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

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

December 31, 2025

Click for the latest version

Abstract

Online marketplaces frequently run experiments on item-side changes such as price updates. A key challenge is interference: items compete for users' limited attention and demand within ranked lists, so treating some items can change outcomes for untreated items, biasing standard item-level A/B tests. Moreover, platforms typically require price coherency (all users see the same price for each item) and full catalog access, ruling out many existing experimental designs. We propose Two-Sided Prioritized Ranking (TSPR), which exploits position bias in ranked search results to create variation in treatment exposure. TSPR randomizes users into two groups and reorders ranked lists so that treated items appear at the top for one group and untreated items for the other, while a placebo set balances item quality across groups. All users see the same items at consistent prices, but differ in exposure to treatment through rank position. Our estimator identifies the proportional effect of universal treatment relative to universal control. In simulations calibrated to Expedia hotel search data, TSPR substantially reduces interference bias relative to item-level A/B tests and outperforms cluster randomization while preserving a coherent user experience.

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 6 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 aver-

age 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

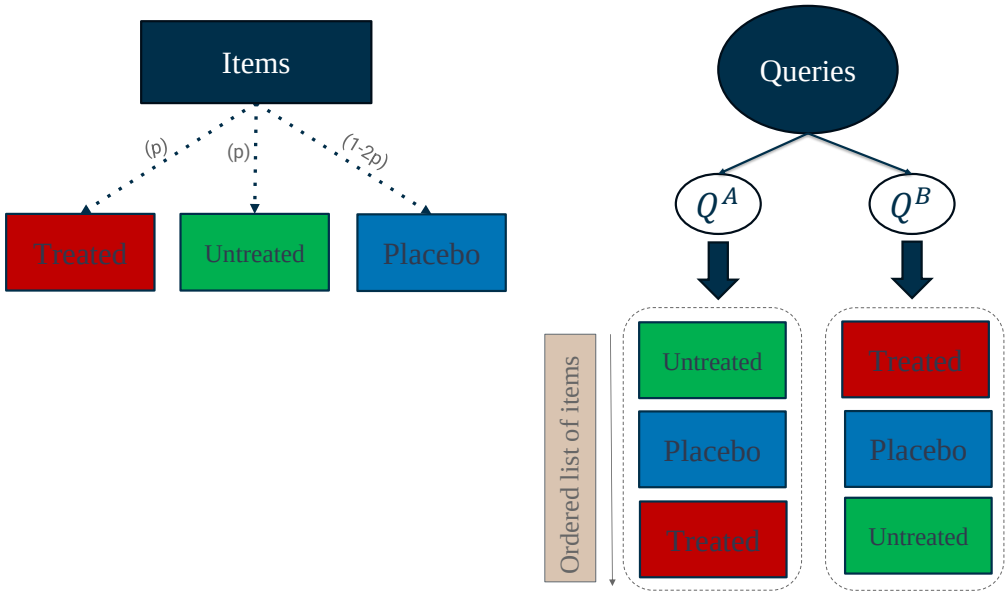
[4]Extending the design to settings with binding capacity constraints or other forms of user-side interference across queries or users is left to future work.

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

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

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 $l \le n_b$ treated items at the top:

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

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: substitution among items tends to push $\tau(1) < 1$ because demand can shift from treated top items to untreated items below, whereas complementary-purchase or engagement-spillover effects can push $\tau(1) > 1$. In our main application (an Expedia-like marketplace), items are substitutes in expectation: on average, treating top-ranked items diverts demand toward untreated items lower in the list. This substitution channel suggests $\tau(1) < 1$, with $\tau(l)$ increasing 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 $l \le n_u + n_p$ 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:

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

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.

In the remainder of this section, we work in this negligible-contamination regime and set the nuisance term $\nu(l)$ to zero. This allows us to summarize depth-dependent interference using a single function $\tau(l)$ and to write the partial lift as $\phi(l) = \Phi\,\tau(l)$. We then impose a parsimonious parametrization for $\tau(l)$ and estimate (Φ, γ, k) by matching the model-implied partial lifts to their empirical analogs.

3.3 Parametrization

We impose a parsimonious parametric form on the depth function $\tau(l)$ that matches the qualitative restrictions above: $\tau(l)$ is smooth in l and satisfies $\tau(l) \to 1$ as $l \to \infty$. This normalization ensures that the global lift Φ corresponds to the full-treatment proportional effect, and that $\tau(l)$ captures how partial treatment at depth l differs from full treatment.

Using the partial-lift definition in Equation (7), we model

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

where (γ, k) governs the sign and persistence of interference across items. Under Assumptions 1 (treatment–attention separability) and 3 (symmetric distortion), $\tau(l;\gamma,k)$ separates interference from mechanical rank effects and can be interpreted as *net spillovers* beyond baseline rank popularity and position effects.

Interpretation of parameters.

- Φ is the *full-treatment proportional effect*. Because $\tau(l;\gamma,k) \to 1$ as $l \to \infty$, we have $\phi(l) \to \Phi$ as l increases.
- γ controls the *direction* and *magnitude* of interference at small depths through deviations of $\tau(l;\gamma,k)$ from 1. If $\gamma < 0$, then $\tau(l;\gamma,k) < 1$ for small l, consistent with

substitution-dominant lists: treating only the top-ranked items shifts some demand toward untreated items below, so the partial-treatment effect is smaller in magnitude than the full-treatment lift Φ. If $\gamma > 0$, then $\tau(l; \gamma, k) > 1$ for small l, consistent with *complementary-purchase* or *engagement-spillover* effects: treating salient top-ranked items can reduce demand for other items in the query (e.g., complementary basket additions) or increase abandonment, so the partial-treatment effect can be larger in magnitude than Φ.

- When $\gamma = 0$, there is *no interference* and $\tau(l; \gamma, k) \equiv 1$ for all l, implying $\phi(l) = \Phi$ at every depth.
- The parameter $k > 0$ governs the *rate* at which interference dissipates with depth. Larger k implies faster convergence of $\tau(l; \gamma, k)$ to 1, so $\phi(l) \approx \Phi$ even at small l.

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, the estimation compares queries from group B conditional on $n_b = l$ to queries from group A conditional on $n_u = l$, and the partial outcome is calculated up to position l in both cases.

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.

3.4 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 chosen 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.

4 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.

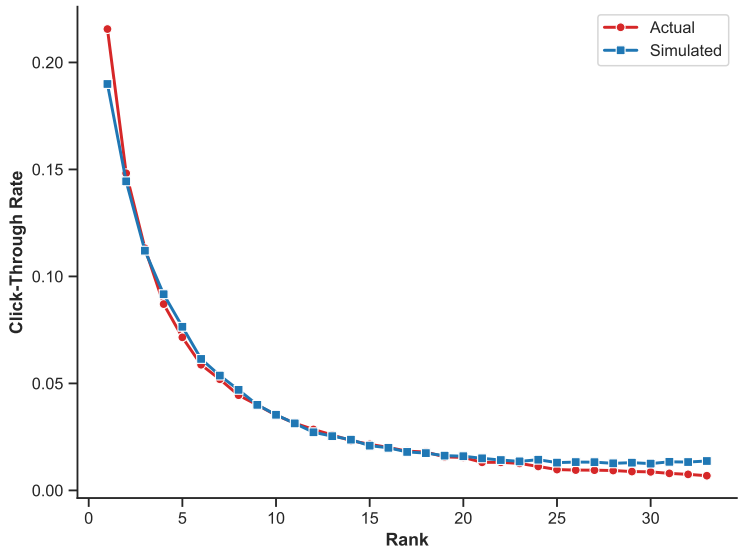
To model user interactions, we assume that each item displayed to a consumer has a net (hidden) utility, denoted as v. The relevance score r, which represents the recommender system's match score between a consumer's query and an item, is modeled as $r = v + \epsilon$ with ϵ following a normal distribution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

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.

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.

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 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

$$P(\text{click}_{ij}) = \text{logit}^{-1}\big(\beta_1 p_{ij} + \beta_2 p_{ij}^2 + \beta_3\,\text{prevclicks}_i + \beta_4\,\text{has\_clicked}_i + \beta_5 v_{ij} + \beta_0\big)\,. \quad (12)$$

where v_{ij} is the latent utility of item j for user i, has_clicked_i indicates whether the user has clicked any earlier item in the list, and prevclicks_i is the number of clicks the user has made on earlier positions within 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 coefficients β_3 and β_4 capture stopping behavior. As the user accumulates earlier clicks (prevclicks_i) or has already clicked at least one item (has_clicked_i), the probability of clicking on the current item decreases. Together, these terms generate a continuation probability similar to the stopping parameter in position–based models with continuation.
- The coefficient β_5 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.

5 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.

Figure 3 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

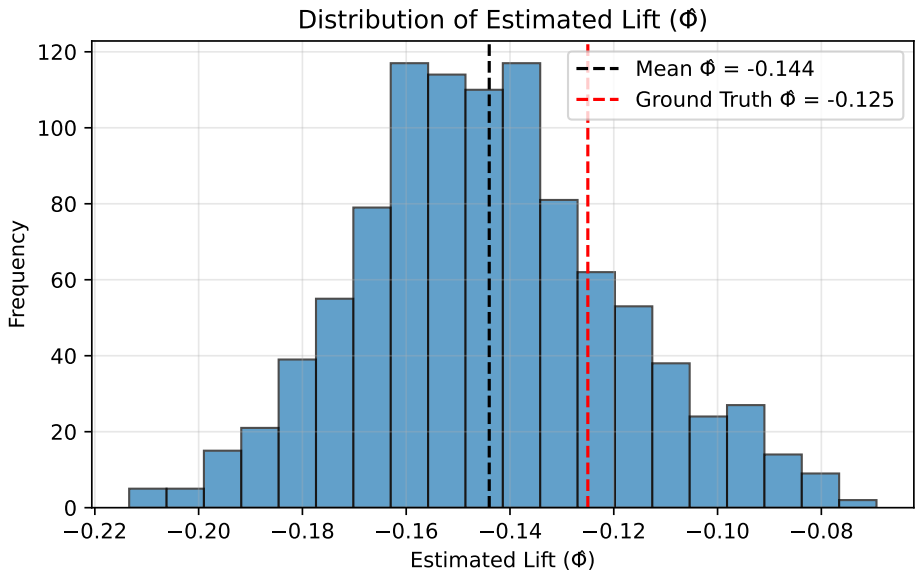


Figure 3: Distribution of estimated lift Φ from TSPR experiments in the semi-synthetic setting. The histogram is based on 1,000 simulation runs based on sampled queries and items, where choice sets are generated synthetically and clicks and bookings follow the model in Subsection 4.1. The dashed vertical line marks the mean estimated lift and the red line marks the ground truth value of Φ.

are generated from the specification in Subsection 4.1.

Figure 4 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, in both a fully synthetic world and a world based on real search impressions.

5.1 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 5 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

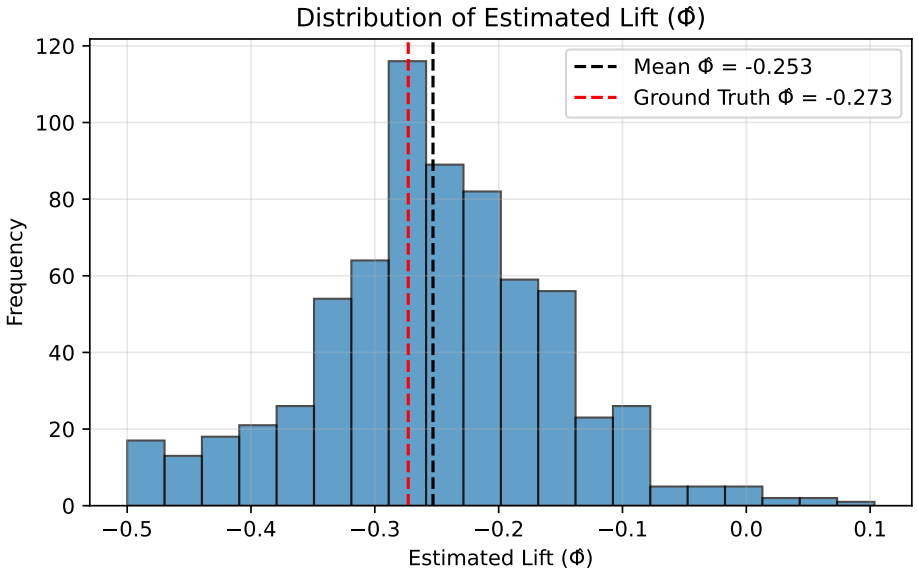


Figure 4: Distribution of estimated lift Φ from TSPR experiments on real search impression data. The histogram is based on 1,000 simulation runs on the `df_single` sample, where the observed impression graph is held fixed and clicks and bookings are simulated according to the model in Subsection 4.1. The dashed vertical line marks the mean estimated lift and the red line marks the ground truth value of Φ.

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 Lift in total outcome, we extend 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

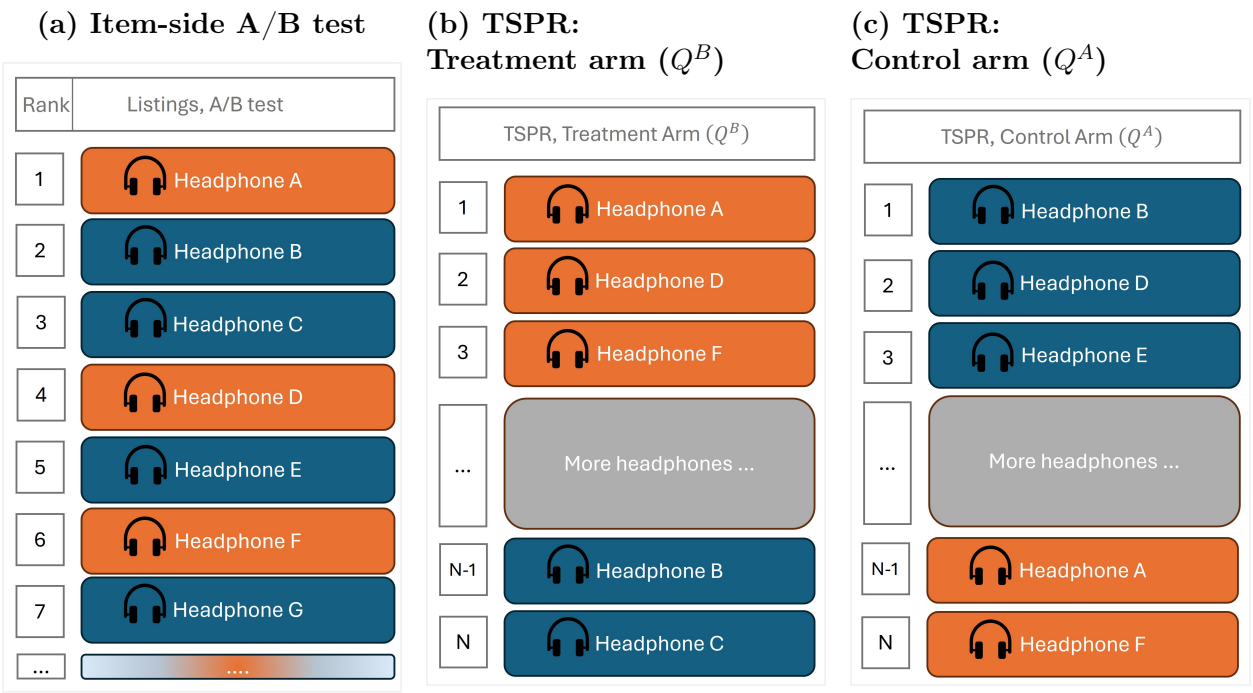


Figure 5: Illustrative comparison of an item-side A/B test and the Two-Sided Prioritized Ranking (TSPR) design. Orange items indicate treated items, blue items indicate untreated items, and gray items indicate placebo items. Panel (a) shows a Bernoulli-randomized item-side experiment in which treated 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.

source of randomization is the item assignment, which affects the distribution of treated and control items that appear in each query.

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

$$\Phi_{IS} = \frac{\mu_B}{\mu_A} - 1.$$

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 Bernoulli-randomized experiment as a baseline that uses only item-level randomization and ignores query-level randomization.

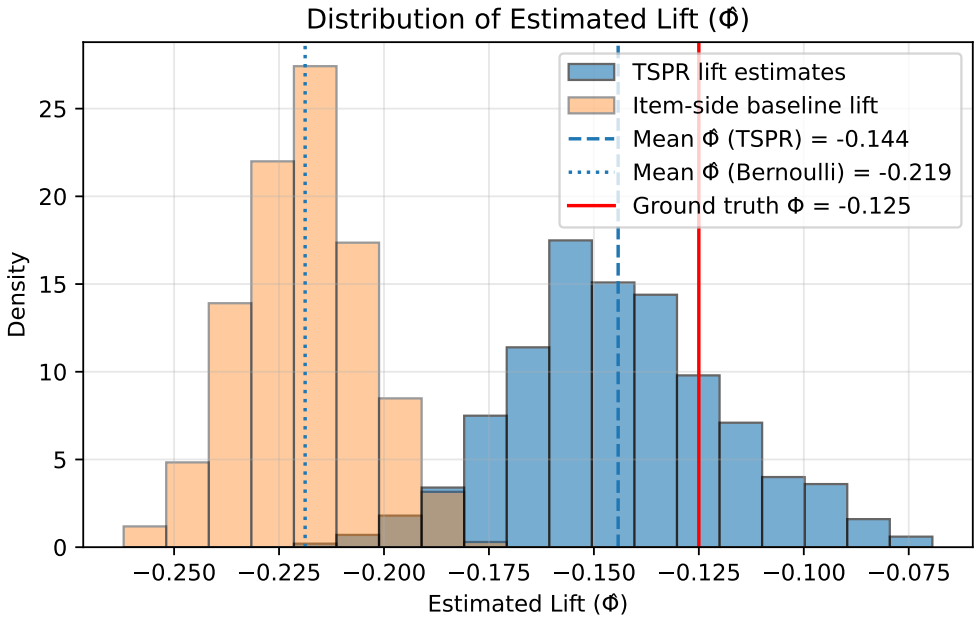


Figure 6: Distribution of lift estimates from 1,000 simulations under the TSPR design and the Bernoulli-randomized A/B test. The vertical lines show the mean estimate for each method and the ground truth lift, computed as the proportional change in total outcomes between counterfactual all-treated and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

The item-side baseline 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.

5.2 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 4. 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, while all properties in a control destination remain untreated. This design respects geographic segmentation and, by construction, eliminates spillovers between treated and control clusters.

Restricting attention to properties linked to a single destination yields an unusually clean clustering structure that is well aligned with the true pattern of spillovers. In practice, properties often appear in multiple destinations, and geographic or demand-based clusters are substantially more overlapping and difficult to delineate. As a result, this implementation represents a favorable setting for cluster randomization and provides an upper bound on its performance in more realistic marketplace environments.

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

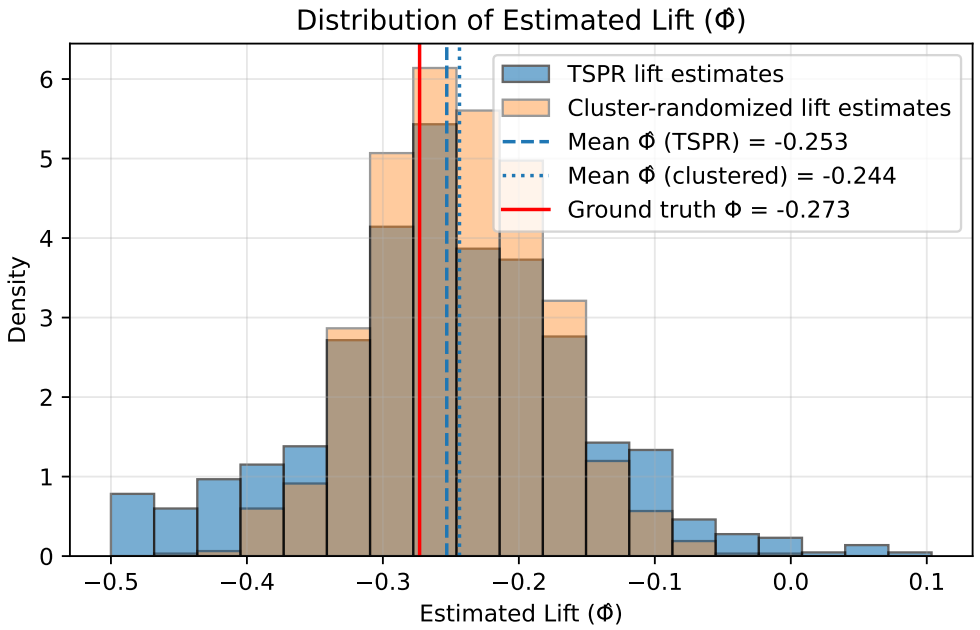


Figure 7: Distribution of estimated lift Φ under TSPR and cluster-randomized designs. Each histogram is based on 1,000 simulation runs using the same impression data and outcome model. Vertical lines mark the mean estimated lift for each design and the true value of Φ.

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-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 reflects a modest variance cost in exchange for a substantially smaller bias. Importantly, the cluster-randomized design performed here operates in a favorable setting with

access to clean, well-aligned geographic clusters and therefore represents an upper bound on the performance of cluster randomization in practice. Even in this setting, TSPR delivers estimates that are closer to the ground truth while preserving coherent user experience.

6 Conclusion

This paper introduces Two-Sided Prioritized Ranking (TSPR), an experimental design for item-side interventions in online marketplaces that maintains price coherency 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 item-level A/B tests and outperforms cluster randomization even when clusters are cleanly defined.

Practical implications. Our findings have direct implications for platform experimentation. First, practitioners can credibly estimate treatment effects for price changes and other item-level interventions without showing different prices to different users or partitioning the item catalog. TSPR requires only the ability to modify ranking priorities, which is straightforward to implement in existing recommender systems. Second, the design leverages position bias (typically viewed as a nuisance) as an instrument for identification. This suggests that platforms with strong rank-dependent attention can exploit this feature for causal inference, turning a limitation into an asset. Third, while TSPR induces higher variance than some biased alternatives, the bias-variance tradeoff is favorable: estimates remain centered near the truth rather than systematically overestimating or underestimating effects.

Limitations and scope. TSPR is best suited to settings with three features: strong position bias, slack supply (so one user's actions do not constrain availability for others), and short experiment horizons that limit dynamic feedback effects. The design may perform poorly when users exhibit weak substitution between items, or when ranking perturbations substantially degrade the user experience. Our identifying assumptions (treatment-attention separability and symmetric re-ranking distortion) are plausible in many marketplace settings but may fail when treatment fundamentally changes user engagement or when treatment

assignment probabilities are large. Practitioners should validate these assumptions through pre-tests or sensitivity analyses.

Future directions. Several extensions warrant investigation. Adapting TSPR to environments with binding capacity constraints or user-side interference across queries would broaden its applicability to supply-constrained platforms. Extending the framework to continuous treatments or multiple treatment arms would address more complex experimental needs. More broadly, this paper contributes to the growing literature on causal inference under interference by showing that experimental designs can credibly recover treatment effects even when standard randomization schemes fail, provided the design exploits features of the environment (in our case, ranked attention) to create identifying variation. We view the coherency constraints that motivate TSPR not as limitations but as realistic requirements that should guide methodological development in platform experimentation, bridging the gap between theoretical advances and industry practice.

Appendix: Empirical Support for Assumptions

Empirical support for the attention function. Figure 8 provides empirical support for Definition 3 by illustrating the shape of the cumulative attention function implied by the data. Across both full-treatment ($p = 1$) and no-treatment ($p = 0$) counterfactuals, the expected partial outcome $\mathbb{E}[Y^l]$ increases with the rank length l but at a decreasing rate. Normalizing by total expected outcomes yields a cumulative attention function $F(l) = \mathbb{E}[Y^l]/\mathbb{E}[Y]$ that is increasing and concave in l, consistent with strong position bias and diminishing marginal attention at lower-ranked positions.

Importantly, the attention profiles under full treatment and no treatment are nearly identical in shape. This indicates that treatment affects the level of outcomes without altering the allocation of attention across ranks, providing empirical support for Assumption 1.

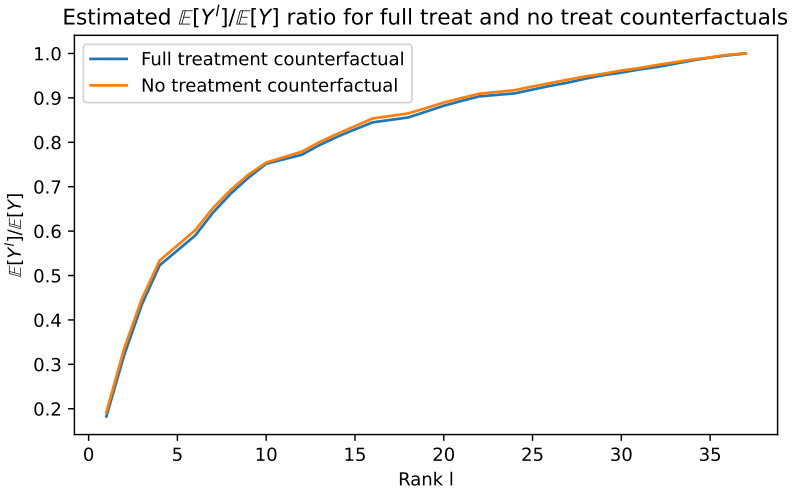


Figure 8: Estimated cumulative attention functions under full-treatment and no-treatment counterfactuals. The figure plots $\mathbb{E}[Y^l]/\mathbb{E}[Y]$ as a function of rank l, where Y^l denotes the cumulative partial outcome up to rank l. In both counterfactuals, cumulative outcomes increase with l at a decreasing rate, yielding a concave and increasing attention profile. The near overlap of the two curves indicates that treatment scales outcomes without altering the shape of attention.

Empirical support for symmetric distortion. Assumption 3 is plausible in the experimental environment we study for two structural reasons. First, the experiment is conducted over a short horizon, which limits dynamic feedback effects such as popularity accumula-

tion, learning, or relevance drift that could differentially affect attention across experimental arms. Second, the marketplace operates in a slack-supply regime, so increased exposure or engagement in one arm does not constrain item availability or ranking opportunities in the other.

Under these conditions, any attenuation in user attention induced by the re-ranking mechanism is mechanical and driven by list position rather than by endogenous cross-arm interactions. As a result, the expected distortion at a given rank depends on the treatment assignment probability p but not on the experimental arm, motivating Assumption 3.

Assumption 3 requires that, conditional on the treatment assignment probability p, the ranking mechanism induces the same expected attention attenuation at each depth l in both experimental arms. To assess this implication, we conduct a simulation-based diagnostic that compares rank-specific average item quality across arms.

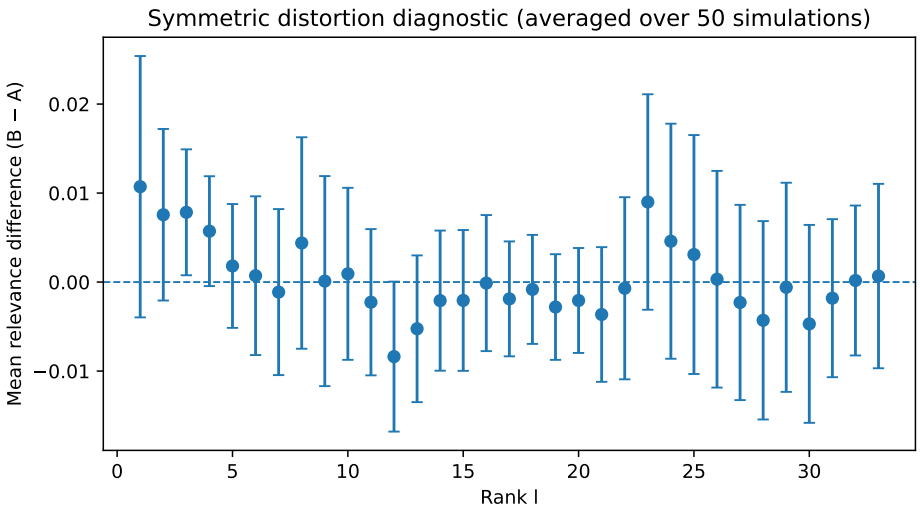


Figure 9: Rank-specific difference in mean pre-treatment relevance between arms Q^B and Q^A, averaged over 50 independent simulation draws. Each point shows the across-simulation mean of $\mathbb{E}[r \mid Q^B, l] - \mathbb{E}[r \mid Q^A, l]$ at rank l. Error bars denote ± 1.96 standard errors computed across simulations. The dashed horizontal line at zero corresponds to symmetric distortion.

For each simulation draw, we compute the difference in mean pre-treatment relevance at rank l between arm B and arm A, pooling across all queries within each arm. We then average these rank-specific differences across 50 independent simulation draws and compute standard errors using the across-simulation variation. Figure 9 plots the resulting mean differences with pointwise 95% confidence intervals.

Across all ranks, the estimated differences are tightly centered around zero and exhibit no systematic rank-dependent pattern. The maximum standardized deviation across ranks is $|z| = 2.17$, which is consistent with chance variation under multiple testing. Overall, this diagnostic provides empirical support for Assumption 3 by showing no evidence of asymmetric distortion across experimental arms.

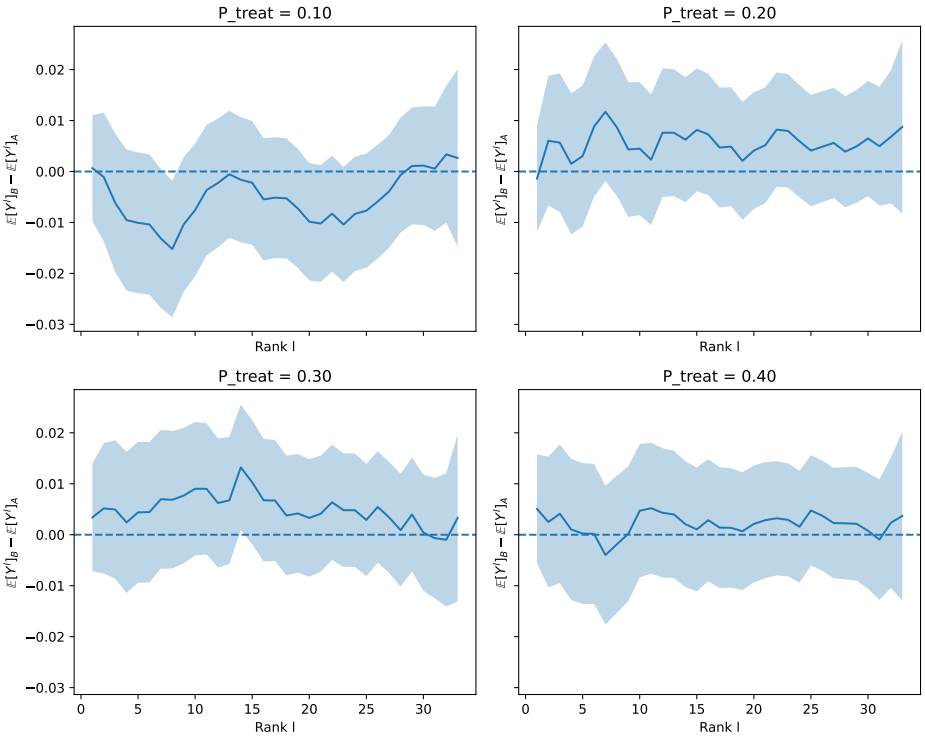


Figure 10: A/A diagnostic for symmetric distortion across treatment probabilities. The figure plots the difference in mean cumulative partial outcomes, $\mathbb{E}[Y^l]_B - \mathbb{E}[Y^l]_A$, by rank l in an A/A experiment where treatment labels are assigned but no treatment is applied ($\tau = 0$). Panels correspond to different treatment assignment probabilities $p \in \{0.10, 0.20, 0.30, 0.40\}$. Shaded regions denote pointwise 95% confidence intervals. The dashed horizontal line at zero corresponds to symmetric distortion.

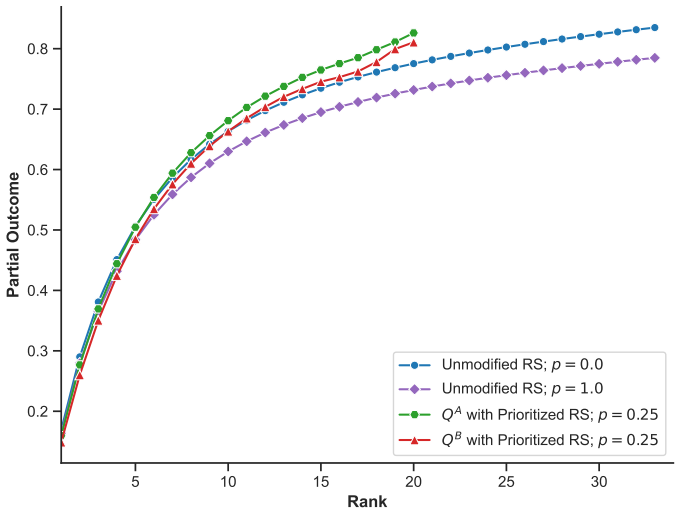
As an additional robustness check, we conduct an A/A stress test in which treatment labels are randomly assigned but no treatment is applied. In this setting, any systematic difference in outcomes across arms must arise from asymmetric distortion rather than treatment effects. Figure 10 plots differences in mean cumulative partial outcomes across arms for several treatment assignment probabilities p.

We focus on intermediate values of p for which the TSPR design is well defined: placebo blocks exist but do not dominate the ranked list. Across all values of p, the estimated

differences remain small, fluctuate around zero, and exhibit no systematic rank-dependent pattern. The widening confidence intervals at deeper ranks reflect the cumulative nature of the outcome and diminishing effective sample size rather than asymmetric attenuation. Overall, this stress test provides corroborating evidence for Assumption 3 and confirms that the TSPR mechanism does not induce economically meaningful asymmetric distortion across experimental arms.

As a further sanity check, we compare the distribution of baseline relevance scores for placebo items across experimental arms. Because placebo items are unaffected by treatment, any differences in their baseline quality would reflect compositional imbalance rather than distortion. We find that the distributions of query-level mean placebo relevance are nearly identical across arms, and a Kolmogorov–Smirnov test fails to reject equality of distributions, supporting the comparability of the item pools faced by users in both arms.

Figure 11: 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 $p = 0.25$.

Additionally, Figure 11 shows partial outcomes under the modified recommender system close to the baseline case. It 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.

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.

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* .

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* .