across eight different levels of p. In every scenario, the TSPR distribution is more tightly centered around the ground truth ($\Phi = -0.0691$). As shown in Figure 12, TSPR consistently outperforms the naive experiment in both bias and variance.

TSPR achieves significantly lower bias than the naive estimator across all tested values of p. The naive estimator exhibits substantial negative bias, with an absolute relative bias ranging from 41.9% to 55.3% of the ground truth. This magnitude reflects the severe interference that occurs when treated and untreated items compete for finite user attention within the same ranking list. In contrast, TSPR dramatically reduces this error, maintaining an absolute relative bias between 11.0% and 22.6%. By decoupling the competition through prioritized ranking, TSPR yields estimates that are consistently closer to the true treatment effect, regardless of the specific allocation.

The variance results similarly favor TSPR, which exhibits lower estimation variance than the naive approach across all values of p. Both estimators show a general downward trend in variance as p increases, as larger values of p increase the sample size of treated and control units, thereby improving the precision of the point estimates.

These results confirm that TSPR is robust to the choice of p and provides reliable, interference-corrected estimates across a wide range of experimental configurations.

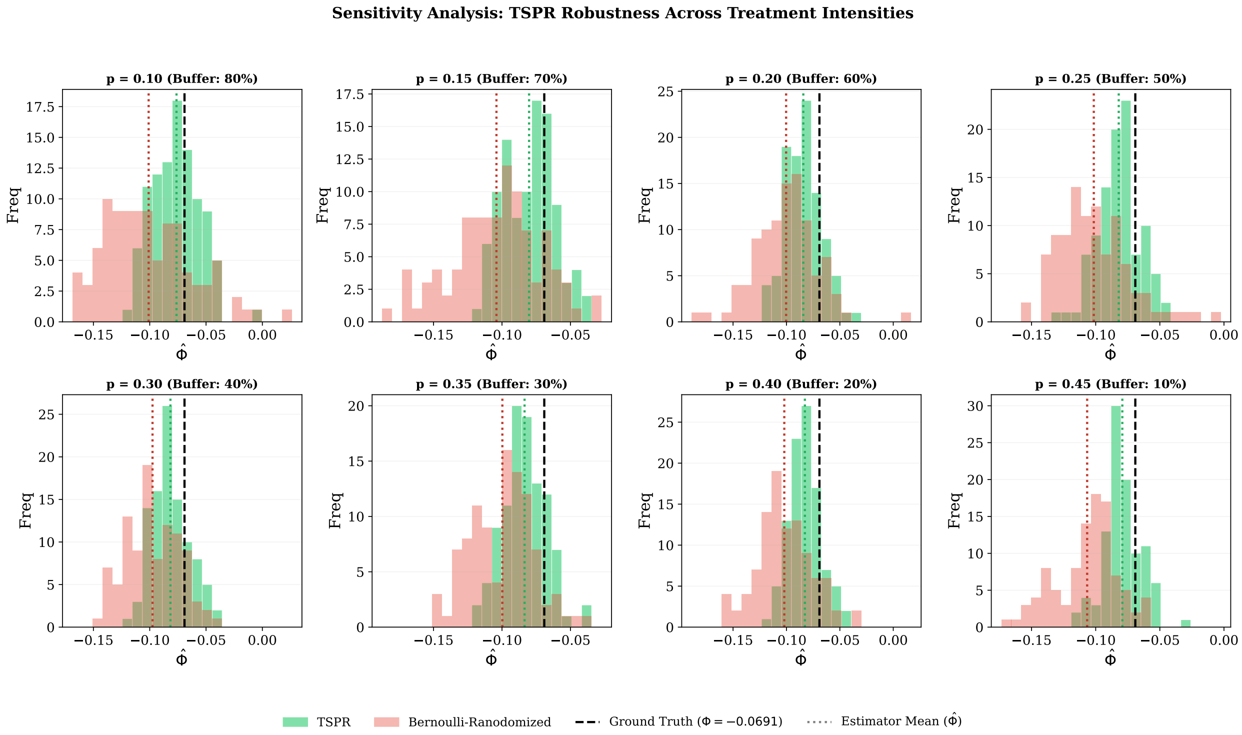
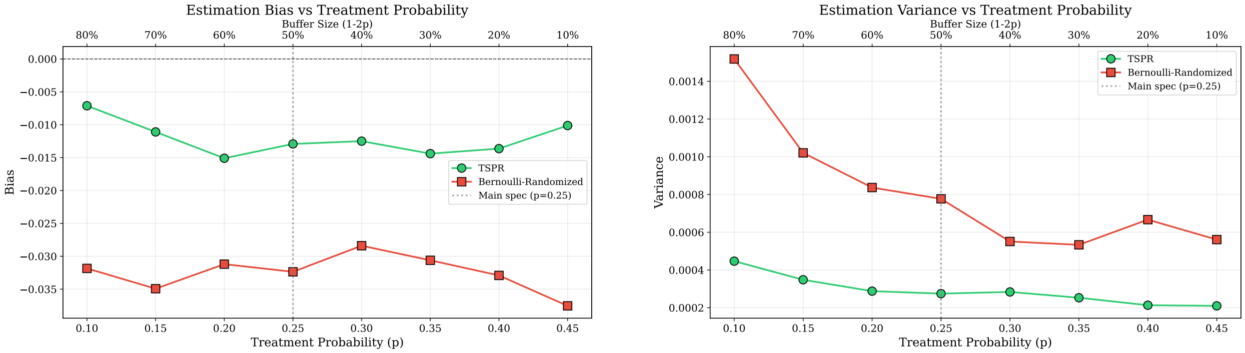


Figure 11: Distribution of Estimates ($\hat{\Phi}$) across different treatment probabilities (p). The dashed black line represents the ground truth value of -0.0691. TSPR (green) consistently exhibits lower spread and better centering on the ground truth compared to the Naive estimator (red) across all buffer sizes.



(a) Estimation Bias

(b) Estimation Variance

Figure 12: Sensitivity to treatment probability p. Panel (a) shows estimation bias and Panel (b) shows estimation variance for TSPR and the Bernoulli-Randomized estimator across different values of p. The buffer size (1-2p) is shown on the secondary axis on top. The vertical dotted line indicates our main specification ($p = 0.25$). TSPR consistently achieves lower bias and variance than the naive estimator across all tested configurations.

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.

Barlow, R. E. (1972), 'Statistical inference under order restrictions: The theory and application of isotonic regression', *(No Title)* .

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.

Brown Jr, B. W. (1980), 'The crossover experiment for clinical trials', *Biometrics* pp. 69–79.

Çakır, M., Liaukonyte, J. and Richards, T. J. (2025), 'Price gouging, greedflation, and price fairness perceptions', *Cornell SC Johnson College of Business Research Paper* .

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.

Consumer Reports (2025), 'American experiences survey: A nationally representative multimode survey: September 2025 omnibus results'. Prepared by Consumer Reports Survey Research Department.

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.

European Commission (2025), 'Application of article 102 tfeu', European Commission, Competition Policy (DG COMP).
URL: *https://competition-policy.ec.europa.eu/antitrust-and-cartels/legislation/application-article-102-tfeu$_e$n*

European Union (2008), 'Consolidated version of the treaty on the functioning of the european union, article 102', EUR-Lex. CELEX:12008E102. Article 102(c) prohibits applying dissimilar conditions to equivalent transactions by dominant undertakings.
URL: *https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:12008E102*

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* .

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.

Instacart (2025), 'Ending item price tests on instacart', Instacart Company Updates. Accessed 2025-12-26.
URL: *https://www.instacart.com/company/updates/ending-item-price-tests-on-instacart*

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.

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.

Kravitz, D. (2025), 'Instacart's ai pricing may be inflating your grocery bill', Consumer Reports. Updated Dec. 22, 2025. Accessed Dec. 30, 2025.

URL: *https://www.consumerreports.org/money/questionable-business-practices/instacart-ai-pricing-experiment-inflating-grocery-bills-a1142182490/*

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.

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.

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.

WIRED Staff (2000), 'Amazon makes price amends', *WIRED* . Accessed 2025-12-26.
URL: *https://www.wired.com/2000/09/amazon-makes-price-amends/*

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* .