Algorithmic Design for Shipping Efficiency: Lessons from Field Experiments

Malika Korganbekova
The University of Chicago
Booth School of Business

Alibek Korganbekov
Boston University
Questrom School of Business

Aliya Korganbekova
Harvard University
Harvard Business School

Vinny DeGenova
Wayfair LLC

Yasaman Khazaeni
Wayfair LLC

Cole Zuber
Wayfair LLC

July 6, 2025

Abstract

Fast and free shipping has become a standard expectation in online retail, but it imposes substantial logistical and environmental costs. This paper evaluates whether platform-side interventions can reduce shipping costs and emissions without harming consumer outcomes. Using two large-scale field experiments on a major e-commerce platform, we first show that removing fast-shipping promises does not improve shipping efficiency. On the contrary, given the current infrastructure and demand for fast shipping, reducing these promises increases shipping distances and costs and leads to less consolidation. We then propose and test an alternative approach: algorithmic re-ranking of product assortments to (i) show well-liked products that are also (ii) cheaper-to-ship. This randomized intervention reduces shipping distances, increases fulfillment from centralized warehouses, and lowers shipping costs by up to 7.3 percent, without degrading conversion rates or affecting product returns. Our findings demonstrate that low-cost, demand-side levers—such as modifying product rankings—can achieve operational and environmental gains, offering a scalable alternative to traditional supply-chain optimization or monetary consumer-incentive approaches.

Keywords: fast shipping, shipment consolidation, platform algorithms, ranking nudges, logistics costs, environmental externalities

1. Introduction

The normalization of fast and free shipping in retail has reshaped consumer expectations and introduced significant operational and policy challenges. Same- and next-day delivery, once exceptional,

are now standard for many shoppers.[1] To meet this demand, platforms have invested heavily in fulfillment infrastructure and absorbed substantial shipping costs - large online retailers Amazon and Wayfair reported $89.5 billion and $2.2 billion in 2023 shipping costs (14% of their annual reported revenue),[2] respectively. According to MIT Review, last-mile delivery expenses account for up to 53% of total shipping costs.[3] Yet faster delivery may reduce opportunities for consolidation, raise per-order costs, and intensify last-mile inefficiencies. As deliveries become more fragmented and reliant on large vehicles, the resulting traffic externalities—including road wear & tear and urban congestion—have attracted growing regulatory attention. Several U.S. states, including Colorado and Minnesota, now impose retail delivery fees ranging from $0.29 to $0.50 per transaction to recover infrastructure costs and incentivize consolidation.[4] The logistical efficiency and environmental footprint of shipping costs are tightly linked,[5] which makes it even more important to optimize shipping efficiency. To reduce shipping costs, firms try (i) to optimize their supply chain by building micro-fulfillment centers to reduce logistics costs and use Reinforcement Learning and other Machine Learning algorithms to optimize re-routing (e.g., Andaz et al., 2023, Madeka et al., 2022). Moreover, some firms introduced monetary and behavioral nudges to consumers to encourage delayed delivery—such as Amazon's "Low carbon option" that encourages customers to select slower shipping. However, the evidence on the effectiveness of these type of measures remains limited. This paper proposes and tests platform-side algorithmic interventions, including adjustments to product rankings or shipping promises, to investigate whether they can reduce fulfillment costs and emissions without degrading consumer outcomes. If successful, these algorithmic interventions could be cost-effective and scalable alternatives to the traditional supply-chain and monetary nudging interventions.

This paper has two goals. First, we investigate whether reducing the number of fast shipping promises improves last-mile shipping efficiency, i.e., shipping costs and shipment consolidation. Executives from UPS claimed that the fast delivery promises creates challenges to shipping efficiency, and that they would be more efficient per mile had they had more time.[6,7] If successful, such efficiency could also lower environmental externalities. Using experimental variation, we find that removing fast shipping promises does not yield significant cost savings or emissions reductions. Over time, platforms have built extensive centralized distribution networks to support fast shipping, and sellers have increasingly relied on these networks rather than expanding their own warehouse capacity. As a result, the existing logistics structure

[1] 2020 Flexe Omnichannel Consumer Survey shows that 90% of consumers see 2- to 3-day shipping as the baseline delivery promise.

[2] See Amazon's annual financial report and Wayfair's annual financial report.

[3] See this article for additional details.

[4] See Colorado website and Minnesota website for details on the retail delivery fee.

[5] See MIT study for more information.

[6] See this CNN article for more details.

[7] Specifically, the director of the UPS global sustainability said: 'I don't think the average consumer understands the environmental impact of having something tomorrow vs. two days from now. The more time you give me, the more efficient I can be.'

constrains the platform's ability to achieve shipping cost or consolidation gains through reductions in fast shipping offerings.

Second, we propose and test whether algorithmic interventions can achieve consolidation and cost reduction objectives. Platforms have traditionally relied on operational improvements or behavioral incentives to influence shipping outcomes. In contrast, we evaluate a strategy based on *algorithmic nudges*. Specifically, we consider augmenting the current product ranking algorithm by (i) replacing expensive-to-ship items with cheaper-to-ship substitutes on product pages without altering the original ranking rule, and (ii) re-ranking products so that cheaper-to-ship items among well-liked products are prioritized. We find that simple replacement without re-ranking has negligible effect, while algorithmic re-ranking significantly improves shipment consolidation and reduces shipping costs. The results highlight the potential of algorithmic demand-side interventions to complement existing logistics infrastructure in achieving operational and environmental objectives. Next, we describe the setting and findings in more detail.

In the first part of the paper, we examine whether the removal of fast-shipping promises from online retail platforms can reduce shipping costs and facilitate greater shipment consolidation. Ideally, to address this question, one would compare two distinct counterfactual environments: one in which fast shipping is entirely unavailable, with no supporting infrastructure, and another in which fast shipping is fully implemented across the platform. Such a comparison would allow for a direct evaluation of consumer behavior, seller strategies, and platform outcomes across these two regimes. However, conducting an experiment of this scale is not feasible in practice. As a practical alternative, we partnered with Wayfair, a large US-based online retailer of home goods and furniture, and use an experimental intervention in which a subset of fast-shipping badges was randomly removed from ranking pages. In the control condition, consumers observed all fast-shipping badges as usual. In contrast, for treated consumers, the platform randomly removed between 25 and 75% of fast-delivery badges, delayed the displayed "get-it-by" dates for affected products, and eliminated the search ranking boost previously associated with fast-shipping items. This randomized experiment ran for 1.5 months and spanned the entire website traffic.

During the experiment, consumers assigned to the treatment group—who observed a 25–75% reduction in fast-shipping badges and slight adjustments in product rankings—became 2.7% less likely to purchase products with fast-shipping options. As a result, demand shifted toward products stored in warehouses located farther from the consumers. This shift increased average shipping distances by approximately 1.2–2.2%, equivalent to 6–35 additional last-mile delivery miles per order or 279 miles for mid- and last-mile delivery per purchasing customer. Moreover, this led to the increase in per-order shipping costs for the platform by 3.2%. Given that the platform's annual shipping costs were $2 billion, that would account for approximately additional $64 million in shipping costs.

Using FedEx and UPS tracking data, we further document that the platform's ability to consolidate shipments deteriorated. Shipment consolidation is measured by the number of consumer orders departing a warehouse within a given dispatch window. The consolidation rates fell by 2.4%, and the number of outbound trucks increased by 0.5%. In simpler terms, the platform needed 1 additional truck on average to fulfill the same number of orders per day.

Notably, these effects are either consistent across the intensity of treatment (i.e., 25-75% of removal) or grow in magnitude with the intensity of treatment (e.g., shipping distances exhibit an increasing pattern with the share of the fast badges removed).

Moreover, we find that consumer satisfaction declined following treatment: product returns increased by 3.1%. This effect is primarily driven by returns of non-fast-shipped products, suggesting that consumers who substituted away from fast-shipped goods were less satisfied with their purchases. The increase in returns further contributes to shipping inefficiencies, as additional return shipments and replacement orders increase the total distance traveled by packages by an additional 1.8-2.7%.

These estimates represent a lower bound on the total cost impact, as they do not account for the large fixed investments made by the platform to support its fast-shipping infrastructure.[8]

Thus, the experimental evidence indicates that simply removing fast-delivery promises, absent broader adjustments to platform design and infrastructure, is unlikely to generate meaningful improvements in shipping efficiency or cost savings.

As described above, platforms typically seek to optimize shipping costs through two primary channels: (i) operational improvements, such as investing in fulfillment infrastructure or route optimization, and (ii) behavioral or monetary nudges, such as incentivizing consumers to select slower shipping options or adding climate-friendly badges. In contrast, we propose a different approach: using algorithmic nudges embedded in the platform's ranking algorithm to guide consumers toward lower-cost shipping options that are also well-liked products. Specifically, we propose a ranking policy that balances two objectives: (i) maximizing the probability of consumer conversion (as in the business-as-usual ranking), and (ii) minimizing expected shipping costs by preferentially promoting products that are less expensive to deliver to each consumer. Under this policy, costly-to-ship products are systematically replaced with lower-cost alternatives, subject to preserving conversion likelihood.

To evaluate the proposed approach, we leverage a second large-scale field experiment conducted on the platform. The experiment ran for 12 days and encompassed the entirety of platform traffic. Consumers were randomly assigned to one of three conditions. In the control group, consumers were exposed to the business-as-usual product rankings. In the first treatment condition, consumers viewed the same rankings as in the control group, except that products that were expensive to ship were systematically replaced with close substitutes that were less costly to deliver. In the second treatment

[8]For instance, the construction of fulfillment centers can entail costs upward.

condition, consumers were shown a re-ranked version of the same set of substituted products from the first treatment, where the ordering prioritized products with lower shipping costs. Thus, while the product assortment was held constant, the re-ranking algorithm optimized the product display sequence to further reduce expected shipping costs, conditional on maintaining high conversion probabilities.

Using this experiment we find that replacing products with cheaper-to-ship alternatives without re-ranking them has negligible effect on the shipping efficiency. However, the treatment condition in which products were re-ranked, i.e., consumers are algorithmically nudged towards these products, had significant positive effects on the shipping efficiency.

In the re-ranking treatment group, consumers shifted toward better-placed products, which were disproportionately stocked in centralized distribution centers, resulting in a 16.8% increase in orders fulfilled from centralized locations. Consequently, average shipping distances per-order decreased by 34 miles (a 4.6% reduction) and average delivery times fell by 1.8%. Although we do not observe significant changes in product return rates, incident rates—such as delayed or damaged shipments—declined by 11% in the re-ranking condition. Since the distances traveled by the products have decreased, we hypothesize that the reduction in incident rates is attributable to the shorter travel distances. Overall, the platform realized shipping cost savings of between 1.3 and 7.3%, and consolidation rates improved by 4.4%.

The main takeaway from this experiment is that re-ranking products to guide consumer demand toward items that jointly maximize conversion probability and minimize shipping costs can yield meaningful operational efficiencies. Importantly, algorithmic re-ranking may represent one of the most cost-effective interventions available to platforms: unlike fulfillment center optimization or consumer-facing incentives, re-ranking requires minimal infrastructure changes and avoids the need to offer costly rewards for slower shipping.

More broadly, our results suggest that subtle modifications to choice architecture, rather than costly supply-side investments or direct consumer incentives, may offer scalable and effective pathways for retailers seeking to balance speed, cost, and sustainability objectives.

2. Related Literature

This paper contributes to the growing literature on shipping logistics, shipment consolidation, and platform optimization. A substantial body of work in operations research examines shipment consolidation under various logistical constraints. Xu, Li and Chen (2024) develop analytical and computational frameworks for consolidation under multiple shipping methods and delivery deadlines. Wei, Kapuscinski and Jasin (2021) study consolidation across two warehouses, characterizing how delivery timing shapes optimal policies. Sonntag, Schrotenboer and Kiesmüller (2023) introduce a chance-constrained inventory routing model that balances shipment timing with inventory control. Other contributions, such as Dror, Ball and Zäpfel (2007), explore cost-sharing mechanisms in multi-product consolidation settings. More

recent work incorporates machine learning into consolidation decisions—for example, Cheng, Hijazi, Konak, Erera and Hentenryck (2025) identify spatio-temporal patterns to optimize logistics efficiency.

In parallel, research in operations and transportation economics has emphasized the environmental dimensions of logistics. Kang, Song, Xu, Miller, Wang, Zhang, Liu, Zhou, Ren, Zhong et al. (2021) propose low-carbon delivery pathways in China's express sector, while Tran and Lam (2022) and Cristea, Hummels, Puzzello and Avetisyan (2013) analyze emissions in container shipping. Juan, Faulin, Pérez-Bernabeu and Jozefowiez (2014) show that collaborative fulfillment across firms can reduce emissions, and Van Loon, Deketele, Dewaele, McKinnon and Rutherford (2015) provide a comparative analysis of carbon footprints in retail logistics. Closely related to our study, Muñoz-Villamizar, Velázquez-Martínez, Haro, Ferrer and Mariño (2021) simulate how fast shipping alters product routing and hampers consolidation in the mattress category.

By contrast, relatively little research in economics or marketing directly addresses shipping costs or consolidation. A growing marketing literature examines how sustainability claims affect consumer behavior. Brecko and Kim (2023) and Bollinger, Kronthal-Sacco and Zhu (2023) study sustainability labeling in the packaged goods market. Recent work has also examined Amazon's sustainability claims and their influence on consumer choice (e.g., Proserpio et al., 2025, Wang and He, 2024). Our second experiment relates to this literature: we prioritize lower-cost (and often more environmentally friendly) products through re-ranking, but do so without disclosing the ranking logic to consumers. A natural extension would be to test whether explicitly informing consumers of the ranking criteria affects outcomes—a direction we leave for future research.

Lee and Donohue (2024) conduct a lab experiment with 3,000 participants to evaluate how messaging strategies influence willingness to choose slower, more sustainable shipping options. They find that emphasizing environmental impact and offering transparent information about fulfillment increases the likelihood of choosing slower delivery. Their results suggest that informational nudges can shift demand without financial incentives.

More broadly, our work connects to literature in transportation economics and supply chain management. For example, Cristea, Hummels, Puzzello and Avetisyan (2013) study the role of transportation costs in trade flows, while Kellogg and Reguant (2021) examine the interaction between environmental regulation and supply chain operations. Our contribution is distinct in focusing on shipment consolidation at the consumer-platform interface, using experimental variation rather than structural modeling.

Other tangentially related studies address grocery-store waste (Sanders, 2020) and post-consumer packaging waste (Duan et al., 2019), contributing to a broader agenda on sustainable logistics.

This paper makes two main contributions to the existing literature. First, to our knowledge, it provides the first experimental evidence that removing fast-shipping options from the platform is unlikely

to yield significant cost or environmental savings. While simulation-based studies often suggest gains from delayed shipping, our findings highlight the limitations imposed by fulfillment infrastructure designed for speed. In practice, this infrastructure constrains the platform's ability to consolidate shipments ex post.

Second, we shift the focus from supply-side logistics optimization to a demand-side intervention embedded in the platform's ranking algorithm. We show that subtle re-ranking to favor cheaper-to-ship items can reduce shipping costs and improve consolidation—without changes to infrastructure or reliance on consumer-facing incentives or disclosures. This choice architecture approach offers a scalable and cost-effective complement to traditional logistics strategies.

The remainder of the paper is organized as follows. Section 3 describes the empirical setting. Section 4 presents the data and descriptive statistics. Section 5 discusses the first experiment, which partially removes fast-shipping options, and presents our main findings. Section 6 reports results from a second large-scale field experiment. Section 7 concludes.

3. Empirical Setting

In this paper, we use data generously provided by Wayfair (hereafter, the 'platform'), a large US-based online platform of furniture and home goods. Wayfair does not own sellers' inventory, instead acting as a platform (intermediary) between product sellers and consumers. In this section, we walk the reader through the platform's interface design, explain typical consumer search process, product ranking rules, supplier and platform decisions, and the role of fast delivery promises.

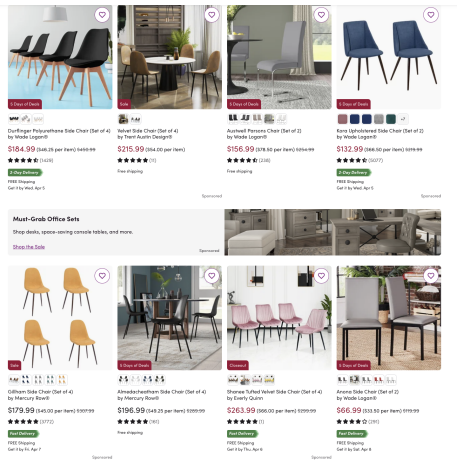
3.1. Platform interface

When consumers arrive on the platform, they can either enter a search query via a keyword box or select from predefined product categories (e.g., chairs, kitchen appliances). Once a category is selected, consumers are directed to ranking pages. Each ranking page displays 48 products, organized in a 12×4 grid (Figure 1a). Each product tile on a ranking page includes an image, sale price, average rating, number of reviews, and—if applicable—a fast delivery badge indicating 1- or 2-day shipping. These green badges, shown beneath the price, also display the expected calendar delivery dates: 'Get-it-by' certain date. Fast-shipped products typically qualify for free shipping, and therefore, would have 'Free 1-Day' or 'Free 2-Day' labels. A lot of products on the website would also have 'Free Shipping' labels, which means that consumer does not pay separately for shipping, but it does not guarantee the delivery date. Starting 2022, the platform started displaying new badge 'Free Fast Shipping', which are for products that have Free Shipping and, typically, will arrive within five days . These badges are introduced as additional product differentiation tool to promote products that have guaranteed delivery, albeit not 1-2 days.

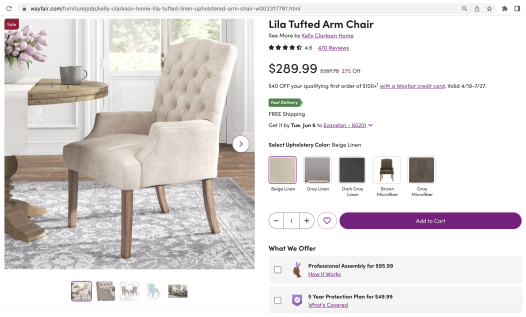
Once on a ranking page, consumer has an option to scroll down the ranking page or click on the

FIGURE 1: Platform Interface

(a) Product Ranking Page



(b) Product Detail Page



Notes. The figure illustrates a typical product ranking page (left) and product detail page (right) on wayfair.com.

next page to see the next set of 48 products. Clicking on a product on the ranking page takes the user to a product detail page (Figure 1b), where they can view more comprehensive information, such as specifications and customer reviews. The fast delivery badge remains visible on this page, accompanied by an inferred consumer ZIP code that determines delivery eligibility. Users can manually edit this ZIP code—either to reflect their actual location or the destination ZIP if the item is being shipped elsewhere. We will discuss the importance of consumer recognition for identifying these ZIP codes.

Following exposure to the ranked product assortment, consumers may continue browsing, complete a purchase, or exit the platform. These interactions follow a standard pattern common to most e-commerce environments.

Refinement behavior. Additionally, consumers can also refine search results using sorting or filtering tools (e.g., by price, color). In 2019, only about 10% of users employed such refinement actions, with the vast majority applying a single filter with a single value (see Figure E9). Given this limited usage, we abstract from refinement behavior in the analysis of the first experiment but incorporate it explicitly in the analysis of the second experiment, where these actions became slightly more prevalent (18%). Even then, still most people would apply a single filter with a single value (see Figure E10).

Below, we highlight several platform-specific features that are relevant to understand the experimental interventions.

3.2. Product Ranking Algorithm

In 2019, during the timing of the first experiment, product rankings were determined by a scoring rule based on consumer engagement and delivery eligibility. Each product j received a base score:

$$score_{jt} = clicks_{jt} + 100 \times orders_{jt}, \tag{1}$$

where $score_{jt}$ is the score of product j at day t, $clicks_{jt}$ is the number of clicks the product j got within the last 90 days before day t, and $orders_{jt}$ are the number of orders that product j had within the last 90 days before day t.

This score was then adjusted depending on the product's fast shipping status. Products eligible for 2-day delivery received a 30% score boost, while those with 1-day delivery received an additional 10% boost, resulting in the final rule:

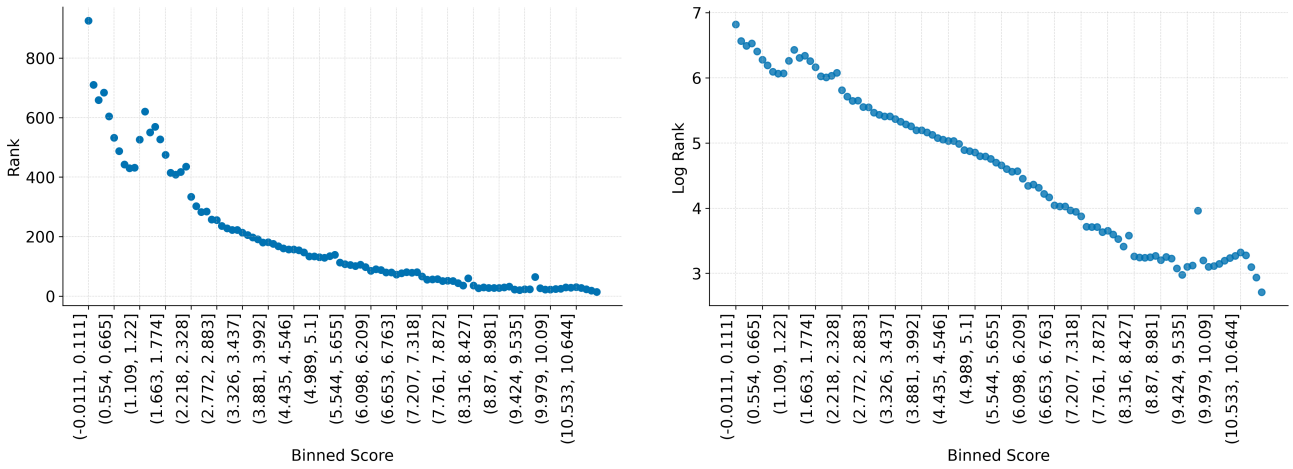
$$\begin{cases} score_{jt} = clicks_{jt} + 100 \times orders_{jt} \\ score_{jt} \times 1.3 \qquad \text{if 2-day delivery eligible} \\ score_{jt} \times 1.3 \times 1.1 \quad \text{if 1-day delivery eligible} \end{cases} \tag{2}$$

To verify that the platform's stated ranking algorithm was indeed implemented during the study period, we reconstructed product scores using the formula provided in Equation2 and compared these to the actual product ranks observed in the data. Figure 2 plots binned, rescaled scores on the x-axis against observed product ranks. Panel (a) displays absolute ranks, while Panel (b) shows the logarithm of ranks. In both cases, the monotonic and approximately linear relationship confirms that higher scores correspond to better (i.e., lower numerical) ranking positions, validating that the score-based ranking algorithm was operating as intended.

FIGURE 2: Validation of the Ranking Algorithm

(a) Rank by Binned Scores

(b) Log(Rank) by Binned Scores



Notes. These plots validate the platform's ranking algorithm. The x-axis displays rescaled score bins (for anonymization), and the y-axis shows absolute or log ranks. Higher scores correspond to better (i.e., lower) ranks, consistent with the platform's ranking logic.

In 2022, during the second experiment, the ranking algorithm became dynamic. The logic of ranking products based on their clicks and orders remained, however, now the algorithm became Bayesian. While we cannot disclose the actual algorithm, at a high level, the platform would construct a prior distribution of the product quality based on historical clicks and purchases. Next, as new clicks and orders arrive, the platform updates the posterior score for each product. The rank of the product is determined by a score drawn randomly from the product's posterior distribution using the Thompson Sampling algorithm (Thompson, 1933; Russo et al., 2018). The algorithm re-calculates scores several times within a day to account for contemporary clicks and orders.

Theoretically, the platform could just rank the products according to the scores in 2019 and according to the posterior distributions in 2022. However, the platform also has to take into account the product inventory to make sure the product is available. Moreover, they have to make sure that the badges are displayed correctly. Thus, the platform calculates (i) product scores, boosts them, (ii) checks product availability, and (iii) delivery guarantees. This entire process—from supplier filtering to badge assignment—occurs dynamically as consumers interact with the site.

Personalization. Product rankings on the platform were not personalized in 2019. Personalized ranking was introduced in 2020, after the completion of our first experiment.[9] Although in 2019, there was no ranking personalization, the rankings could differ across the consumers due to the fast shipping options.

In contrast, our second experiment was conducted after the introduction of personalization. Throughout the analysis, we explicitly note where and when personalization may influence outcomes in the second experiment.

Next, we describe the logic and implementation of the fast shipping badges.

3.3. Fast Shipping Badges

Historically, online platforms would match the consumers with the third-party sellers through the platform's website, and once the match is made, the third-party sellers would send the product to the consumers directly. This way of shipping is called *Dropship*. Under the *Dropship* model, the third-party seller retains the inventory and ships the item directly to the consumer from its own warehouse. Though the platform doesn't participate in the shipping process, they usually handle the shipping costs themselves.[10]

However, to facilitate shipping, in recent years, Wayfair has built multiple distribution centers that are located across the U.S., e.g., in California, Texas, and Georgia. These platform-owned warehouses are called *Castlegate* and create an alternative shipping option.[11] See Figure 3 for publicly available information on the locations of the Castlegate warehouses across the United States. To enable centralized

[9] Korganbekova and Zuber (2023) study personalization on Wayfair and confirm this.

[10] The seller sends the invoice to the platform, and the platform pays for shipping.

[11] See the Equity Research for the map of Castlegate warehouses.

logistics, third-party sellers have the option to pre-position inventory within Castlegate facilities. In operations management, this arrangement is commonly referred to as *forward positioning*, whereby sellers allocate inventory to the platform prior to demand realization. Although sellers retain ownership of the goods, they pay Wayfair a fee for storage and delegate fulfillment to the platform. Once an order is placed, Wayfair handles the shipping directly from the Castlegate warehouse.

Table 1 illustrates that while the Dropship method remains the primary fulfillment approach for the majority of orders (63%), there is a significant proportion of shipments (36%) accounted for by Castlegate or mixed Castlegate-shipped items.[12]

TABLE 1: Summary statistics

Fulfillment Type	**%Orders**	**Travel metrics**			
		Miles	**Weight**	**Girth**	**Locations**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**
Dropship	63.2%	1123.9	11.7%	5.0%	59,101
Castlegate	8.5%	793.8	7.5%	3.3%	25,894
Mixed	28.0%	936.8	11.5%	10.8%	17,950
Other	0.3%	852.6	11.7%	5.0%	8,637
Overall	100%	931.0	24.3%	10.7%	72,098

Notes. % Orders indicate the fraction of orders for each fulfillment type. Miles, Weight, Girth, Locations are corresponding metrics by fulfillment type.

FIGURE 3: Castlegate warehouses across the US



Source: Screenshot taken on April 20, 2025, from sell.wayfair.com website.

Products housed in Castlegate warehouses typically get fast shipping badge because the platform can ensure timely delivery to the consumer. These products are frequently eligible for "Free 1-Day" or "Free 2-Day" delivery badges, which are displayed prominently to consumers. However, Dropship sellers can also get the badge if (i) they can guarantee that the product will be delivered within promised number of days, 1 or 2 days, and (ii) if the seller maintains a good history of on-time delivery for fast badged products.

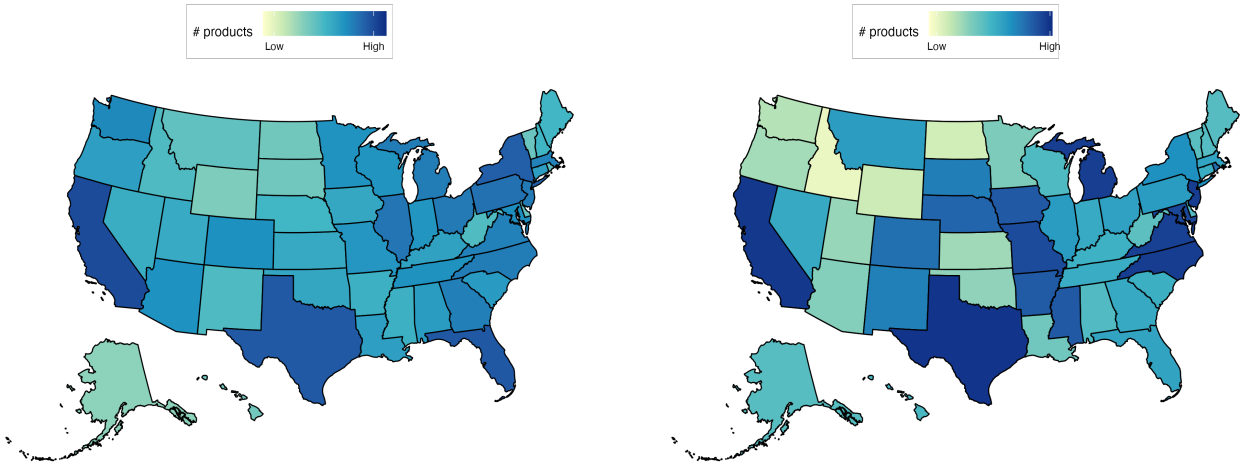
[12] *Other* refers to cases where the platform purchases inventory from sellers for various business reasons and handles the shipping itself—typically through Castlegate, though not exclusively.

While a lot of regions of the US are eligible for free shipping, the 1- or 2-day shipping badges would depend on whether the product is located in a warehouse (either Castlegate or Dropship) that can realistically ship in 1- or 2-days. Figure 4 shows the distribution of the number of products that have Free Shipping or Free 1- or 2-Day shipping. Figure 4a shows that the free shipping is more or less uniformly distributed across the US. However, Figure 4b shows that there are some states that have high concentration of fast shipping and some states have less.

FIGURE 4: Distribution of product types across the US

(a) Free Shipping

(b) Free 1- and 2-day Shipping



Notes. This figure shows the distribution of the number of products across the US by product types. Figure 4a shows the number of Free Shipping products across different states. Figure 4b shows the number of Free 1- and 2-day badged products across the US.

Consumer Recognition. To be able to reliably display the shipping badge the crucial part is also determining the zipcode of whether the consumer resides. A key challenge for the platform is recognizing individual consumers across devices and sessions. Approximately 40% of users log in, allowing for deterministic identification (Korganbekova and Zuber, 2023). For the remaining users, the platform relies on a combination of first-party cookies (issued by the platform itself) and third-party cookies (from affiliated domains) to probabilistically link browsing behavior across sessions and devices. In addition, a backfilling procedure is used to retroactively associate sessions with known identities once a user becomes identifiable. This mechanism enables the platform to map device-browser pairs to a unified consumer profile.

Accurate consumer recognition is particularly important for two reasons. First, it ensures the integrity of randomized experiments by maintaining consistent user assignments to treatment or control groups

across browsing sessions. Second, it enables accurate estimation of delivery eligibility by linking users to ZIP code–level geographic locations. ZIP codes were inferred either from user-provided information upon login or via IP-based geolocation services. While consumers could manually update their ZIP code on the product detail page, the platform preserved the product ranking shown on the search page for a 30-minute window to maintain a consistent user experience. This freeze prevented rank reshuffling that might otherwise result from changes in inferred delivery eligibility.

Importantly, fast-shipping badges were only displayed to consumers located within the estimated two-day delivery radius of the relevant warehouse.[13] Thus, a product eligible for fast shipping may show a badge for some users but not others, depending on geographic proximity. For instance, a user in Hawaii would not see the badge for a product stored in a continental U.S. warehouse. This logic ensures that fast-delivery promises are shown only when the platform can confidently guarantee fulfillment within the stated time frame.

3.4. Supplier Choice and Order Fulfillment

As described above, consumers who choose to purchase a product receive fulfillment through one of two distinct channels. Under the *Dropship* model, the third-party seller holds inventory and ships directly to the consumer from its own warehouse. Alternatively, orders may be fulfilled through *Castlegate*, Wayfair's proprietary network of distribution centers strategically located across the United States (e.g., California, Texas, Georgia).

While consumers see the product's brand on the website, they do not observe which seller will fulfill the order. Once a purchase decision is made, the platform's fulfillment algorithm selects the seller. A given product may be offered by multiple sellers, each with inventory at several warehouse locations. The algorithm filters the candidate set to sellers and locations capable of meeting the delivery promise shown at checkout.

For each candidate, the platform evaluates available shipping routes via major carriers (e.g., FedEx, UPS), estimating transit times and shipping costs based on historical data. Outbound shipping costs are entirely variable and include carrier base rates as well as transportation costs across the first, middle, and last mile—particularly for third-party logistics partners. Additional charges include parcel surcharges (e.g., for oversized or irregular packages), packaging-related fees (e.g., for cylindrical boxes), and other variable costs such as fuel and delivery area surcharges.

Among eligible suppliers, the algorithm compares expected delivery times and costs to identify the most efficient option. For bundled or kit products, it prioritizes sellers able to fulfill all components together, reducing the likelihood of split shipments. The decision process balances speed, cost, and

[13] That is, the warehouse location at which the item is stored determines the visibility of the two-day shipping badge, conditional on the consumer's inferred location.

fulfillment feasibility at the time of order allocation.

If the consumer updates the shipping ZIP code on the cart page, the algorithm is rerun and the process restarts.

In summary, the platform operates as a search-based e-commerce marketplace. Product rankings are influenced by historical performance, inventory availability, and delivery speed. Products with stronger delivery promises receive higher placement. Once a product is selected, a separate fulfillment algorithm assigns the seller based on shipping routes and costs. Given the scale of operations, shipping expenditures are substantial, and are further amplified by fast-delivery guarantees.

We next describe how both private actors and regulators have responded to these fulfillment practices.

3.5. Policy Setting

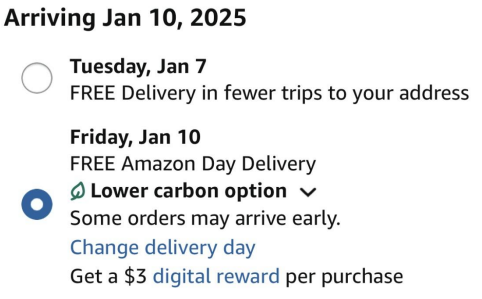
Fast-delivery badges serve as a salient signal of fulfillment quality and can influence consumer purchase decisions. Fast delivery commitments—such as 1,2-day, next-day or same-day shipping—constrain the platform's ability to pool items across consumer orders and optimize delivery routes. In contrast, slower delivery windows allow for more efficient batching, warehouse assignment, and last-mile routing. The trade-off is operational: faster delivery improves consumer experience but increases fragmentation in shipment-level logistics.

To counteract this inefficiency, some retailers have begun deploying incentive-compatible mechanisms that nudge consumers toward lower-impact choices. For example, Amazon offers a "No-Rush" or "Low-carbon" delivery option that delays shipment in exchange for monetary or credit-based rewards. These delays give the platform additional flexibility to consolidate products and assign them to lower-cost, lower-emission routes. See Figure 5a for an example. These interventions reflect an increasing convergence between logistics optimization and environmental signaling.

In addition to no rush options, platforms have also introduced sustainability labels to influence purchasing decisions. Amazon's "Climate Pledge Friendly" badge highlights products certified to meet third-party sustainability standards (Figure 5b).

Figure 5: Sustainability Options on Amazon

(a) Lower Carbon Delivery Option



(b) Climate Pledge Friendly Badge



Similarly, many retailers invest in micro-fulfillment centers and route-optimization software reflect broader efforts to manage the logistics-environment trade-off without compromising delivery quality.

However, fast shipping also has implications for supply chain efficiency and environmental outcomes. Regulators have expressed concern that the proliferation of 1-day and 2-day delivery promises may lead to fragmented logistics and higher emissions.

Several U.S. states have introduced or are considering retail delivery fees to address the negative externalities of e-commerce logistics. These policies are motivated by growing concerns over infrastructure wear and urban congestion resulting from the expansion of last-mile delivery networks. Colorado and Minnesota have implemented per-order surcharges to fund road maintenance and transportation infrastructure. Other states, including Hawaii, New York, Maryland,[14] are evaluating similar proposals. Platforms may absorb these fees as part of their shipping costs or pass them through to consumers. In economic terms, these policies represent an attempt to internalize the social costs of commercial delivery, analogous to congestion pricing mechanisms commonly studied in transportation economics.

Taken together, these trends suggest a broadening set of institutional levers—both regulatory and platform-based—for influencing fulfillment practices. In the empirical sections that follow, we examine whether fast delivery promises generate meaningful differences in shipping costs and environmental impact, and to what extent platform design and algorithmic changes can mediate these effects.

4. Data

In this section, we describe the main components of our data. The main samples consist of two large-scale field experiment. The first experiment ran for 1.5 months across the entire website and all product categories. The second experiment ran for 12 days. With small exceptions, we have very similar data for both experiments.

Clickstream data. At the consumer-device-session-level, we observe full consumer browsing history on the website. Starting from the moment when consumers land on the website, we observe which categories they searched for and the ordered list of products they observed on the ranking pages. For each product on each ranking page, we observe a product's price, ratings, numbers of ratings, and the availability of the fast shipping badge, the get-it-by dates, and promised delivery days. Thus, we can fully re-create the view of the page as it was shown to the consumers. Next, we observe whether consumers clicked on an item, added it to cart, or purchased.

Transactions and shipping data. At the consumer-level, we observe which products were ordered by a consumer, time of transaction, quantity, price, destination state, city, and zip code. From the clickstream data, we know whether the chosen product had fast delivery promises. We also observe which supplier was chosen to fulfill the order, and the address of a warehouse from which a product was

[14] See this article for more details.

shipped and consumers' destination address.

Shipment tracking data. We observe shipment tracking information provided by FedEx and UPS. Specifically, we observe the timestamp and origin scan of the shipment and all subsequent tracking scans up to the destination together with the timestamps. This data allows us to analyze whether multiple shipments were consolidated and shipped together from the warehouse or added en route to the same destination. Moreover, it allows us to credibly estimate how loaded a truck was which is crucial to analyze the shipping efficiency metrics, such as order consolidation. Moreover, this data allows us to calculate full travel history of the product from the seller's warehouse to the Castlegate and to the final consumer.

Product characteristics. We also use data on product characteristics, such as weight and girth of the products. This allows us to distinguish small, medium, and large parcel items.

Seller characteristics. For each seller on the platform, we observe the complete set of products they offer, the introduction date of each product, and the date the seller first joined the platform and the date they joined Castlegate (if applicable). Sellers are typically companies operating multiple warehouses; our data are at the seller-warehouse level. For each warehouse, we observe the opening date and, if applicable, the contract signature date corresponding to the seller's enrollment in the platform's centralized distribution network.

Prices and product wholesale and shippings costs. We observe a daily panel of each product's prices, wholesale costs set by the seller, any discounts/allowances provided by the seller and shipping costs. We use these data to calculate product markups and margins to assess the platform's profitability.

Product returns and incidents data. For each product ordered on the platform, we observe a post-purchase return indicator, capturing whether the product was ultimately returned by the consumer. Additionally, for the second experiment—which was conducted at a later date—we observe an incidence indicator, identifying whether the product experienced damage or other delivery-related issues during shipment.

Score calculations. We have access to the full historical data on product performance so that we are able to calculate the product scores for the first field experiment. Moreover, we have a field daily panel of algorithm posterior product scores, to calculate the posterior scores for the second field experiment.

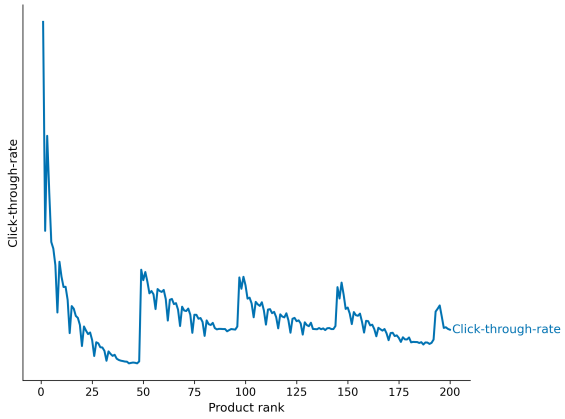
Overall, we have rich datasets that track consumer actions, platform design, and seller and characteristics, and shipping information.

4.1. Descriptive Statistics

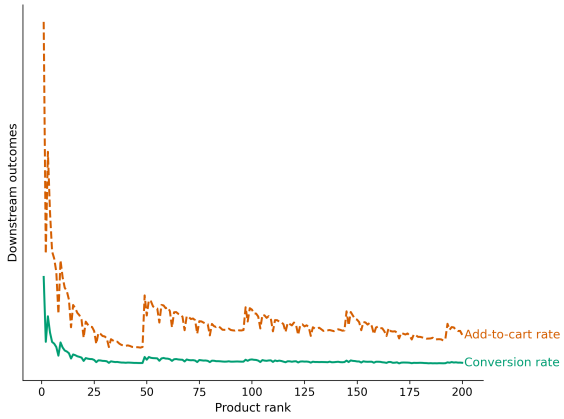
Importance of Product Rankings. First, we show that product rankings are important in guiding consumer choice. Figure 6a shows that the click-through-rate[15] is highest at the lower ranks (i.e., for top-ranked products). The discontinuous jumps are associated with the next pages. For example, the first product on page 2 has rank 49 and it has a higher click-through rate than the last product on page 1. Figure 6b shows that most add-to-cart rate and conversion actions also happen for better-ranked products.

Figure 6: Descriptive Ranking effect

(a) Click-through-rate by product rank



(b) Add-to-cart and conversion rate by product rank



Notes. This figure shows the click-through-rate, add-to-cart, and conversion rates per product rank.

Similar findings are reported by Ursu (2018) in the context of Expedia's hotel rankings, highlighting that product rankings can meaningfully influence consumer choice. However, identifying the underlying mechanism is inherently a causal question: are consumers responding to the ranking itself, to their beliefs about how the platform generates rankings, to underlying preferences, or to some other factor? In our setting, observational data alone cannot distinguish whether the observed patterns reflect these mechanisms or are primarily driven by fast delivery promises. This underscores the importance of conducting a large-scale field experiment to isolate and identify the causal effects of interest.

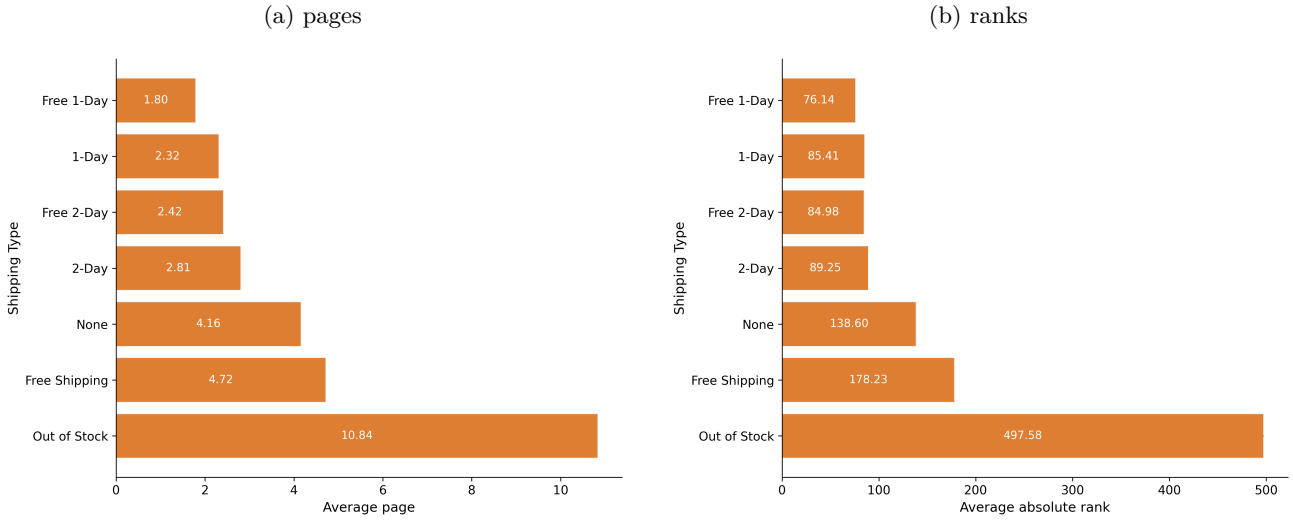
Fast Shipping Products on the Ranking pages. As was described in the previous section, fast delivery products get a boost in the ranks. Figure 7 shows that indeed, fast delivery products are consistently ranked higher than the non-fast shipped ones both in terms of the pages and the absolute ranks. Figure 7a shows that fast shipped products are on average shown on pages 1.8-2.8 (up to the third page). Figure 7b shows the similar patterns in the absolute rank terms.[16]

These descriptive patterns could arise due to several factors: (i) fast-shipped products are boosted by the platform's ranking algorithm (see Subsection 3.2) and therefore receive better placement and

[15] Click-through-rate is calculated as the number of clicks divided by the number of impressions.

[16] Recall that each page contains 48 products.

FIGURE 7: Average pages and absolute ranks by shipping type



Notes. TThe figure illustrates how products with different shipping types are positioned in the platform's ranking. It is based on one year of pre-first-experimental data across all product categories and includes only unfiltered, unsorted ranking pages.

greater consumer attention; (ii) consumers may have an inherent preference for fast shipping as a product attribute; and (iii) fast-shipped products may be higher-quality offerings overall. The experimental variation in our data will help to disentangle these potential mechanisms. For now, Table 2 shows that fast-shipping products are, on average, cheaper and have higher ratings and more reviews than products offering free shipping, those without shipping labels, or those that are out of stock.[17]

TABLE 2: Summary Statistics by shipping types

	Panel A: Price ($)					
	Observations	Min	Mean	Median	Max	SD
	(1)	(2)	(3)	(4)	(5)	(6)
1-Day	684,049	7.93	159.58	79.96	1,305.80	213.74
2-Day	223,557	3.48	27.14	24.99	229.40	25.75
Free 1-Day	505,419	14.99	195.03	107.83	1,499.99	242.38
Free 2-Day	882,031	13.99	199.33	103.99	1,719.99	267.13
Free Shipping	5,590,983	14.99	336.83	123.48	3,765.00	585.93
None	2,195,033	7.23	239.34	59.33	2,945.44	465.09
Out of Stock	1,541,654	11.11	447.15	164.99	4,298.28	730.11

[17] Out-of-stock products tend to be more expensive. One possible explanation is that these products are higher priced and therefore supplied more conservatively to match demand. Alternatively, limited availability may reflect high consumer demand or logistical constraints.

	Panel B: Ratings					
	Observations	Min	Mean	Median	Max	SD
	(1)	(2)	(3)	(4)	(5)	(6)
1-Day	525,597	1.00	4.57	4.67	5.00	0.51
2-Day	165,307	1.00	4.56	4.68	5.00	0.55
Free 1-Day	405,853	1.00	4.58	4.67	5.00	0.49
Free 2-Day	650,081	1.00	4.57	4.68	5.00	0.52
Free Shipping	1,942,796	1.00	4.52	4.71	5.00	0.67
None	587,419	1.00	4.52	4.69	5.00	0.66
Out of Stock	724,141	1.00	4.49	4.67	5.00	0.67

	Panel C: Reviews					
	Observations	Min	Mean	Median	Max	SD
	(1)	(2)	(3)	(4)	(5)	(6)
1-Day	525,597	1.00	170.67	13.02	149,868.35	1,508.09
2-Day	165,307	1.00	202.21	10.76	150,167.39	2,365.22
Free 1-Day	405,853	1.00	192.09	16.00	152,937.43	1,531.80
Free 2-Day	650,081	1.00	153.00	12.00	153,507.12	1,299.59
Free Shipping	1,942,796	1.00	78.82	5.00	151,905.00	876.02
None	587,419	1.00	122.24	6.00	149,988.39	1,425.58
Out of Stock	724,141	1.00	69.10	6.00	109,569.20	751.63

Notes: This table shows summary statistics, i.e., prices, ratings and # of reviews, by the product type.

We further document systematic differences between Dropship and Castlegate products. Table 3 shows that, on average, Castlegate products have higher ratings, a greater number of ratings and reviews, and—critically—lower return rates compared to Dropship products. These patterns suggest higher levels of customer satisfaction with Castlegate-fulfilled items. Although incident rates also differ statistically between the two groups, the magnitudes are similar and distinct from return behavior, indicating that lower return rates are not merely driven by fewer reported incidents. All differences reported in the table are statistically significant.

Thus, fast shipping products seem to be better quality products and they are ranked higher. Next,

TABLE 3: Differences between Dropship and Castlegate products

Variable	Mean Castlegate	Mean Dropship	Difference	t-stat	p-value
Average Rating	4.48	4.39	0.09***	116.77	0.0
# Ratings	99.72	61.23	38.5***	53.51	0.0
# Reviews	70.46	45.27	25.18***	44.86	0.0
Return Rate	7.66	13.56	-5.89***	-311.33	0.0
Incident Rate	0.34	0.34	-0.0***	-13.20	0.0

Notes. This table reports mean comparisons between Castlegate and Dropship sellers across key performance metrics. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

we want to briefly describe the types of sellers that choose to be part of Castlegate.

Seller characteristics. Sellers that eventually join Castlegate differ systematically from those that do not. Prior to joining, these sellers exhibit higher revenues, lower prices—both in raw terms and when weighted by quantity sold—and sell a larger number of products. Table 4 reports t-test results comparing pre-joining characteristics of Castlegate and Dropship sellers.[18]

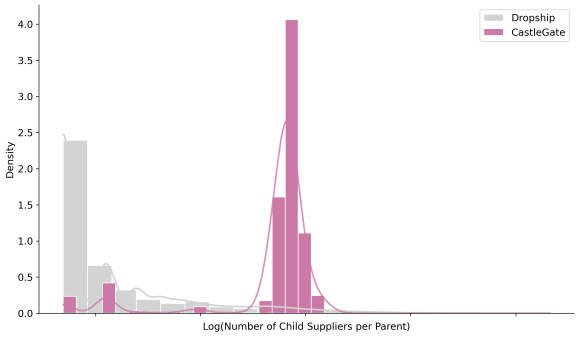
TABLE 4: Differences between Dropship and Castlegate products

Variable	Pre-Mean Castlegate	Pre-Mean Dropship	Difference	t-stat	p-value
Pre Revenue	287.9	29.8	258.11***	3.72	0.0002
Pre Price	300.06	364.2	-64.14*	-2.24	0.03
Pre Weighted Price	299.21	363.01	-63.80*	-2.23	0.03
Pre Products	30.53	2.89	27.64***	140.6	0

Notes. Table presents pre-period summary statistics for Castlegate and Dropship sellers across revenue, price, and purchased product count. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Moreover, Castlegate suppliers tend to be larger sellers, as reflected in their broader warehouse footprint. Figure 8 plots the distribution of the log number of warehouses for Castlegate sellers (purple histogram) and Dropship sellers (gray histogram), showing that Castlegate sellers maintain a larger number of warehouses on average.

FIGURE 8: Distribution of the log(# of warehouses) per parent supplier by Castlegate status



Notes: This figure shows the # of warehouses per parent supplier by Castlegate status.

[18] See Table D3 for additional descriptives.

Product characteristics. While we cannot disclose exact figures, the distribution of small versus large parcel items is broadly similar between Dropship and Castlegate channels. The majority of orders in both cases qualify as Small Parcel deliveries, with Castlegate handling a modestly higher share of large parcels—on the order of a single-digit percentage point difference relative to Dropship.

Overall, the descriptive statistics underscore the role of ranking algorithms in shaping consumer behavior. Fast-shipping products are prominently featured in rankings and appear to be of higher quality. Similarly, sellers participating in Castlegate tend to be larger and more established. In light of growing cost pressures from logistics providers and emerging state-level regulations (as discussed in prior sections), we first examine the effects of reducing fast delivery promises. We then evaluate an algorithmic re-ranking of product assortments designed to enhance shipping efficiency without compromising the consumer experience.

5. Field Experiment #1: Reducing Fast Delivery Promises

Motivated by the fact that both corporate partners (e.g., UPS) and regulators (e.g., the states of Colorado and Minnesota) have expressed concerns that fast delivery promises create shipping inefficiencies, we begin by examining the consequences of reducing such promises. The main experiment ran for 41 days, from mid-June to the end of July 2019, and spanned the entire traffic on `wayfair.com`. The experiment was paused during the July 4th sales period.[19,20]

During the main experiment, 30% of traffic was assigned to the control group, which experienced business-as-usual product rankings and fast delivery promises. The remaining 70%, assigned to the treatment group, saw a reduced number of fast delivery promises. Specifically, for each treated visitor, the platform: (i) removed between 25% and 75% of fast shipping badges from products on ranking pages; (ii) pushed back the corresponding get-it-by dates; and (iii) removed the fast shipping sort boosts.

Figure 9 illustrates the control condition, in which fast-shipped products display the badge and delivery promise. In the treatment condition (Figure 10), the fast delivery promises (green badges) were removed, and delivery dates were extended by 1–3 days to reflect standard (non-fast) shipping timelines. Additionally, score boosts used to promote fast-shipping products in the ranking algorithm were removed.

As a result, the treatment effect reflects a combination of removing delivery promises and altering product rankings. Later in this section, we detail the specific changes implemented during the experiment and present evidence suggesting that most of the treatment effect can be attributed to badge removal, rather than rank changes.

[19] The exact dates used in the analysis are June 12–29, 2019, and July 7–29, 2019.

[20] The platform retained 6% of traffic in a holdout group to continue the experiment through November 2019, but our analysis focuses on the main experiment period only.

FIGURE 9: Control condition

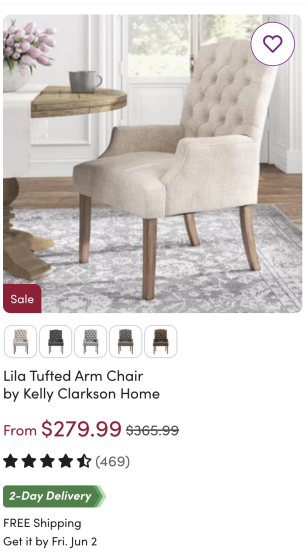
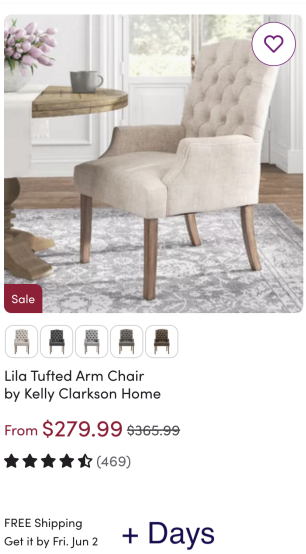


FIGURE 10: Treatment condition



Operationally, randomization was conducted at the device level. Each device was randomly assigned an h-value, where $h \in [25, 75]$, indicating the percentage of eligible fast delivery promises to be hidden from the consumer's ranking page. The h-values were assigned at the beginning of the experiment and remained fixed throughout its duration at the device level. Following this assignment, each device was then randomized into either the treatment or control group. Hidden badge means that there was no green badge and the product's delivery date was slowed - we will use the words hidden and slowed interchangeably throughout this section.

Treatment Assignment. Consumers may access the platform using multiple devices, each of which is linked to a unique identifier. While treatment assignment was implemented at the device level—primarily due to practical constraints associated with website infrastructure—the platform algorithmically clusters devices to consumers via a shared email address. Importantly, the platform attempts to ensure that all devices associated with a given consumer are assigned to a consistent treatment condition. In our data, we verify that fewer than 0.2% of consumers appear in both treatment and control groups, a contamination rate well within acceptable bounds for field experiments of this type. Such cross-assignment arises primarily when consumers clear cookies or reset devices during the experimental window, thereby circumventing persistent identifiers. Given the low incidence of such events, we regard consumer-level contamination as negligible for identification purposes.

Treatment Intensity Assignment. While the binary treatment versus control assignment is cleanly implemented, some complications arise in the assignment of treatment intensity. Specifically, approximately 1.72% of consumers are associated with multiple treatment intensity values due to device-level variation: although all devices linked to a given consumer were assigned to the same treatment group, they occasionally received different treatment intensity levels, indexed by the parameter h (see

Figure A1). To address this discrepancy, we adopt the following strategy. We define the consumer-level treatment intensity as the most frequently observed h-value, i.e., the mode of h-value across multiple sessions. We break ties by taking the last session's $h-$value. The rationale is that if there were purchases made, they would be made during the last session consumer made, which increases the importance of the last session's h-value. This procedure ensures that treatment intensity is aggregated in a manner that accurately reflects consumer-level exposure.

Stable Unit Treatment Value Assignment: Rankings. A potential concern is that the experiment may violate the Stable Unit Treatment Value Assumption (SUTVA, Athey and Imbens, 2017), whereby actions taken by consumers in one experimental arm (e.g., treatment) may influence outcomes in the other arm (e.g., control), particularly through endogenous adjustments in product rankings driven by user engagement, e.g., clicks or purchases . However, anecdotal evidence from the platform indicates that the score calculation procedure—detailed in Subsection 3.2—is implemented independently across treatment conditions. In particular, scores used to rank products within the treatment (control) group do not incorporate engagement data from the control (treatment) group. We verify this claim by examining the score construction directly. We show that the scores from the control condition did not have any statistically significant effect on the scores in the treatment condition. See Appendix F.1 for the detailed analysis.

Stable Unit Treatment Value Assignment: inventory. Another way SUTVA could be violated in the experiment is if the demand for some product in one condition is so high that it depletes the product's inventory so that the product becomes out of stock in another condition. In the data, we observe that approximately 21% of the products were out-of-stock during the expeirment, with a lot of products being out-of-stock even before the experiment start date. As a result, only 1.85% of observations were marked as out-of-stock during the experiment on the ranking pages.[21] We further do some descriptive checks whether the orders in the control condition had statistically significant effect on the out-of-stocks that treatment consumers saw. We do not find significant effects. See Appendix F.2 for additional details.

Thus, we alleviate the concerns related to SUTVA violations and treatment assignment. However, we still need to take care of the multiple treatment types, which we do next.

Treatment contamination. The experimental intervention altered both the visibility of fast-shipping badges and the ranking of products, as it removed the product score boosts associated with fast delivery. However, Table 5 shows that, even prior to the experiment, fast-shipping products had substantially higher raw scores than non-fast-shipping products. To assess the potential impact on rankings, we group products into average rank bins in increments of 48—the number of products displayed per page. Across all bins, the raw score gap between fast-shipping and non-fast-shipping products exceeds 200%,

[21] It takes time to remove the out-of-stock products completely from the ranking pages. Moreover, there are a lot of situations where the main option of the product, e.g., white couch, is out-of-stock but there are other options that are still in stock.

suggesting that the removal of boost factors—typically between 30% and 43% (see Subsection 3.2)—is unlikely to meaningfully alter product rankings.

TABLE 5: Score differences b/w Fast Shipped and Non-Fast Shipped products

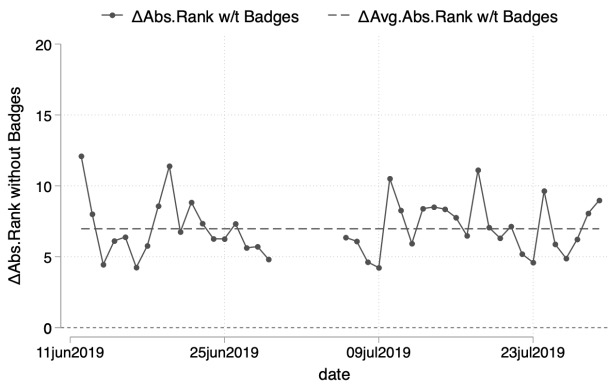
Rank Bin	Non-Fast Avg Score	Fast Avg Score	Score Gap (Fast - Non)	% Score Gap (Fast - Non)
0–48	842.00	3754.65	2912.64	345.92
49–96	277.92	1065.97	788.05	283.55
97–144	171.17	642.10	470.93	275.12
145–192	127.47	475.43	347.96	272.96
193–240	102.91	353.34	250.43	243.36
241–288	88.05	281.37	193.32	219.56
289+	47.25	160.76	113.51	240.22

Notes. This table shows the average scores for fast-shipped and non-fast-shipped products by rank bin, along with the absolute and percentage differences in scores between the two groups.

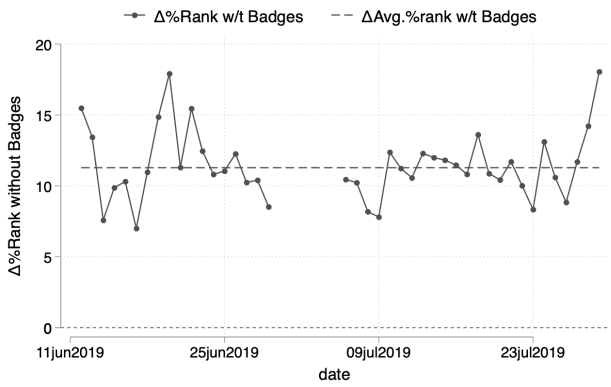
Indeed, Figure 11a shows that absolute rankings increased by only 5-10 positions, and in relative terms this corresponds to approximately 12% increase in rankings (Figure 11b).

FIGURE 11: Effect of Hidden Badge on Product Rankings

(a) Δ absolute rank



(b) Δ % rank



Notes. Figure 11a shows the difference between the absolute rankings of the eligible fast shipped products when the badge was hidden versus when it was shown. Figure 11b shows the percent difference between the rankings of the eligible fast shipped products when the badge was hidden versus when it was shown.

Next, we examine how product rankings changed at the product level, comparing instances in which a product's fast-shipping badge was hidden to those in which it remained visible.

We ran the following specification

$$\text{AbsRank}_{it} = \alpha + \beta \cdot \text{HiddenBadge}_{it} + \gamma_i + \delta_t + \varepsilon_{it}, \tag{3}$$

where the dependent variable AbsRank_{it} denotes the absolute rank of product i at time t. The indicator variable HiddenBadge_{it} equals 1 if product i displays a hidden badge at time t, and 0 otherwise. γ_i and δ_t denote product and date fixed effects, respectively. ε_{it} is the error term. The regression is estimated separately by absolute rank bin.

TABLE 6: Estimated effect of hidden badges on product rank across rank bins

	abs. rank						
	(1) bin: 1-48	(2) bin: 49-96	(3) bin: 97-144	(4) bin: 145-192	(5) bin: 193-240	(6) bin: 241-288	(7) bin: 289+
Hidden Badge	0.334***	3.916***	3.115***	2.369***	1.718***	1.236***	307.808***
	(0.088)	(0.095)	(0.074)	(0.070)	(0.065)	(0.055)	(24.201)
Intercept	30.743***	68.807***	117.693***	166.198***	214.631***	262.905***	783.564***
	(0.040)	(0.041)	(0.032)	(0.031)	(0.029)	(0.023)	(13.197)
Product FE	✓	✓	✓	✓	✓	✓	✓
Date FE	✓	✓	✓	✓	✓	✓	✓
Observations	2,489,815	2,158,302	1,368,084	961,199	691,775	518,372	3,042,439
Product Clusters	123,932	139,777	133,595	121,380	105,341	91,985	132,844
Date Clusters	41	41	41	41	41	41	41

Notes. This table reports estimates from separate regressions of absolute rank on a Hidden Badge indicator across seven rank bins. Data is at the product-date level for fast shipped products. All specifications include product and date fixed effects. Standard errors clustered by product and date. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 6 reports changes in product rankings following the removal of fast-shipping badges. For products initially shown in higher-ranking positions (pages 1–6; Columns 1–6), the changes in rank were minimal. The primary effects appear beyond page 6, starting at rank 289, where the average product experienced a rank increase of 308 positions, indicating a more substantial drop in visibility for hidden badge items.

It is important to note that the vast majority of consumer search activity—over 90% of traffic—occurs within the first six pages of search results.

Table 7 presents the same specification using the log of product ranks. The results are consistent with the levels specification: on page 1, the estimated effect is slightly negative, while on pages 2 through 6 (Columns 2–6), rank changes range from 0.5% to 5.7%. As with the level-based results, the most pronounced changes occur beyond page 6, where product ranks increase more substantially, by 23.9%.

TABLE 7: Estimated effect of hidden badges on product rank across rank bins

	log(rank)						
	(1) bin: 1-48	(2) bin: 49-96	(3) bin: 97-144	(4) bin: 145-192	(5) bin: 193-240	(6) bin: 241-288	(7) bin: 289+
Hidden Badge	-0.012***	0.057***	0.026***	0.014***	0.008***	0.005***	0.239***
	(0.004)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.009)
Intercept	3.346***	4.213***	4.762***	5.110***	5.367***	5.570***	6.363***
	(0.002)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.005)
Product FE	✓	✓	✓	✓	✓	✓	✓
Date FE	✓	✓	✓	✓	✓	✓	✓
Observations	2,489,815	2,158,302	1,368,084	961,199	691,775	518,372	3,042,439
Product Clusters	123,932	139,777	133,595	121,380	105,341	91,985	132,844
Date Clusters	41	41	41	41	41	41	41

Notes. This table reports estimates from separate regressions of log rank on a Hidden Badge indicator across seven rank bins. Data is at the product-date level for fast shipped products. All specifications include product and date fixed effects. Standard errors clustered by product and date. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Thus, given that the scores of the fast-shipped products were so much higher than the non-fast-

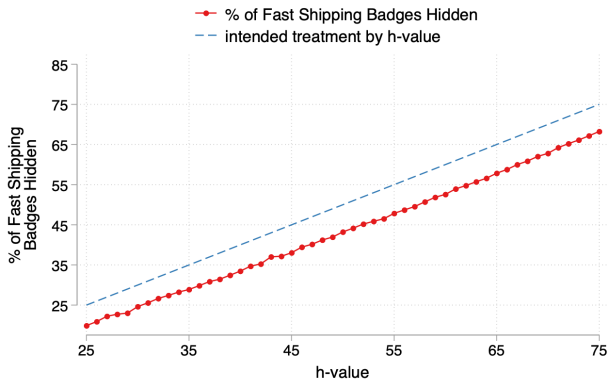
shipped ones, there were minimal changes in the rankings. Therefore, we could attribute the main treatment effect to the removal of the badges. However, to alleviate any remaining concerns, in Appendix G.2, we re-run all the main specifications of the experimental analysis we describe next controlling for the consumer-level ranking changes. For the remainder of the main part of the paper, we will treat the treatment effect as mainly the effect of removing badges because our results are robust to controlling for the rankings too. Next, we validate that the experiment ran correctly, explain the changes in choice architecture, and present the main experimental results.

5.1. Experiment validation

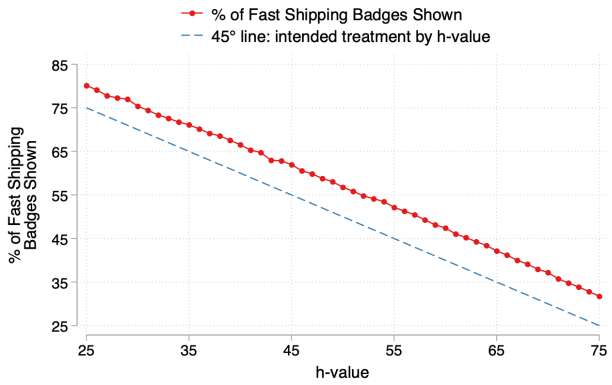
We validate that the experiment was ran correctly as follows. First, we check whether the h-values truly correspond to the h% of hidden badges on the pages. Figure 12 shows the actual percent of badges hidden and shown by h-value compared to the control assignment.

Figure 12: Treatment validation

(a) % of Fast Shipping Badges hidden by h-value



(b) % of Fast Shipping Badges shown by h-value



Notes. This figure validates that the experiment was ran correctly. Figure 12b shows the fraction of Fast Shipping Badges that were shown in the treatment condition compared to the control condition. For example, at $h = 25$, treatment condition had 77% of the eligible badges shown compared to the control group. In Figure 12a, we mirror the result and show the % of badges that were hidden, e.g., 23% at $h = 25$. The blue dashed 45-degree lines correspond to perfect treatment where exactly h% of badges would be hidden (Figure 12a) or $100 - h$% would be shown (Figure 12b). The actual treatment (red dotted line) shows that the treatment was not exact but still very close to perfect.

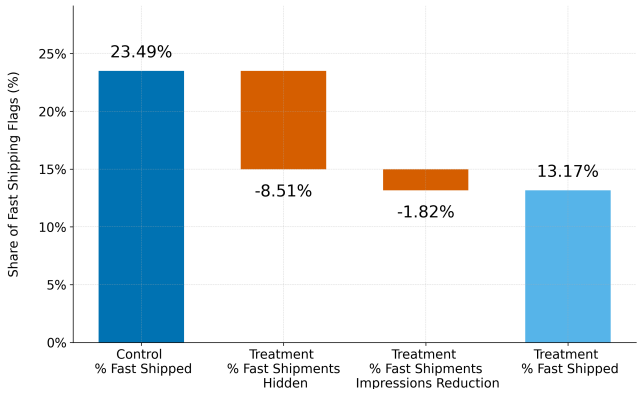
5.2. Changes in choice architecture

Next, we describe the changes in the choice architecture that resulted from the experimental intervention.

Fast Shipping Badges. As a result of the experiment, the number of fast shipping badges displayed to consumers changed significantly. Figure 13 shows that 23.49% of products in the control group were labeled with fast shipping badges. In contrast, only 13.17% of products in the treatment group were badged—representing a reduction of nearly 43%. Of this reduction, 8.51 percentage points can

be attributed to the direct removal of badges, while only 1.82 percentage points are due to changes in product ranking.

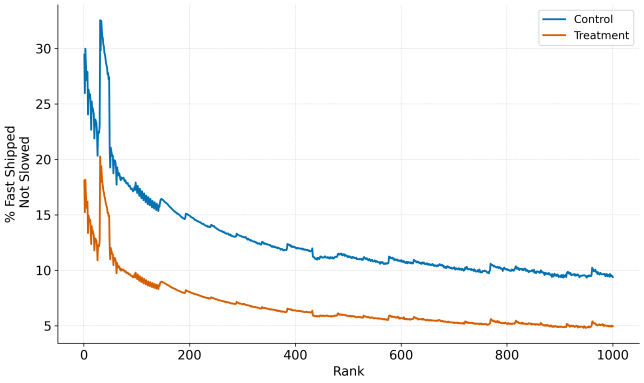
FIGURE 13: Share of fast shipped products between control and treatment groups



Notes: This figure displays the share of Fast Shipping badges shown in the control and treatment groups. We report both the overall share as well as disaggregated shares by two dimensions: (i) whether the Fast Shipping badge was intentionally hidden as part of the treatment condition, and (ii) whether the product experienced a reduction in impressions.

This pattern is also consistent across product rank positions. Figure 14 shows that the share of not-slowed, fast-shipped products decreases with product rank. The treatment group consistently featured a lower share of fast-shipping products across the ranking spectrum.

FIGURE 14: Share of fast shipped products between control and treatment groups by ranks

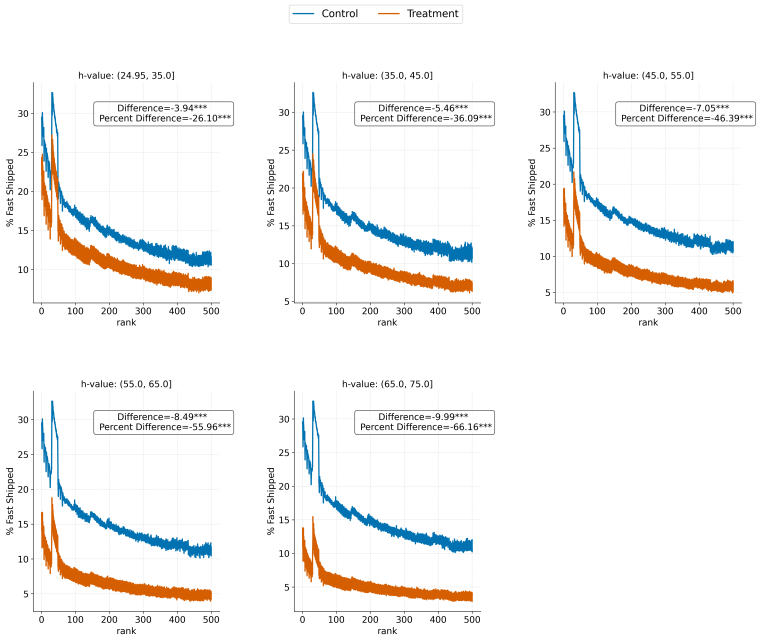


Notes: This figure shows the % of non-slowed fast shipped products by ranks across treatment groups.

Figure 15 plots the share of fast-shipped (i.e., non-hidden badge) products across product ranks and h-value bins. Each panel includes annotations reporting both the absolute difference in means and the corresponding percentage difference between treatment and control groups within each bin. The figure confirms that differences between groups are localized within h-value bins, consistent with the intended design and implementation of the experimental variation.

Prices. The price distribution between the treatment and control groups reveals no economically

FIGURE 15: Share of fast shipped products between control and treatment by rank and $h-$values

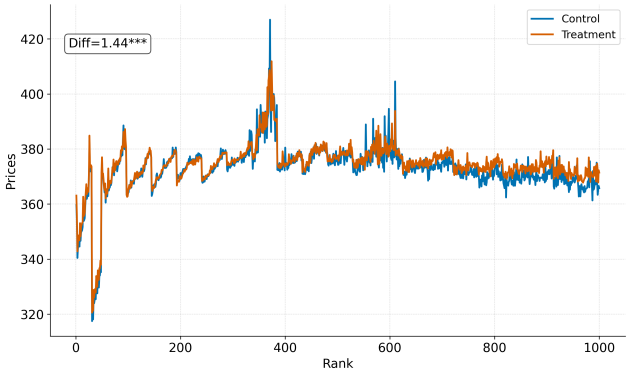


Notes: This figure shows the share of fast-shipped items by rank, separately for treatment and control groups, across five h-value bins. Each subplot corresponds to one bin and plots the mean fast-shipping rate among non-slowed, fast-eligible items over product ranks. Treatment and control groups are plotted separately. The annotations report the difference in means and percent difference between groups within each bin. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Text boxes summarize treatment effects within bins. Confidence intervals are not shown for clarity.

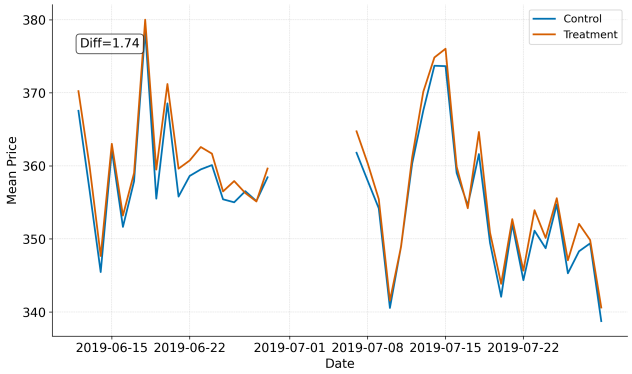
meaningful differences at the time of the experiment. As shown in Figure 16b, the average price difference is $1.74 and is not statistically significant. When examined by product rank, the difference becomes statistically significant but remains modest—averaging $1.44—as illustrated in Figure 16a. Notably, the bulk of this difference arises above rank 800, where consumer traffic is minimal. These patterns suggest that the experimental assignment did not generate meaningful price differences in the set of products that consumers were most likely to encounter.

FIGURE 16: Prices differences b/w control and treatment groups

(a) by rank



(b) by experiment date



Notes. This figure presents the price differences among treatment/control groups by rank and experiment date. Figure 16a plots the prices in control and treatment groups by rank. Figure 16b plots the prices in control and treatment groups by experiment date. The annotation summarizes the average price difference between groups using a t-test with unequal variances. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

In the Appendix Figure D8 , we show that there were no significant price differences between hidden badge and not-hidden-badge fast shipping products.

Delivery dates. The experiment removed fast shipping badges and, obviously, the 'Get-it-by' dates had to change too. We check with the differences in the treatment and control groups. Figure 17 shows that the consumers in treatment group on average saw the delivery dates 1.44 days later than the consumers in the control group.

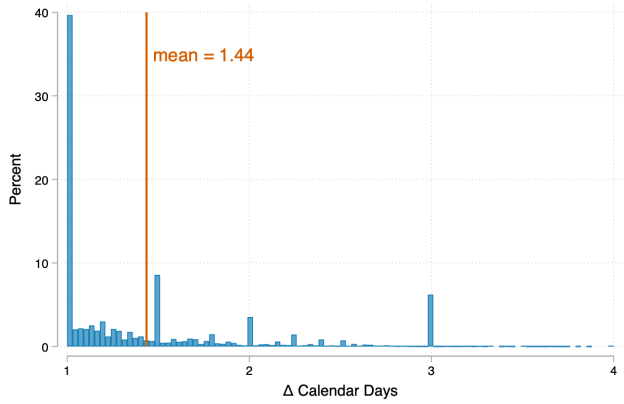
Differences in slowed and not slowed product outcomes. Next, we examine within-product outcome differences between instances where a product had a hidden badge versus when the badge was shown. To do so, we estimate the following specification, where the outcomes of interest are product-level clicks, orders, and conversion rates.

$$Y_{it} = \beta \cdot \text{HiddenBadge}_{it} + \alpha_i + \delta_t + \varepsilon_{it}, \tag{4}$$

where Y_{it} denotes one of the outcome variables. The variable HiddenBadge_{it} is an indicator equal to one if the product's fast-shipping badge was hidden on day t. The term α_i represents product fixed effects, while δ_t captures date fixed effects. The error term is denoted by ε_{it}.

Table 8 reports the results from a two-way fixed effects (TWFE) regression estimating the impact of

FIGURE 17: Delivery dates difference b/w treatment and control groups



Notes: This figure shows the difference in delivery days between treatment and control groups. The vertical line denotes the mean difference.

hidden fast shipping badges on key product-level outcomes. Each column presents a different dependent variable: raw clicks (Column 1), log(clicks) (Column 2), raw orders (Column 3), log(orders) (Column 4), and the conversion rate conditional on a click (Column 5). The main independent variable is a binary indicator for whether the product's fast shipping badge was hidden.

TABLE 8: Effect of hidden badge on clicks, orders, and conversion rates

	TWFE				
	(1) Clicks	(2) Log(Clicks)	(3) Orders	(4) Log(Orders)	(5) Conversion rate (cond-l on click)
Hidden Badge	-1.203***	-0.097***	-0.012***	-0.006***	-0.037***
	(0.084)	(0.002)	(0.000)	(0.000)	(0.010)
Intercept	5.522***	0.634***	0.059***	0.029***	1.265***
	(0.038)	(0.001)	(0.000)	(0.000)	(0.003)
Product FE	✓	✓	✓	✓	✓
Date FE	✓	✓	✓	✓	✓
Observations	11,538,971	11,538,971	11,538,971	11,538,971	4,086,043
Product Clusters	194,540	194,540	194,540	194,540	146,480
Date Clusters	41	41	41	41	41

Notes. This table reports two-way fixed effects (TWFE) estimates of the impact of hidden badges on user engagement metrics. All specifications include product and date fixed effects. Standard errors are clustered by product and date. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Across all specifications, hiding the badge is associated with statistically significant declines in consumer engagement and purchase outcomes. Specifically, products with hidden badges receive fewer clicks, fewer orders, and lower conversion rates. For example, in Columns (1) and (2), hiding the badge reduces average daily clicks by 1.2 or 9.7%, and in Columns (3) and (4), it lowers orders by 0.012 or 0.6% on average. The effect on conversion rate, conditional on a click, is also negative and significant, -3.7% (Column 5).

All specifications include product and date fixed effects, and standard errors are clustered at both

the product and date levels. These results suggest that the visibility of fast shipping badges plays a meaningful role in shaping consumer behavior on the platform. In the next subsection, we investigate the causal impact of reducing fast shipping badges using results from the randomized experiment.

5.3. Experimental Results

Since this is a randomized experiment, we estimate the following OLS specification to examine how being assigned to the treatment group affects delivery and purchase outcomes:

$$y_i = \alpha + \beta \cdot \text{treated}_i + \varepsilon_i, \tag{5}$$

where y_i denotes the outcome of interest for consumer i; treated_i is an indicator equal to 1 if the consumer was assigned to the treatment group and 0 otherwise; and ε_i is the error term.

Given that the treatment involved removing 25% to 75% of fast shipping badges, the coefficient β can be interpreted as the average effect of removing approximately 50% of the badges from the ranking pages. In subsequent analysis, we further explore treatment heterogeneity by the assigned h-value.

Overall Outcomes. The results in this subsection include all product categories and all experimental consumers who visited ranking pages on **desktop devices**.[22]

We analyze data from over 45 million consumers who participated in the experiment and estimate how treatment assignment affected purchasing behavior. Table 9 shows that consumers in the treatment group—those exposed to reduced fast shipping—made fewer purchases: the number of orders decreased by 0.5% (Column 3), revenue decreased by 1.7% (Column 4), post-return revenue decreased by 1.7% (Column 5), and profits declined by 1.2% (Column 6). However, we find no statistically significant differences in the number of clicks or the probability of clicking between the treatment and control groups (Columns 1 and 2).

Table 9: Overall treatment outcomes

	OLS					
	(1) clicks	(2) log(clicks)	(3) log(# orders)	(4) log(revenue)	(5) log(net revenue)	(6) log(profit)
Treated	-0.000	0.000	-0.005***	-0.017***	-0.017***	-0.012***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Intercept	0.505***	0.444***	0.085***	0.336***	0.324***	
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	
Observations	45,561,809	45,561,809	45,561,809	45,561,809	45,561,809	45,561,809

Notes. This table reports the output from the estimation of Equation 5 on **all experimental desktop consumers**. Data is at the consumer-level. Columns (1)-(6) use OLS specification results. Robust standard errors in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

[22] We exclude consumers who accessed the platform exclusively through the mobile app or tablets. However, we still capture cross-device effects, as individual consumers may have used multiple devices.

Thus, consumers who saw fewer fast delivery promises appeared to click on products at similar rates as those in the control group, but ultimately purchased less. As shown in Table 8, conversion rates for the same products were lower when the fast shipping badge was hidden, suggesting that consumers place value on the badge itself—even when the underlying product remain unchanged.

To investigate heterogeneity based on the share of products whose badges were hidden, we estimate the following specification:

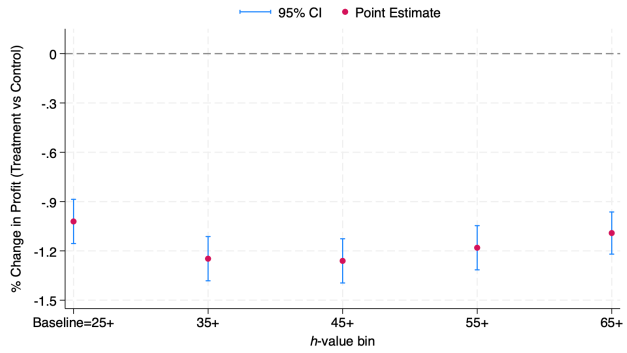
$$y_i = \alpha + \beta_1 \cdot \text{treated}_i + \beta_2 \cdot h\text{-bin}_i + \beta_3 \cdot (\text{treated}_i \times h\text{-bin}_i) + \varepsilon_i, \tag{6}$$

where we construct the variable h-bin$_i$ by partitioning the continuous h-value into five discrete bins. These bins are defined as: 25–34, 35–44, 45–54, 55–64, and 65–75, corresponding to equal-width intervals designed to balance granularity with sufficient sample size in each group.

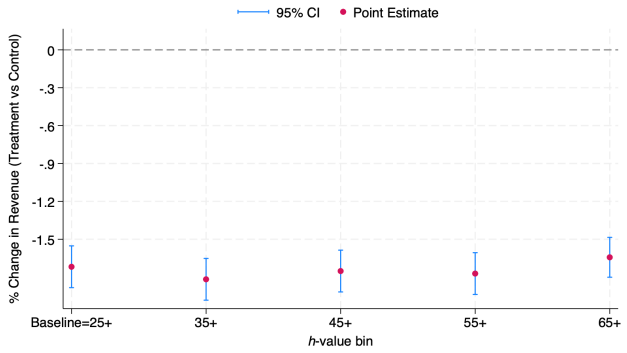
Figures 18a and 18b plot the estimated β_3 interaction coefficients from Equation 6, along with 95% confidence intervals. While the magnitudes of the effects vary slightly across bins, we observe consistently negative impacts on both profits and revenue. One might expect that removing a smaller share of fast shipping badges (i.e., lower h-values) would lead to smaller effects. However, as shown in Figure 15, badge removal was implemented consistently across the product ranking spectrum. Consequently, reduced ordering behavior persisted across bins, contributing to a uniform negative effect on revenue and profits.

FIGURE 18: Heterogeneous Treatment Effects on Profits and Revenue by h-value bins

(a) Profits



(b) Revenue



Notes. This figure shows the % change in profits between treatment and control groups by h-value bins. This figure shows the % change in revenues between treatment and control groups by h-value bins.

Fast Shipping and Distance Outcomes. Next, we examine how treatment assignment affects the probability of purchasing products eligible for fast shipping, as well as the distance those products travel. To do so, we estimate Equation 5 using relevant delivery outcomes.

Table 10 shows that the probability of purchasing a fast-shipped product decreased by 2.4% in the treatment group (Column 1). Additionally, consumers ordered products from more distant warehouses: the average shipping distance per product increased by nearly 6 miles (Column 2), or 1.3% (Column

3). When aggregating across all orders placed by a consumer, the total shipping distance increased by approximately 35 miles, or 2.2% (Column 5), reflecting the cumulative effect for consumers who purchased multiple items.

TABLE 10: Purchase outcomes for consumers

	Logit	*OLS*			
	(1) fast shipped	(2) mean distance	(3) log(mean distance)	(4) sum distance	(5) log(sum distance)
Treated	-0.024***	5.972***	0.013***	35.845***	0.022***
	(0.003)	(1.034)	(0.002)	(4.115)	(0.002)
Intercept	-0.656***	922.888***	6.389***	1720.701***	6.790***
	(0.003)	(0.868)	(0.002)	(3.454)	(0.002)
Observations	2,006,737	2,006,737	2,006,737	2,006,737	2,006,737

Notes. This table reports the output from the estimation of Equation 5 on **purchasing experimental consumers**. Data is at the consumer-level. Column (1) reports Logit results, the rest of the columns report OLS specification results. Robust standard errors in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 11 presents an extended analysis of shipping distances in which we incorporate not only the distance from the warehouse to the consumer, but also the distance from the seller to the Castlegate warehouse and from Castlegate to the consumer.[23]

TABLE 11: Alternative distance measures: upper bound on Castlegate distances

	OLS			
	(1) log(distance)	(2) log(distance+CG)	(3) log(distance+CG):min	(4) log(distance+CG):max
Treated	0.022***	0.012***	0.022***	0.008***
	(0.002)	(0.002)	(0.002)	(0.002)
Intercept	6.792***	7.058***	6.792***	7.203***
	(0.002)	(0.002)	(0.002)	(0.002)
Observations	2,006,737	2,006,737	2,006,737	2,006,737

Notes. This table reports the output from the estimation of Equation 5 on **purchasing experimental consumers**. Data is at the consumer-level. Columns (1)-(4) use OLS specification results. Robust standard errors in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

For comparison, Column 1 replicates the result from Column 5 of Table 10. In Column 2, we compute the average distance from the seller's multiple warehouses to the Castlegate warehouse from which the order was shipped. Column 3 reports results using the minimum distance between the seller's nearest warehouse and the corresponding Castlegate location, while Column 4 uses the maximum distance (i.e., from the farthest seller warehouse to Castlegate). Across all specifications, the treatment effect remains robust, confirming that our findings are not sensitive to alternative ways of measuring inbound shipping distances.

In some sense, these estimates represent lower bounds on the increase in shipping distances, as multiple products are typically consolidated and shipped together in a single truck to Castlegate warehouses. To

[23]We thank the anonymous reviewers for highlighting this point.

provide a more accurate measure, we obtained the platform's internally calculated shipping distances, which incorporate consolidation effects and routing optimizations.

Table 12 presents these results. The treatment effects on total miles traveled (aggregated across all orders per customer) and average miles per order are either similar to or larger than the earlier estimates. Specifically, we observe distance increases ranging from 1.3% to 3.6% (16 miles to 279 miles), reinforcing the conclusion that removing fast shipping badges leads to less efficient fulfillment logistics.

TABLE 12: Alternative distance measures: actual Castlegate distances

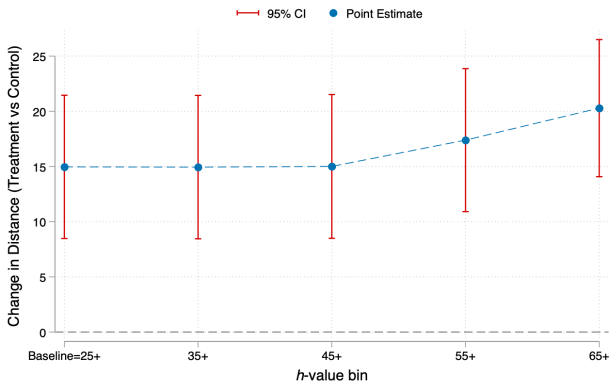
	OLS			
	(1) total miles	(2) log(total miles)	(3) avg miles	(4) log(avg miles)
Treated	279.088***	0.036***	16.808***	0.013***
	(14.818)	(0.002)	(1.556)	(0.001)
Intercept	6098.677***	8.320***	3042.645***	7.878***
	(12.438)	(0.001)	(1.306)	(0.001)
Observations	2,006,737	2,006,737	2,006,737	2,006,737

Notes. This table reports the output from the estimation of Equation 5 on **purchasing experimental consumers**. Data is at the consumer-level. Robust standard errors in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

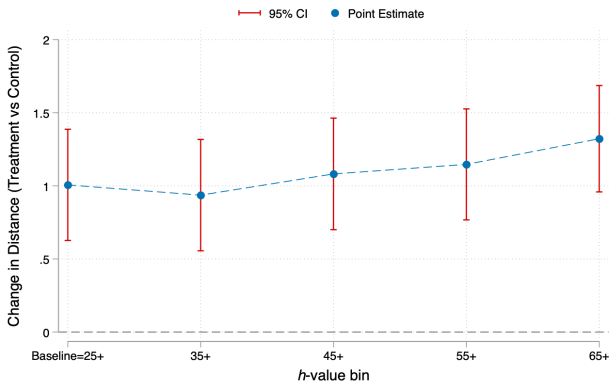
When we examine the heterogeneity of the results, we find that shipping distance increases with the share of products for which fast shipping badges are hidden—that is, with higher h-values. Even when we log-transform the outcome and smooth the estimates, the effect remains more pronounced for h-values above 45 compared to lower h-value bins. This pattern reinforces the conclusion that greater suppression of fast shipping information leads to less efficient fulfillment logistics. See Figure 19.

FIGURE 19: Distance outcomes by h-value bins

(a) Distance



(b) Log(Distance)

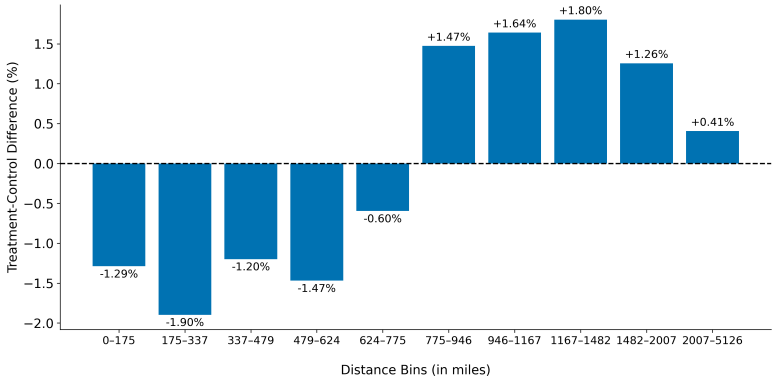


Notes. This figure shows heterogeneous distance outcomes by h-value bins. Figure 19a shows the distance changes in treatment versus control by $h-$value bins. Figure 19b shows the log distance changes in treatment versus control by $h-$value bins.

Figure 20 illustrates that the observed increase in shipping distances is driven by consumers ordering products from warehouses located farther away. The x-axis denotes distance bins (in miles), while the

y-axis shows the percentage difference in purchases between the treatment and control groups. The figure reveals significant negative effects on the sales of products stored closer to the consumer, alongside increased purchases of products from more distant warehouses. This shift in sourcing patterns contributes to the overall rise in shipping distances under the treatment condition.

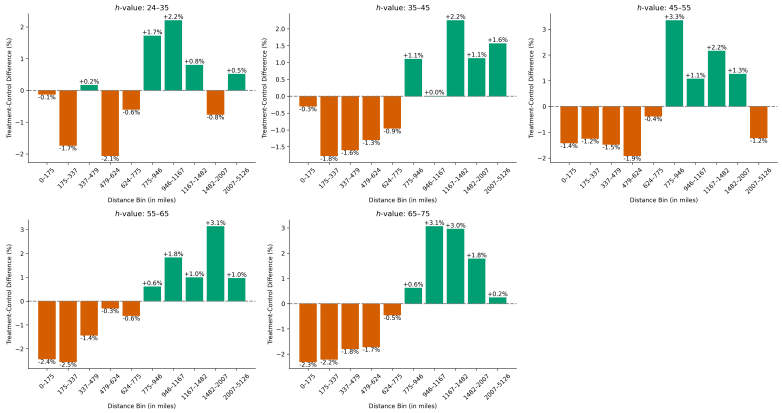
FIGURE 20: % change in orders by shipping distance bin



Notes: This figure shows the % difference in treatment group orders compared to control group orders by shipping distances binned into ten groups.

Figure 21 presents the same analysis, disaggregated by h-value bins. The effects are directionally similar across all bins but appear slightly more pronounced in the bins where a greater share of fast shipping badges was hidden.

FIGURE 21: % change in orders by shipping distance bin and by h-values



Notes: This figure shows the % difference in treatment group orders compared to control group orders by shipping distances binned into ten groups and by h-values.

Shipping and Wholesale Costs. We obtained actual shipping and product cost data for items sold on the platform. Table 13 shows that average shipping costs increased by 3.2% in the treatment group. Additionally, the platform experienced a 2.1% increase in product costs (Column 2), suggesting a loss of access to more favorable wholesale pricing.

As discussed in Section 3, larger and more successful sellers are more likely to participate in Castlegate. These sellers often offer wholesale discounts to the platform due to economies of scale. Thus, the removal of fast shipping badges may have shifted sales away from these preferred sellers.

TABLE 13: Cost outcomes by treatment group

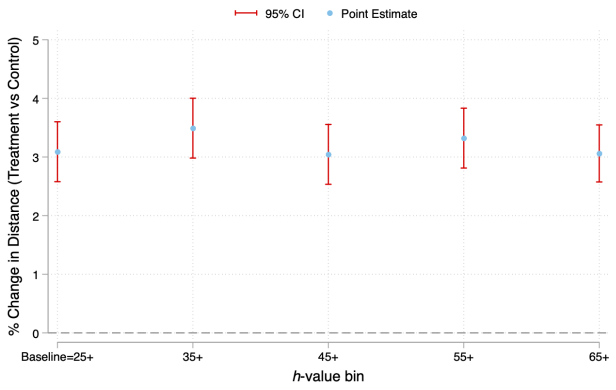
	OLS	
	(1)	(2)
	log(shipping cost)	log(product cost)
Treated	0.032***	0.021***
	(0.001)	(0.002)
Intercept	3.060***	4.273***
	(0.001)	(0.001)
Observations	3,931,011	3,931,011

Notes. This table reports the output from the estimation of equation 5 on **purchasing experimental consumers**. Data is at the consumer-purchased product-level. Robust standard errors in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

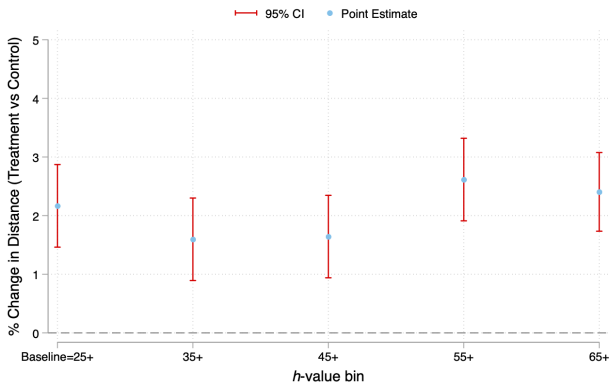
Figure 22 shows that the increase in shipping costs is consistently observed across all h-value bins, with no sharp discontinuities, indicating a robust relationship between badge removal intensity and cost increases.

FIGURE 22: Cost outcomes by h-value bins

(a) Shipping Costs



(b) Product Costs



Notes. This figure shows heterogeneous cost outcomes by h-value bins. Figure 22a shows the shipping cost changes in treatment versus control by $h-$value bins. Figure 22b shows the product cost changes in treatment versus control by $h-$value bins.

While a 3.2% increase in shipping costs may appear modest, the financial implications are substantial. For instance, in 2020, Wayfair reported approximately $2 billion in shipping expenses.[24] A 3.2% increase would translate to roughly $64 million in additional costs—an economically significant amount.

Product Returns. We use product returns as a proxy for customer satisfaction with their purchases. We find that the probability of returning a product increased by 7.6% in the treatment group, while the

[24] See Wayfair's 2020 Annual Report.

number of products returned rose marginally by 0.8%.

To better understand which types of products were returned, we estimate heterogeneous effects by shipping speed. Table 14 shows that the increase in returns is primarily driven by non-fast-shipped products, whereas return rates for fast-shipped products do not differ significantly across treatment and control groups.

TABLE 14: Impact on product returns by product type

	Logit	*OLS*	*Logit*		
	(1) returned	(2) products returned	(3) all	(4) non-fast shipped	(5) fast shipped
Treated	0.076***	0.008***	0.031***	0.029***	0.020
	(0.007)	(0.001)	(0.005)	(0.006)	(0.012)
Intercept	-3.075***	0.064***	-2.903***	-2.835***	-3.132***
	(0.006)	(0.001)	(0.003)	(0.003)	(0.006)
Observations	2,006,737	2,006,737	3,931,011	2,951,807	979,204

Notes. This table reports the output from the estimation of equation **??** on **purchasing experimental consumers**. Data for Columns 1-2 are at the purchasing consumer level, and data for Columns 3-5 are at the purchasing consumer-product-level. Robust standard errors in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 15 shows that, among returned products, those in the treatment group traveled 1.8% to 2.7% more miles compared to the control group. In aggregate terms, this translates into an estimated additional 32 to 64 million miles traveled due to increased product returns in the treatment condition.

TABLE 15: Changes in return distances

	OLS	
	(1) log(mean distance)	(2) log(sum distance)
Treated	0.018**	0.027**
	(0.008)	(0.011)
Intercept	6.492***	7.263***
	(0.007)	(0.009)
Observations	93,262	93,262

Notes. This table reports the output from the estimation of equation 5 on **product returning experimental consumers**. Data is at the consumer-purchased product-level. Robust standard errors in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Overall, the observed increase in product returns may be driven by the fact that products previously labeled with fast delivery promises historically performed better than others in terms of return and incident rates. When demand shifted away from these products due to badge removal, it likely moved toward alternatives with higher historical return rates, contributing to the overall increase in returns.

Consolidation outcomes. We define shipping efficiency in terms of both shipping costs and consolidation outcomes. While shipping cost estimates already capture consolidation effects—since

FedEx and UPS invoices would reflect lower rates for consolidated shipments—we further disentangle the consolidation component using FedEx and UPS tracking data. Specifically, we calculate: (i) the number of other orders delivered alongside a given order on the same truck, and (ii) the number of trucks required to fulfill a given volume of orders.

Table 16 shows that consolidation decreased by 2.4% in the treatment condition (Column 1). Additionally, 0.5% more trucks were required to deliver the same number of orders, which translates to approximately one additional truck per delivery batch (Columns 3 and 4).

TABLE 16: Effect of treatment on consolidation and truck usage

	OLS			
	(1) log(#consolidated)	(2) #consolidated	(3) log(#trucks)	(4) #trucks
Treated	-0.024***	-6.355***	0.005***	0.012***
	(0.003)	(1.091)	(0.001)	(0.002)
Intercept	5.687***	440.276***	0.804***	1.301***
	(0.002)	(0.947)	(0.001)	(0.002)
Observations	700,914	700,914	700,914	700,914

Notes. This table reports OLS estimates of the effect of treatment on consolidation and truck usage outcomes. The dependent variables are the log and level of the number of consolidated shipments (columns 1–2) and the log and level of the number of trucks used (columns 3–4). Data is at the consumer-order level. The reason the number is smaller is because we only have a random sample of the orders for this analysis. Standard errors are robust and reported in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

To summarize, the first experiment demonstrates that, given the current logistics infrastructure, removing fast shipping promises leads to longer shipping distances, higher shipping costs, and reduced order consolidation. It also increases product return rates, further straining the supply chain.

At this stage—when distribution centers have already been built—these infrastructure investments represent sunk costs. As a result, removing fast shipping badges is unlikely to improve shipping efficiency. In fact, when considering the full costs of infrastructure, these estimates represent a lower bound on the true increase in shipping costs, as the platform also incurs substantial expenditures on warehouse construction and ongoing maintenance. For reference, the construction of 90,000 square meters of warehouse space is estimated to cost approximately €110 million,[25] and annual maintenance costs are estimated at $2.9 million.[26]

Figure 23 shows that, over time, more sellers have joined the Castlegate network and have increasingly utilized multiple Castlegate warehouses (orange line). In contrast, non-Castlegate sellers have not significantly expanded their own warehouse infrastructure. Figure 23b further illustrates a decline in the number of child suppliers (i.e., warehouses) per parent supplier, suggesting growing reliance on Castlegate's centralized logistics system.

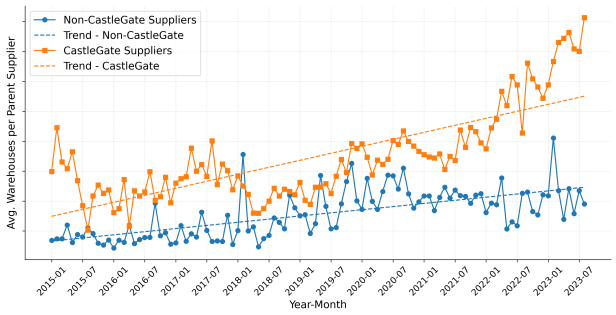
One could argue that, in the absence of Castlegate warehouses, third-party sellers may have built

[25] See this article on en.beyond-ecommerce.com for more details

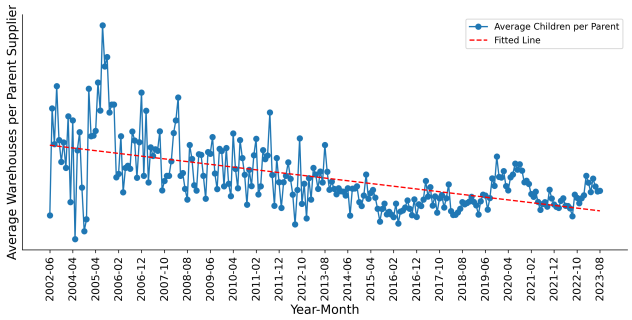
[26] See this article on locusrobotics.com for more details.

FIGURE 23: Number of Warehouses per Parent Supplier by Castlegate status and over time

(a) Castlegate Expansion Over Time



(b) Avg. # of Warehouses per Parent Supplier



Notes: This figure shows the number of warehouses per parent supplier by Castlegate status over time (Figure 23a) and overall over time (Figure 23b).

similar infrastructure themselves, in which case the costs might have netted out. However, given the existing demand for fast shipping, it is reasonable to assume that such infrastructure would have been developed regardless of whether it was platform- or seller-led. Therefore, we interpret our estimates as a lower bound on the increase in shipping costs resulting from the removal of fast delivery promises, since the underlying logistics demand would have persisted even without Castlegate.

Assuming an average retail order travels 900 miles via ground shipping by diesel truck and weighs approximately 5 pounds (0.0025 tons), the baseline carbon emissions are estimated using the GHG Protocol factor of 161.8 gCO_2e[27] per ton-mile, yielding approximately 0.364 kg CO_2e per order[28]. If the shipping distance increases by 6 to 35 miles, emissions rise to 0.366–0.379 kg CO_2e per order. This represents an absolute increase of 2–15 grams CO_2e, or a relative increase of approximately 0.55% to 4.12% per order.[29] While small on a per-order basis, these increases can accumulate significantly at scale across large retail networks.

Given this context, removing fast shipping promises does not appear to be an effective strategy for improving efficiency. We now turn to an alternative approach: algorithmic re-ranking of products.

6. Field Experiment #2: Algorithmic Re-Ranking for Shipping Efficiency

Description of the Experiment. Firms facing high shipping costs typically adopt a combination of strategies, such as optimizing logistics routes, encouraging customers to defer delivery through behavioral nudges (e.g., "low carbon" options), or avoiding fast-shipping commitments altogether—though the latter may come at the expense of competitive positioning. We propose an alternative approach: algorithmic

[27] See the Greenhouse Gas Protocol files.

[28] Baseline emissions are calculated as follows: $900 \times 0.0025 = 2.25$ ton-miles, and $2.25 \times 161.8 = 364.05$ grams of CO_2e. Converting to kilograms: $364.05 \div 1000 = 0.364$ kg CO_2e per order.

[29] Percent increase in emissions is calculated as $\frac{0.366-0.364}{0.364} \times 100 \approx 0.55\%$ for a 6-mile increase and $\frac{0.379-0.364}{0.364} \times 100 \approx 4.12\%$ for a 35-mile increase. These correspond to the emissions rising from 0.364 kg CO_2e (at 900 miles) to 0.366–0.379 kg CO_2e (at 906–935 miles).

nudging that reallocates visibility toward products that are less costly to ship, while continuing to leverage existing warehouse capacity. Rather than relying solely on product popularity, the platform modifies its objective function to explicitly incorporate shipping cost considerations. Importantly, we show that merely substituting the products without re-ranking is insufficient; effective implementation requires additional algorithmic adjustments that amplify the visibility of cost-efficient products. To test this approach, we analyze a field experiment conducted on the platform.

Treatment Assignment. The experiment was conducted over a 12-day period in July 2022, beginning shortly after the Fourth of July holiday and had 4.2 million consumers. Consumers were randomized into two treatment arms. In the first treatment—*Substitute*—high-shipping-cost products were replaced with cheaper-to-ship substitutes (conditional on comparable conversion rates), but the original ranking order was preserved. In the second treatment—*Substitute+Re-Rank*—the same substitution logic was applied, but the resulting product list was re-ranked to prioritize items with lower shipping costs.

Randomization was performed at the consumer level, ensuring that all devices or browsers associated with a given consumer were consistently assigned to a single treatment condition. As in the earlier experiment, the platform cannot account for consumers who delete cookies or otherwise reset their identifiers, but we assume that such behavior is balanced across experimental groups.

The experiment took place under conditions that differed somewhat from those of the earlier intervention. While data privacy restrictions prevent us from disclosing the full details of the updated ranking algorithm, the platform had, by this time, transitioned to a more sophisticated Bayesian-based system. This algorithm adaptively updates product rankings based on observed user behavior—particularly clicks and purchases. Despite these advancements, the core ranking logic remained intact: products continued to be ranked primarily based on historical click-through and conversion rates.

Stable Unit Treatment Value Assumption (SUTVA): Rankings. To assess potential violations of the Stable Unit Treatment Value Assumption (SUTVA)—specifically, whether user interactions in one treatment arm could influence outcomes in the other—we examine both click and purchase data, as well as the posterior distributions used by the platform's Bayesian ranking algorithm. This algorithm updates product rankings over time using a normal posterior distribution, fully characterized by its mean and variance. We leverage panel data on these posterior parameters across products and treatment arms.

Given the relatively short duration of the experiment (12 days), we find no evidence that cross-treatment user behavior (e.g., clicks or purchases in one group) meaningfully influenced the posterior distributions or rankings in the opposite group. See Appendix F.3 for additional details. These findings support the validity of SUTVA in our setting.

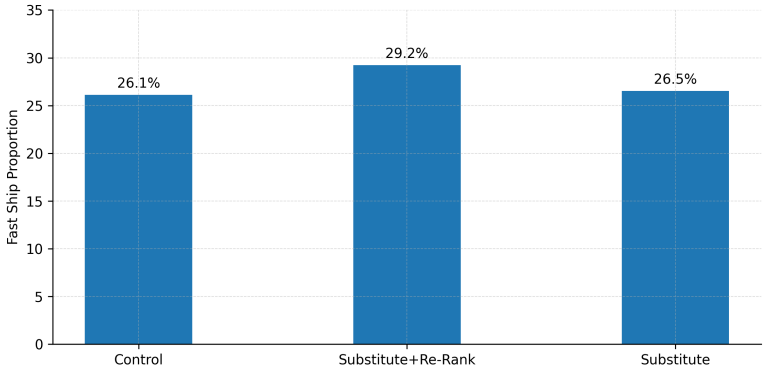
Stable Unit Treatment Value Assumption (SUTVA): Inventory. Owing to the short duration of the experiment, we also find no significant evidence of treatment-induced inventory effects or differences

in stockouts.[30]

Experiment validation. Since the goal of the experiment was to promote cheaper-to-ship items—and given our findings from the previous experiment—it is natural to expect that the intervention would increase the visibility of fast-shipped products, which are typically stored in centralized warehouses. We conduct an analysis to confirm this pattern.

Figure 24 shows that in the control condition, 26.1% of products displayed a fast shipping badge. In contrast, this share increased to 29.2% in the *Substitute+Re-Rank* condition, and to 26.5% in the *Substitute* condition. These results suggest that the re-ranking algorithm indeed shifted visibility toward products more likely to be fulfilled from centralized, fast-shipping infrastructure.

FIGURE 24: Prevalence of fast shipping badges



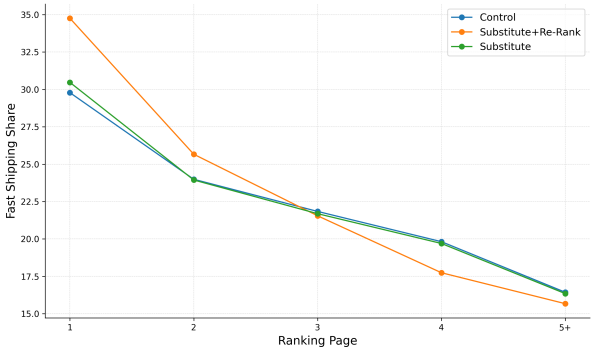
Notes: This figure shows the prevalence of fast shipping badges across experimental groups.

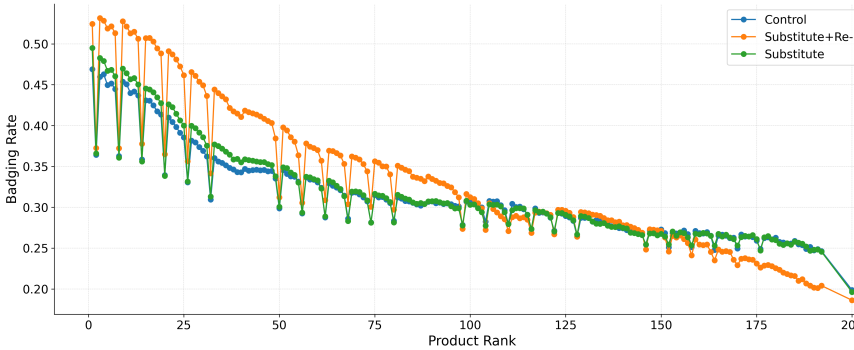
We also confirm that the changes were consistent across ranking pages and ranking positions. See Figure 25.

FIGURE 25: Prevalence of fast shipping badges

(a) by ranking pages

(b) by ranking positions





Notes: This figure shows the prevalence of fast shipping badges across experimental groups by page and product rank.

[30] For internal confidentiality reasons, we are unable to report the full table of results.

6.1. Experimental Results

We now turn to the core experimental results, focusing on how the *Substitute* and *Substitute+Re-Rank* conditions performed relative to the control group. Our primary outcomes of interest include consumer engagement, as measured by click behavior, and platform-level outcomes such as revenue and profit.

Table 17 shows that the probability of clicking was lower in the *Substitute+Re-Rank* condition. However, despite reduced clicking, both overall revenue and profit increased by approximately 0.2%. One possible interpretation is that consumers were able to locate desirable products more quickly in the re-ranked lists, requiring fewer clicks to complete a purchase. In contrast, we observe no statistically significant differences between the *Substitute* condition and the control group, suggesting that re-ranking—rather than substitution alone—was necessary to realize measurable gains in platform performance.

Table 17: Overall experimental outcomes

	Logit	*OLS*	
	(1) clicked	(2) log(revenue)	(3) log(profit)
Substitute+Re-rank	-0.017***	0.002*	0.002**
	(0.002)	(0.001)	(0.001)
Substitute	-0.003	0.000	0.001
	(0.002)	(0.001)	(0.001)
Intercept	0.264***	0.112***	—
	(0.002)	(0.001)	
Observations	4,243,941	4,243,941	4,243,941

Notes. This table reports the output from the estimation of equation 5 on **all experimental consumers**. Column (1) uses Logit specification and Columns (2)-(3) use OLS specification results. Robust standard errors in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

We now analyze the impact of the experimental treatments on shipping-related outcomes. Table 18 presents the results.

While the *Substitute* condition did not produce statistically significant changes in shipping outcomes, the *Substitute+Re-Rank* condition led to meaningful improvements. Specifically, average shipping distances decreased by 4.6% in relative terms, and the share of fast deliveries declined by 1.8%. These changes were primarily driven by an increase in the share of orders that were eventually shipped more quickly—by 6.1% and 4.4%—and a reduction in the number of orders traveling over 1,500 miles.

Although the *Substitute* condition exhibits effects in the same direction, they are smaller in magnitude and not statistically significant. This suggests that re-ranking plays a critical role in realizing shipping efficiency gains.

These improvements in shipping efficiency appear to be driven by a shift in consumer demand

TABLE 18: Experimental impact on order distances

	OLS			*Logit*				
	(1) distance	(2) log(distance)	(3) log(delivery)	(4) < 250 mi	(5) $250 - 1500$ mi	(6) $1500 - 2000$ mi	(7) $2000 - 2500$ mi	(8) > 2500 mi
Substitute+Re-rank	-34.800***	-0.046***	-0.018***	0.061***	0.044***	-0.123***	-0.094***	-0.157***
	(5.299)	(0.008)	(0.004)	(0.018)	(0.015)	(0.027)	(0.024)	(0.051)
Substitute	-3.673	0.004	-0.005	0.006	0.013	-0.030	-0.024	0.008
	(5.321)	(0.008)	(0.004)	(0.018)	(0.015)	(0.026)	(0.024)	(0.049)
Intercept	877.154***	6.298***	1.638***	-1.333***	0.316***	-2.355***	-2.158***	-3.761***
	(3.762)	(0.006)	(0.003)	(0.013)	(0.010)	(0.018)	(0.017)	(0.035)
Observations	112,972	112,972	112,972	112,972	112,972	112,972	112,972	112,972

Notes. This table reports the output from the estimation of equation 5 on **purchasing experimental consumers**. Columns (1)-(3) use OLS specification results and Columns (4)-(8) use Logit specification. Robust standard errors in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

toward Castlegate-fulfilled products. Table 19 shows that the number of orders fulfilled from Castlegate warehouses increased by 16.8% in the *Substitute+Re-Rank* condition, primarily driven by small parcel shipments. The *Substitute* condition also led to an 8% increase in Castlegate orders, though the effect is smaller in magnitude. These results suggest that algorithmic re-ranking effectively nudged consumers toward products that were already integrated into the platform's centralized logistics network.

Table 20 reports the impact of the experimental treatments on the logarithm of order-level shipping costs, estimated separately across four consumer segments defined by two dimensions: whether the ranking page was filtered and whether personalization was active.

In Column (1), which represents consumers viewing unfiltered and non-personalized pages, the *Substitute+Re-Rank* treatment significantly reduced shipping costs by 7.3%, while the *Substitute* treatment reduced them by 5.6%. In Column (2), where personalization was active but the ranking page remained unfiltered, neither treatment had a statistically significant effect, possibly due to small sample size. In Column (3), among consumers who saw filtered and personalized pages, the *Substitute+Re-Rank* condition led to a 5.1% reduction in shipping costs, while the *Substitute* condition had no measurable impact. Finally, in Column (4), among those with filtered but non-personalized views, both treatments significantly reduced shipping costs—by 5.3% for *Substitute+Re-Rank* and 4.5% for *Substitute*.

Taken together, these results show that the re-ranking algorithm consistently reduced shipping costs across most consumer types, especially when personalization was not active or when product filters were applied. The effects are strongest and most robust for the *Substitute+Re-Rank* treatment, reinforcing the importance of actively adjusting product visibility to favor cheaper-to-ship items.

Product Returns and Incidents. Table 21 reports the effects of the experimental treatments on product return rates and incident rates. While the *Substitute+Re-Rank* condition did not result in a meaningful reduction in return rates, it did significantly reduce incident rates by 11%, potentially due to shorter shipping distances and improved fulfillment logistics.

In contrast, the *Substitute* condition did not lead to a reduction in incidents but did significantly

TABLE 19: Experimental impact on Castlegate products

	Logit		
	(1) CG order	(2) CG order-Small Parcel	(3) CG order-Large Parcel
Substitute+Re-rank	0.168***	0.170***	0.090
	(0.016)	(0.016)	(0.095)
Substitute	0.080***	0.081***	0.067
	(0.016)	(0.017)	(0.095)
Intercept	-0.980***	-0.982***	-0.929***
	(0.012)	(0.012)	(0.067)
Observations	112,972	109,737	3,235

Notes. This table reports the output from the estimation of equation 5 on **purchasing experimental consumers**. Columns (1)-(3) use Logit specification results. Robust standard errors in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE 20: Order shipping costs by refinement and personalization outcomes

	log(order shipping costs)			
	(1) filtered=0, personalization=0	(2) filtered=0, personalization=1	(3) filtered=1, personalization=1	(4) filtered=1, personalization=0
Substitute+Re-rank	-0.073**	-0.013	-0.051*	-0.053**
	(0.031)	(0.035)	(0.027)	(0.026)
Substitute	-0.056*	-0.037	-0.001	-0.045*
	(0.031)	(0.036)	(0.027)	(0.026)
Intercept	3.076***	3.189***	3.271***	3.184***
	(0.022)	(0.025)	(0.019)	(0.018)
Observations	3,343	2,981	5,318	5,388

Notes. This table reports the output from the estimation of equation 5. Data is at the product-level because we account for the within-product shipping cost differences across consumers. Columns (1)-(4) use OLS specification results. Robust standard errors in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

lower product return rates relative to the control group. These results suggest that different components of the algorithmic intervention—substitution versus re-ranking—may differentially affect post-purchase outcomes.

TABLE 21: Experimental impact on product returns and incidents

	Logit	
	(1) returned	(2) incident
Substitute+Re-rank	-0.007	-0.110***
	(0.037)	(0.036)
Substitute	-0.072*	-0.051
	(0.038)	(0.036)
Intercept	-3.191***	-3.053***
	(0.027)	(0.025)
Observations	112,972	112,972

Notes. This table reports the output from the estimation of equation 5 on **purchasing experimental consumers**. Columns (1)-(4) use OLS specification results. Robust standard errors in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 22 shows that the *Substitute+Re-Rank* treatment significantly reduced the likelihood of returns by 13.7% in the unfiltered, non-personalized segment (Column 1). However, this effect is not observed in the other segments, where estimates are smaller and statistically insignificant. The *Substitute* condition does not produce statistically significant effects in any consumer group, though the direction of the coefficients is generally negative, which explains

In the incidents panel, the *Substitute+Re-Rank* treatment significantly reduced the likelihood of incidents by 29.7% in the unfiltered, non-personalized segment (Column 1). There are also modest but statistically insignificant reductions in incidents across the remaining segments. The *Substitute* treatment shows a smaller and marginally significant reduction in incidents (15.6%) only in Column 1.

Taken together, these results suggest that the *Substitute+Re-Rank* treatment had the strongest impact in contexts where product filtering and personalization were not active—likely because the algorithmic intervention had greater visibility and influence on consumer behavior in those cases.

Consolidation outcomes. Next, we show that despite the short-term experiment, we observe significant improvements in order consolidation rates. Table 24 shows that consolidation increased by 4.4%. However, we do not find significant improvements in the number of trucks that were used, possibly due to the fact that the experiment only took 12 days.

When we analyze the results by region, we find that all regions experienced an increase in the Castlegate share and a corresponding decrease in average shipping distance (Figures 26 and 27). However, realized shipping costs continue to depend heavily on the geographic distribution of Castlegate

TABLE 22: Experimental impact on returns by refinement and personalization

	Logit: returns			
	(1) filtered=0, personalization=0	(2) filtered=0, personalization=1	(3) filtered=1, personalization=1	(4) filtered=1, personalization=0
Substitute+Re-rank	-0.137*	-0.045	0.056	0.038
	(0.082)	(0.092)	(0.068)	(0.066)
Substitute	-0.072	-0.155	-0.049	-0.052
	(0.081)	(0.096)	(0.070)	(0.067)
Intercept	-3.122***	-3.170***	-3.252***	-3.191***
	(0.056)	(0.065)	(0.049)	(0.047)
Observations	23,535	18,559	34,891	35,987

Notes. This table reports the output from the estimation of equation 5 on **purchasing experimental consumers**. Columns (1)-(4) use Logit specification results. Robust standard errors in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE 23: Experimental impact on incidents by refinement and personalization

	Logit: incidents			
	(1) filtered=0, personalization=0	(2) filtered=0, personalization=1	(3) filtered=1, personalization=1	(4) filtered=1, personalization=0
Substitute+Re-rank	-0.297***	-0.075	-0.089	-0.050
	(0.086)	(0.093)	(0.063)	(0.061)
Substitute	-0.156*	-0.000	-0.050	-0.017
	(0.083)	(0.093)	(0.062)	(0.061)
Intercept	-3.144***	-3.183***	-2.995***	-2.993***
	(0.057)	(0.065)	(0.044)	(0.043)
Observations	23,535	18,559	34,891	35,987

Notes. This table reports the output from the estimation of equation **??** on **purchasing experimental consumers**. Columns (1)-(4) use Logit specification results. Robust standard errors in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

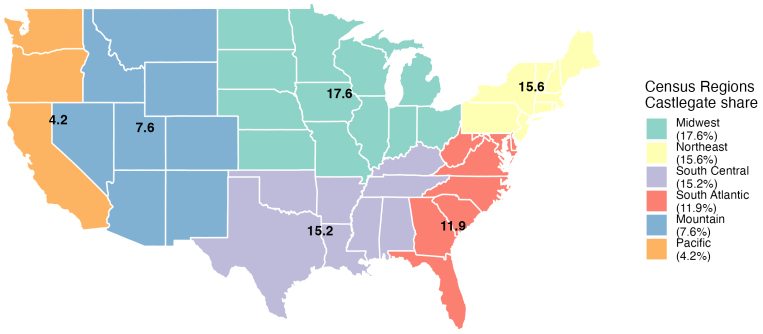
TABLE 24: Consolidation outcomes

	OLS		
	(1) log(#consolidated)	(2) #trucks	(3) #trucks
Substitute+Re-rank	0.044***	0.001	0.001
	(0.013)	(0.001)	(0.001)
Substitute	0.019	0.001	0.001
	(0.013)	(0.001)	(0.001)
Intercept	3.425***	0.687***	0.992***
	(0.009)	(0.001)	(0.001)
Observations	28,393	28,424	28,424

Notes. This table reports OLS estimates of the effect of treatment on consolidation and truck usage outcomes. The dependent variables are the log number of consolidated shipments (column 1) and the log and level of the number of trucks used (columns 2-3). Data is at the consumer-order level. Standard errors are robust and reported in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

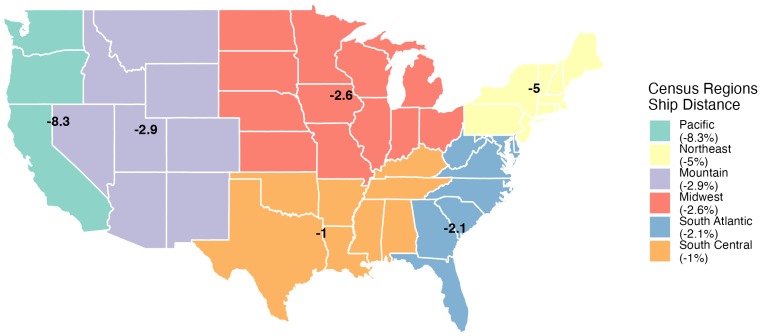
and Dropship warehouses. Regions that include states such as California and New York tend to benefit the most in terms of shipping costs, as they are better served by both Castlegate and Dropship warehouse infrastructure. This suggests that proximity to well-positioned fulfillment centers continues to play a key role in shaping the cost-efficiency of logistics across the country.

FIGURE 26: % change in Castlegate share across census regions



Notes: This figure shows the % change in Castlegate share across census regions.

FIGURE 27: % change in shipping distance across census regions



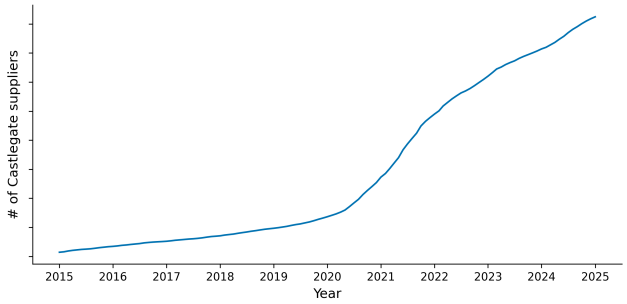
Notes: This figure shows the % change in shipping distance across census regions.

Long-Term Platform Outcomes. Following the conclusion of the experiment, the platform adopted the *Substitute+Re-Rank* condition as its default ranking policy. We examine the longer-term outcomes associated with this change in platform strategy. Figure 28a shows that the cumulative number of sellers participating in the Castlegate fulfillment program continued to increase steadily throughout 2022, suggesting sustained growth in adoption of centralized logistics infrastructure. In addition, we

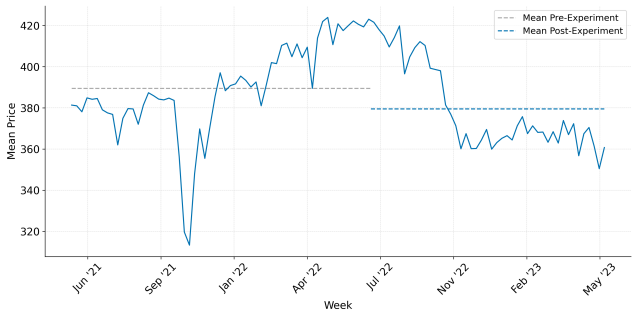
observe suggestive evidence of a decline in average prices on the platform during this period (Figure 28b), potentially reflecting increased competition for visibility among sellers under the new ranking algorithm. While these trends are descriptive, they point to possible equilibrium adjustments in both supply chain participation and seller pricing behavior following the re-ranking intervention.

FIGURE 28: Long-term warehouse and price changes

(a) Additions to the Castlegate warehouses over time



(b) Prices over time



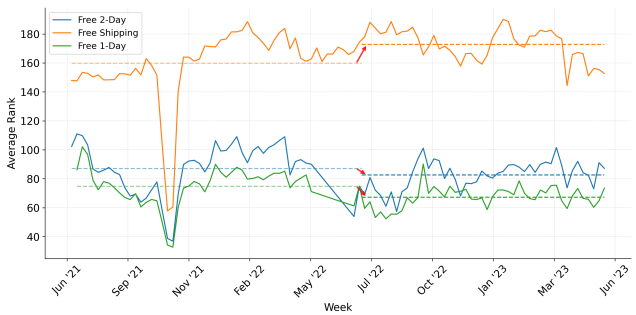
Notes. This figure shows the long-term platform warehouse and price outcomes. Figure 28a shows the cumulative number of Castlegate suppliers over time (y-axis is hidden for data anonymity purposes). Figure 28b shows the average prices on the ranking pages over time. Dashed lines on Figure 28b represent the pre- and post-second experiment mean prices.

When disaggregating products by shipping type, Figure 29a reveals that the absolute rankings of products eligible for free 1- and 2-day shipping declined modestly following the adoption of the re-ranking policy in 2022. In contrast, the rankings of standard free shipping products increased slightly over the same period, as indicated by the orange line. A parallel pattern emerges in pricing: products offering free fast shipping experienced a gradual decline in average prices, while prices of standard free shipping products rose marginally.

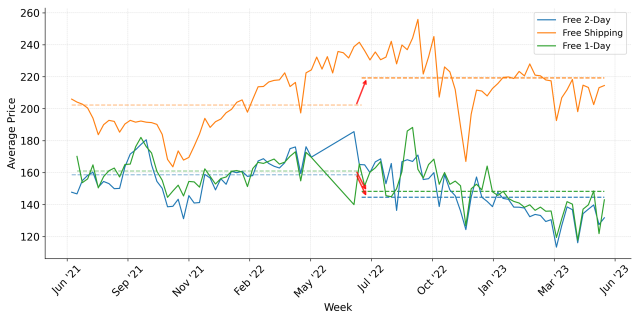
These trends suggest a reallocation of visibility and competitive pressure across shipping types under the new algorithm. Notably, both rankings and prices exhibit sharp short-term fluctuations around major promotional periods, such as the "hot deal" season and the Thanksgiving holiday, consistent with temporary shifts in platform-wide merchandising strategies.

FIGURE 29: Long-term rank and price changes by shipping types

(a) Absolute ranks by shipping types



(b) Prices by shipping types



Notes. This figure shows the long-term rank and price changes by shipping types on the platform. Figure 29a shows the absolute ranks of products by shipping types over time. Figure 29b shows the average prices by shipping types over time.

While we cannot causally attribute these longer-term trends to the experiment alone, the timing and direction of the changes are consistent with plausible post-experimental adjustments. Specifically, the adoption of the re-ranking policy may have incentivized greater participation in the Castlegate program, as sellers sought to maintain or improve visibility under the new ranking algorithm. The resulting increase in competition for higher-ranked positions may have exerted downward pressure on prices for fast-shipping products. In contrast, products relegated to lower-ranked positions—often reached only by a subset of more persistent consumers—may have experienced weaker competitive pressure, allowing for modest price increases in those segments.

Overall, the results of the second experiment indicate that re-ranking is essential for improving shipping outcomes. In the absence of re-ranking, adjustments to the product set yield limited effects on logistics metrics. By contrast, re-ranking products to favor geographic proximity leads to a reduction in average shipping distances, as consumers are more likely to purchase items located nearer to them. This reduction in distance does not increase return rates; in fact, return rates decline in non-filtered and non-personalized contexts. Incident rates also decrease, consistent with improved reliability in fulfillment from proximate warehouses. These patterns hold across all Census-defined regions, suggesting a broad-based effect of re-ranking on supply chain efficiency.

7. Conclusion

This paper examines whether and how online retail platforms can reduce shipping costs and improve shipment consolidation without sacrificing consumer outcomes. Motivated by concerns over the environmental and infrastructural impacts of fast shipping, we use two large-scale randomized experiments to provide causal evidence on the consequences of removing fast-delivery promises and the effectiveness of algorithmic interventions in product rankings.

First, we show that simply removing fast-delivery promises backfires: it leads to longer shipping distances, slower delivery times, higher shipping costs, and increased product return rates. Critically, removing fast shipping reduces the platform's ability to consolidate shipments, thereby increasing emissions and operational costs—contrary to regulatory hopes that curbing fast delivery would yield environmental benefits.

Second, we propose and evaluate an alternative strategy based on algorithmic choice architecture. We design ranking policies that balance purchase probability with shipping efficiency by systematically promoting cheaper-to-ship products. Algorithmic re-ranking significantly improves consolidation, reduces shipping distances and delivery times, lowers incident rates, and achieves shipping cost savings of up to 7.3%—all without harming conversion rates or product return rates.

Overall, our results suggest that platforms aiming to reduce the environmental and financial costs of e-commerce logistics should prioritize shaping consumer demand through algorithmic nudges rather

than blunt restrictions on delivery speed. More broadly, our findings highlight the potential of low-cost, demand-side interventions to complement traditional operational improvements. Future research could examine whether incorporating transparent environmental messaging or bundling strategies into ranking algorithms might further amplify these efficiency gains.

References

Andaz, Sohrab, Carson Eisenach, Dhruv Madeka, Kari Torkkola, Randy Jia, Dean Foster, and Sham Kakade, "Learning an inventory control policy with general inventory arrival dynamics," *arXiv preprint arXiv:2310.17168*, 2023. (Cited on page(s) 2)

Athey, Susan and Guido W Imbens, "The state of applied econometrics: Causality and policy evaluation," *Journal of Economic perspectives*, 2017, *31* (2), 3–32. (Cited on page(s) 23)

Bollinger, Bryan, Randi Kronthal-Sacco, and Levin Zhu, "Sustainable Product Demand and Price Elasticity," 2023. (Cited on page(s) 6)

Brecko, Kristina and Yewon Kim, "Voluntary Provision of Sustainability Claims: Evidence from Consumer Packaged Goods," Technical Report, Working paper 2023. (Cited on page(s) 6)

Cheng, Sikai, Amira Hijazi, Jeren Konak, Alan Erera, and Pascal Van Hentenryck, "SPOT: Spatio-Temporal Pattern Mining and Optimization for Load Consolidation in Freight Transportation Networks," *Transportation Science*, 2025, *59* (2), 345–360. (Cited on page(s) 6)

Cristea, Anca, David Hummels, Laura Puzzello, and Misak Avetisyan, "Trade and the greenhouse gas emissions from international freight transport," *Journal of environmental economics and management*, 2013, *65* (1), 153–173. (Cited on page(s) 6)

Dror, Moshe, Michael Ball, and Günther Zäpfel, "Shipment Consolidation: Who Pays for It and How Much?," *Management Science*, 2007, *53* (4), 632–645. (Cited on page(s) 5)

Duan, Huabo, Guanghan Song, Shen Qu, Xiaobin Dong, and Ming Xu, "Post-consumer packaging waste from express delivery in China," *Resources, Conservation and Recycling*, 2019, *144*, 137–143. (Cited on page(s) 6)

Juan, Angel A, Javier Faulin, Elena Pérez-Bernabeu, and Nicolas Jozefowiez, "Horizontal cooperation in vehicle routing problems with backhauling and environmental criteria," *Procedia-Social and Behavioral Sciences*, 2014, *111*, 1133–1141. (Cited on page(s) 6)

Kang, Peng, Guanghan Song, Ming Xu, Travis R Miller, Haikun Wang, Hui Zhang, Gang Liu, Ya Zhou, Junshu Ren, Ruoyu Zhong et al., "Low-carbon pathways for the booming express delivery sector in China," *Nature Communications*, 2021, *12* (1), 450. (Cited on page(s) 6)

Kellogg, Ryan and Mar Reguant, "Energy and environmental markets, industrial organization, and regulation," in "Handbook of Industrial Organization," Vol. 5, Elsevier, 2021, pp. 615–742. (Cited on page(s) 6)

Korganbekova, Malika and Cole Zuber, "Balancing user privacy and personalization," *Work in progress*, 2023, *6*. (Cited on page(s) 10, 12)

Lee, Yeonjoo and Karen Donohue, "Nudging Green-but-Slow Shipping Choices in Online Retail," *Available at SSRN*, 2024. (Cited on page(s) 6)

Loon, Patricia Van, Lieven Deketele, Joost Dewaele, Alan McKinnon, and Christine Rutherford, "A comparative analysis of carbon emissions from online retailing of fast moving consumer goods," *Journal of Cleaner Production*, 2015, *106*, 478–486. (Cited on page(s) 6)

Madeka, Dhruv, Kari Torkkola, Carson Eisenach, Anna Luo, Dean P Foster, and Sham M Kakade, "Deep inventory management," *arXiv preprint arXiv:2210.03137*, 2022. (Cited on page(s) 2)

Muñoz-Villamizar, Andrés, Josué C Velázquez-Martínez, Perla Haro, Ana Ferrer, and Roger Mariño, "The environmental impact of fast shipping ecommerce in inbound logistics operations: A case study in Mexico," *Journal of Cleaner Production*, 2021, *283*, 125400. (Cited on page(s) 6)

Proserpio, Davide, Ali Goli, Tyler Mangini, Ken Lau, and Daniela Yu, "The impact of sustainability programs on consumer purchase behavior: Evidence from Amazon," *International Journal of Research in Marketing*, 2025. (Cited on page(s) 6)

Russo, Daniel J, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen et al., "A tutorial on thompson sampling," *Foundations and Trends® in Machine Learning*, 2018, *11* (1), 1–96. (Cited on page(s) 10)

Sanders, Robert Evan, "Dynamic Pricing and Organic Waste Bans: A Study of Grocery Retailers' Incentives to Reduce Food Waste," *Available at SSRN 2994426*, 2020. (Cited on page(s) 6)

Sonntag, Danja R., Albert Schrotenboer, and Gudrun P. Kiesmüller, "Stochastic Inventory Routing with Time-Based Shipment Consolidation," *European Journal of Operational Research*, 2023, *306* (3), 1186–1201. (Cited on page(s) 5)

Thompson, William R, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, 1933, *25* (3/4), 285–294. (Cited on page(s) 10)

Tran, Nguyen Khoi and Jasmine Siu Lee Lam, "Effects of container ship speed on CO2 emission, cargo lead time and supply chain costs," *Research in Transportation Business & Management*, 2022, *43*, 100723. (Cited on page(s) 6)

Ursu, Raluca M, "The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions," *Marketing Science*, 2018, *37* (4), 530–552. (Cited on page(s) 17)

Wang, Caroline and Sherry He, "Do Judge a Product by Its Sustainability Label: Evidence from the Amazon Marketplace," *Available at SSRN*, 2024. (Cited on page(s) 6)

Wei, Lai, Roman Kapuscinski, and Stefanus Jasin, "Shipping Consolidation Across Two Warehouses with Delivery Deadline and Expedited Options for E-commerce and Omni-channel Retailers," *Manufacturing & Service Operations Management*, 2021, *23* (6), 1634–1650. (Cited on page(s) 5)

Xu, Zhou, Feng Li, and Zhi-Long Chen, "Shipment Consolidation with Multiple Shipping Methods Under Nonlinear Cost Structures," *Management Science*, 2024, *70* (5), 2823–2841. (Cited on page(s) 5)

Online Appendix

Table of Contents

A Data Cleaning **55**

B Retail Delivery Fees **56**

C Randomization Checks **57**

D Additional Descriptives: Supply side **59**

E Additional Descriptives: Refinement Behavior **61**

F Addressing potential Stable Unit Treatment Value Assignment **62**

F.1 Experiment #1: ranks 62

F.2 Experiment #1: inventory 64

F.3 Experiment #2: ranks 66

G Additional Experimental Results **67**

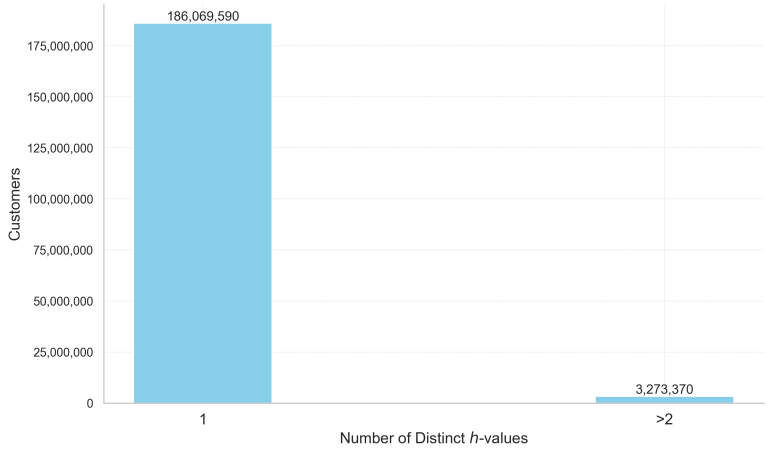
G.1 Robustness checks: Distance measures 68

G.2 Robustness checks: rankings 69

H Additional Experimental Results: Experiment # 2 **71**

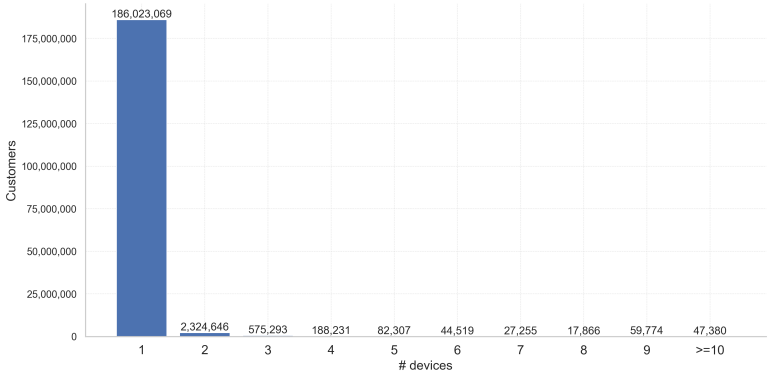
A. Data Cleaning

Figure A1: Consumer-Level Compliance with Treatment Intensity Assignments (return to p. **??**)



Notes: This figure checks the treatment intensity compliance at the consumer-level. x-axis shows the number of h-values a consumer was assigned to. Majority of consumers were assigned only one h-value, and a small portion had more than 2 h-values.

Figure A2: Distribution of Devices per Consumer



Notes: This figure shows the number of devices a consumer has. x-axis shows the number of devices, y-axis has the number of consumers. Majority of consumers have only one device, while 1.75% have more than one device.

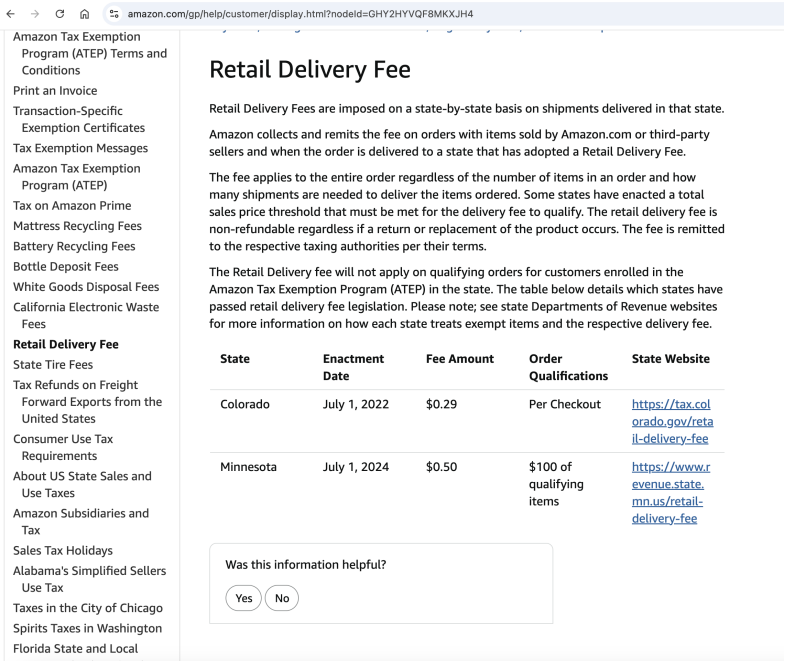
B. Retail Delivery Fees

Figure B3: Retail Delivery Fees in Minnesota and Colorado

	Fee	Applicability	Exemptions
State			
Minnesota	50 cents per delivery	Tangible property subject to sales tax or clothing of value >$100	Retailers with sales <$1 million
			Marketplace deliveries for sellers with sales <$100K
			Drugs
			Medical devices, accessories, and supplies
			Food, food ingredients, or prepared food
			Baby products
			Deliveries to a purchaser that is exempt from sales tax
			Deliveries by a food and beverage establishment
Colorado	28 cents per delivery	All retail delivery by motor vehicles of tangible items subject to sales tax	Retailers with sales <$500K New businesses

Source: Colorado Department of Revenue; Minnesota Department of Revenue.

Figure B4: Retail Delivery Fee explanations to the consumers



Source: Amazon website - link

C. Randomization Checks

Table C1: Randomization checks: Experiment # 1 (return to p. **??**)

	Control (1)	Treatment (2)	Difference (3)	p-value (4)
Mean Historical Purchases ($)	12.38	12.36	0.01 (1.65)	0.10
Historical Quantity Bought	1.97	1.97	0.00 (0.02)	0.98
Estimated networth	13.01	13.01	0.00 (0.28)	0.78
Estimated income	11.64	11.64	0.00 (1.65)	0.10
Age	3.96	3.96	-0.00 (-0.93)	0.35
Home Value	12.55	12.55	0.00 (0.43)	0.67
Observations	12,821,840	32,739,969		

Notes: This table shows randomization checks in the first experiment.

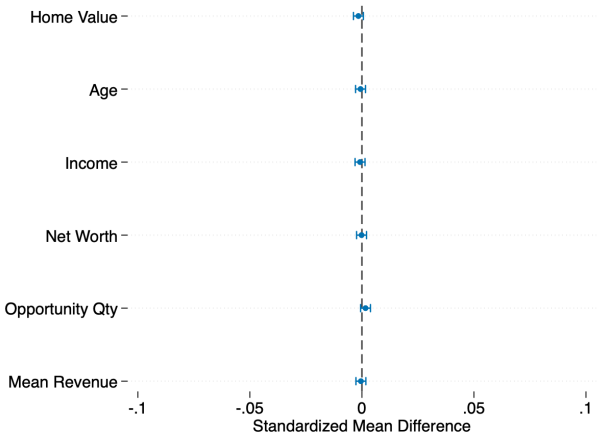
Table C2: Randomization checks: Experiment # 2 (return to p. **??**)

	Control (1)	Treatment (2)	Difference (3)	p-value (4)
Mean Historical Purchases ($)	15.44	15.41	0.02 (1.22)	0.22
Historical Quantity Bought	2.17	2.17	-0.00 (-1.18)	0.24
Estimated networth	12.90	12.90	-0.00 (-0.12)	0.90
Estimated income	11.64	11.64	0.00 (0.23)	0.82
Age	3.91	3.91	0.00 (0.38)	0.71
Home Value	12.56	12.56	0.00 (0.60)	0.55
Observations	1,565,497	3,128,060		

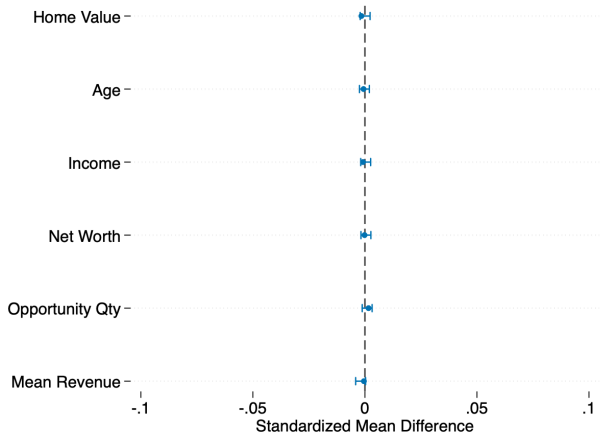
Notes: This table shows randomization checks in the second experiment.

FIGURE C5: Randomization checks: Experiment #2

(a) Substitute+ReRank vs Control



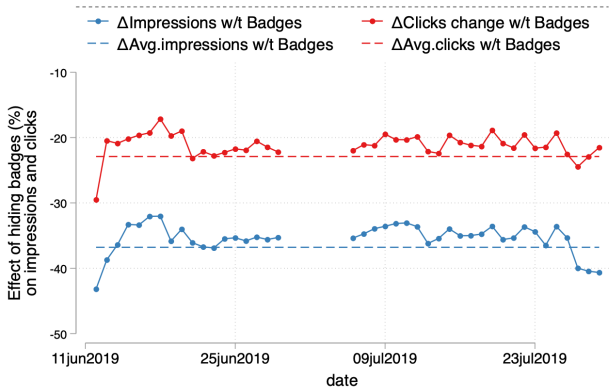
(b) Substitute vs Control



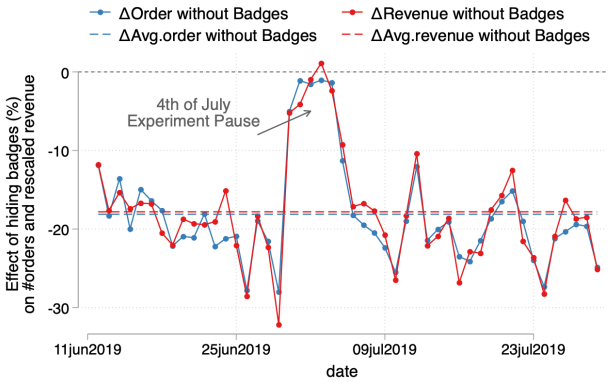
Notes. This figure shows the randomization checks between Control and Substitute+ReRank and Substitute conditions.

FIGURE C6: Changes in product outcomes

(a) Δimpressions and clicks



(b) Δorders and revenue

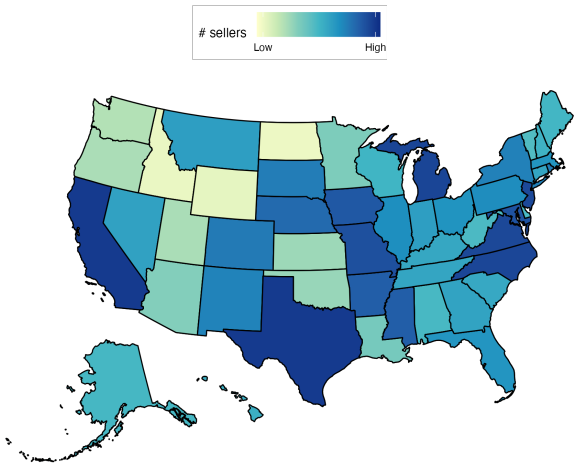


Notes. Figure C6a plots the change in impressions and clicks, and Figure C6b plots the change in orders and revenue, comparing periods when the product's badge is hidden to when it is visible. Dashed lines indicate mean differences.

D. Additional Descriptives: Supply side

Figure D7: Distribution of 1- and 2-day sellers across the US

(a) Number of Sellers



(b) Average shipping Costs across the US

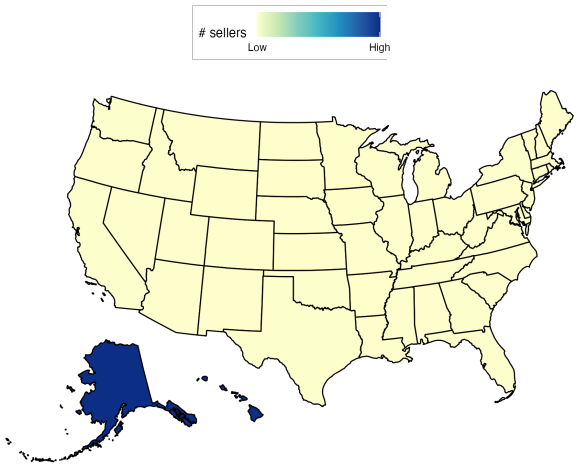
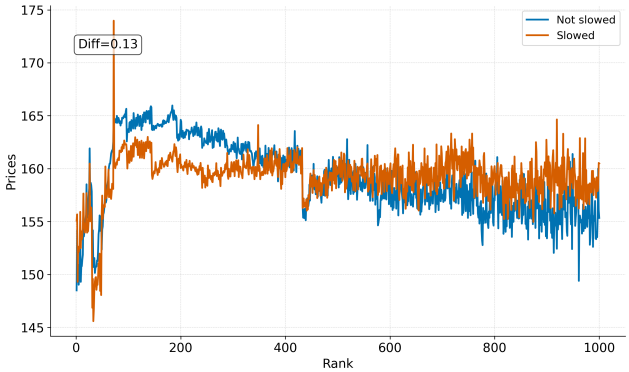
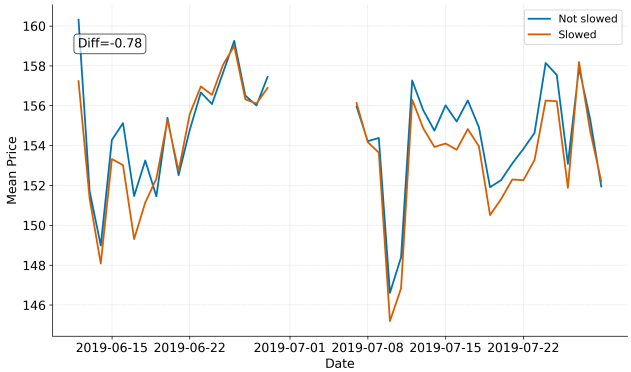


Figure D8: Price differences b/w slowed and not slowed products (return to p. 29)

(a) by rank



(b) by experiment date



Notes. This figure presents the price differences among slowed and not-slowed fast shipping items by rank and experiment date. Figure D8a plots the prices of slowed and not-slowed fast-shipping items by rank. Figure D8b plots the prices of slowed and not-slowed fast-shipping items by experiment date. The annotation summarizes the average price difference between groups using a t-test with unequal variances. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE D3: Summary Statistics by fulfillment type

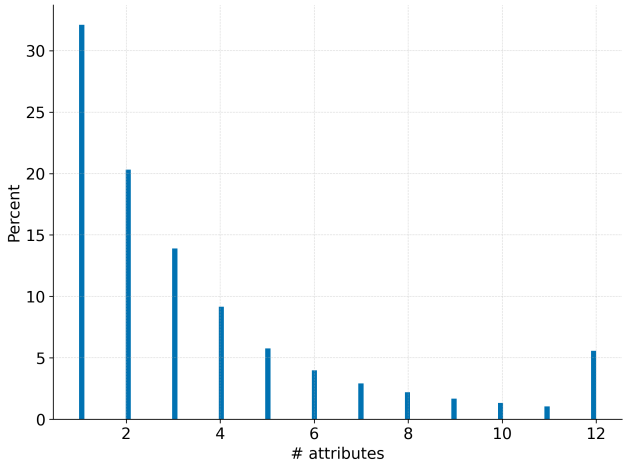
	Min	Mean	Median	Max	SD
	(1)	(2)	(3)	(4)	(5)
Panel A: Rating					
Dropship	1.00	4.39	4.63	5.00	0.81
Castlegate	1.00	4.48	4.65	5.00	0.67
Panel B: Review Count					
Dropship	0.00	45.27	3.00	57081.00	530.58
Castlegate	0.00	70.46	5.00	107038.00	706.96
Panel C: Return Share					
Dropship	0.00	13.56	1.77	100.00	25.40
Castlegate	0.00	7.66	1.75	100.00	16.79
Panel D: Incident Rate					
Dropship	0.00	0.34	0.00	1.00	0.48
Castlegate	0.00	0.34	0.00	1.00	0.47

Notes: This table shows summary statistics by fulfillment type. Each panel reports descriptive statistics for a key variable by fulfillment type.

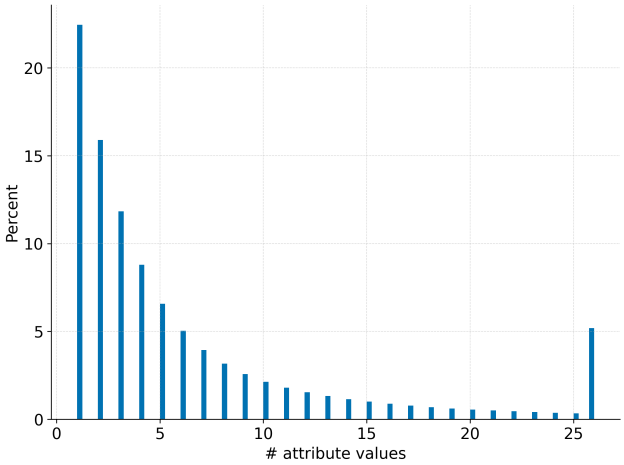
E. Additional Descriptives: Refinement Behavior

Figure E9: Filtering Behavior: Experiment # 1 (return to p. 8)

(a) # attributes



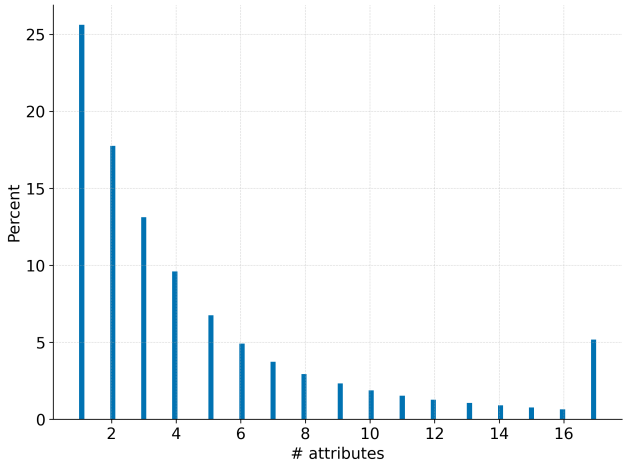
(b) # attribute values



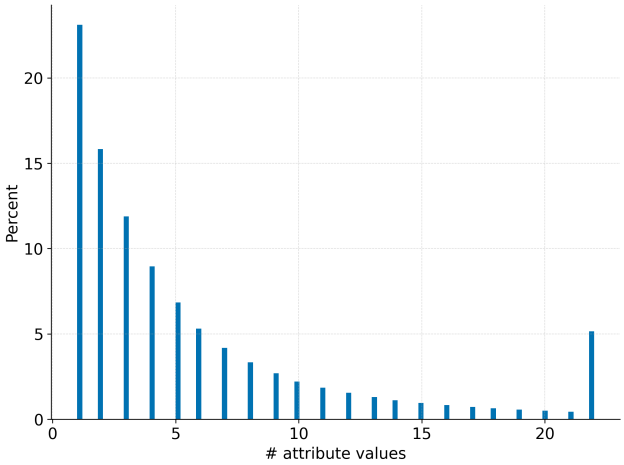
Notes. This figure shows the filtering behavior in Experiment # 1 among filtering consumers. Figure E9a shows the percent of consumers by # of attributes they applied. Figure E9b shows the percent of consumers by # of attribute values they applied when filtering.

Figure E10: Filtering Behavior: Experiment # 2 (return to p. 8)

(a) # attributes



(b) # attribute values



Notes. This figure shows the filtering behavior in Experiment # 1 among filtering consumers. Figure E10a shows the percent of consumers by # of attributes they applied. Figure E10b shows the percent of consumers by # of attribute values they applied when filtering.

F. Addressing potential Stable Unit Treatment Value Assignment

F.1. Experiment #1: ranks

$$\log(\text{TreatmentRank}_{it}) = \alpha + \beta_1 \cdot \log(\text{TreatmentScore}_{it}) + \beta_2 \cdot \log(\text{ControlScore}_{it}) + \varepsilon_{it} \tag{7}$$

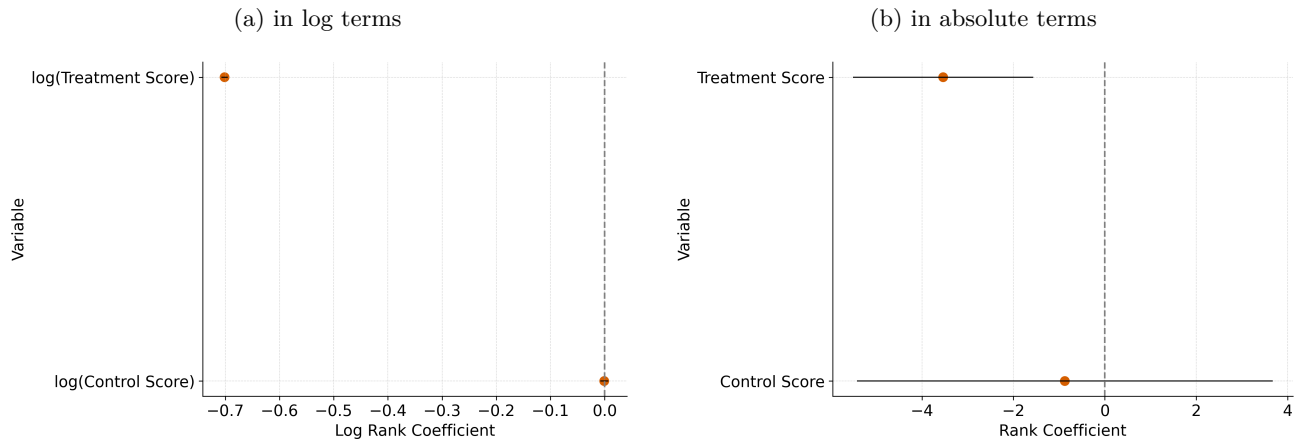
The dependent variable $\log(\text{TreatmentRank}_{it})$ denotes the natural logarithm of the rank of product i at time t within the treatment group. The variable $\log(\text{TreatmentScore}_{it})$ is the logarithm of the product's own score, while $\log(\text{ControlScore}_{it})$ captures the logarithm of the average score of control group products at the same time. The coefficient β_1 captures how sensitive a product's rank is to changes in its own score, controlling for the performance of control products. The term α denotes the intercept, and ε_{it} is the error term.

Table F4: Rank Score spillover test: treatment vs. control ranks

	Log(Treatment Rank)
Intercept	6.891***
	(0.001)
log(Treatment Score)	-0.701***
	(0.003)
Log(Control Score)	-0.001
	(0.003)
Observations	2,645,705

Notes. This table reports OLS estimates of the relationship between the treatment and control scores and the log treatment ranks. Standard errors are reported in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

FIGURE F11: Rank Score spillover test: treatment vs. control ranks



Notes. This figure plots coefficients from an OLS regression of log transaction rank in treatment bins on the log of predicted scores in treatment and control bins. Points indicate coefficient estimates; horizontal bars show 95% confidence intervals. The test assesses whether scoring in one group predicts outcomes in the other, providing a check for potential cross-arm interference.

F.2. Experiment #1: inventory

$$\text{OutOfStocks}_{it} = \alpha + \beta_1 \cdot \text{TreatmentOrders}_{it} + \beta_2 \cdot \text{ControlOrders}_{it} + \varepsilon_{it} \tag{8}$$

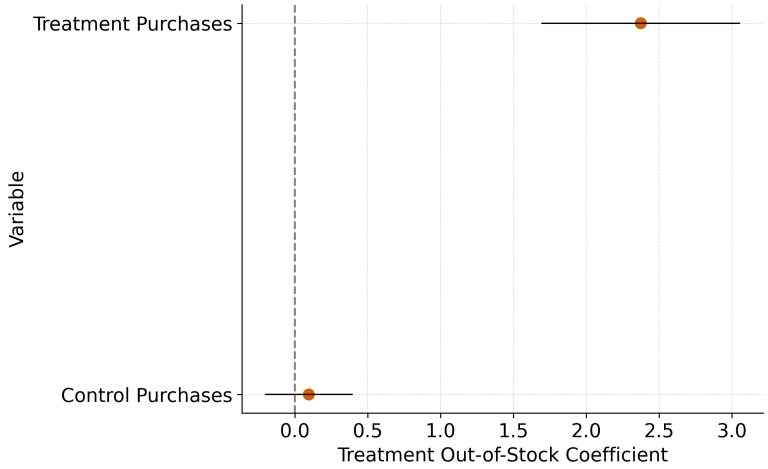
The dependent variable OutOfStocks_{it} denotes the number of out-of-stock incidents for product i at time t. The variable $\text{TreatmentOrders}_{it}$ measures the number of orders placed for the product by treated units, while $\text{ControlOrders}_{it}$ captures the number of orders placed by control units. The coefficient β_1 thus reflects the marginal effect of treatment-related demand on out-of-stock events, controlling for concurrent demand from the control group. The term α denotes the intercept, and ε_{it} is the error term.

Table F5: Out-of-stock spillover test: treatment vs. control purchases

	Treatment Out-of-Stocks
Intercept	1.771^{***}
	(0.007)
Treatment Orders	2.375^{***}
	(0.345)
Control Orders	0.097
	(0.151)
Observations	2,645,705

Notes. This table reports OLS estimates of the relationship between order volume and out-of-stock occurrences among treated products. The dependent variable is the number of out-of-stock events. Regressors include the number of orders from treated and control products. Standard errors are reported in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

FIGURE F12: Out-of-stock spillover test: treatment vs. control purchases



Notes: This figure plots coefficients from an OLS regression of the out-of-stock rate on purchase counts in control and treatment groups. Points show estimates; horizontal bars denote 95% confidence intervals. The analysis tests for potential interference between units by examining whether purchases in one group predict stockouts in the other.

F.3. Experiment #2: ranks

Table F6

	Treatment Out-of-Stocks
Intercept	-0.001***
	(0.000)
Substitute+Re-Rank	0.000
	(0.000)
Substitute	0.000
	(0.000)
Observations	9,908,635

Notes. This table reports OLS estimates of the relationship between order volume and out-of-stock occurrences among treated products. The dependent variable is the number of out-of-stock events. Regressors include the number of orders from treated and control products. Standard errors are reported in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

G. Additional Experimental Results

Table G7: Estimated effect of hidden badges on log clicks across rank bins

	log(clicks)						
	(1) bin: 1-48	(2) bin: 49-96	(3) bin: 97-144	(4) bin: 145-192	(5) bin: 193-240	(6) bin: 241-288	(7) bin: 289+
Hidden Badge	-0.096***	-0.102***	-0.092***	-0.082***	-0.080***	-0.070***	-0.054***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Intercept	1.009***	0.828***	0.666***	0.552***	0.476***	0.416***	0.252***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Product FE	✓	✓	✓	✓	✓	✓	✓
Date FE	✓	✓	✓	✓	✓	✓	✓
Observations	2,489,815	2,158,302	1,368,084	961,199	691,775	518,372	3,042,439
Product Clusters	123,932	139,777	133,595	121,380	105,341	91,985	132,844
Date Clusters	41	41	41	41	41	41	41

Notes. This table reports estimates from separate regressions of log clicks on a Hidden Badge indicator across seven rank bins. Data is at the product-date level for fast shipped products. All specifications include product and date fixed effects. Standard errors clustered by product and date. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table G8: Estimated effect of hidden badges on clicks across rank bins

	clicks						
	(1) bin: 1-48	(2) bin: 49-96	(3) bin: 97-144	(4) bin: 145-192	(5) bin: 193-240	(6) bin: 241-288	(7) bin: 289+
Hidden Badge	-3.160***	-1.051***	-0.582***	-0.395***	-0.330***	-0.315***	-0.216***
	(0.259)	(0.074)	(0.033)	(0.023)	(0.021)	(0.023)	(0.013)
Intercept	15.889***	5.300***	3.084***	2.206***	1.743***	1.489***	0.826***
	(0.101)	(0.030)	(0.012)	(0.009)	(0.008)	(0.009)	(0.006)
Product FE	✓	✓	✓	✓	✓	✓	✓
Date FE	✓	✓	✓	✓	✓	✓	✓
Observations	2,489,815	2,158,302	1,368,084	961,199	691,775	518,372	3,042,439
Product Clusters	123,932	139,777	133,595	121,380	105,341	91,985	132,844
Date Clusters	41	41	41	41	41	41	41

Notes. This table reports estimates from separate regressions of clicks on a Hidden Badge indicator across seven rank bins. Data is at the product-date level for fast shipped products. All specifications include product and date fixed effects. Standard errors clustered by product and date. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table G9: Estimated effect of hidden badges on log orders across rank bins

	log(orders)						
	(1) bin: 1-48	(2) bin: 49-96	(3) bin: 97-144	(4) bin: 145-192	(5) bin: 193-240	(6) bin: 241-288	(7) bin: 289+
Hidden Badge	-0.015***	-0.006***	-0.003***	-0.002***	-0.002***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Intercept	0.082***	0.031***	0.018***	0.012***	0.010***	0.008***	0.004***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Product FE	✓	✓	✓	✓	✓	✓	✓
Date FE	✓	✓	✓	✓	✓	✓	✓
Observations	2,489,815	2,158,302	1,368,084	961,199	691,775	518,372	3,042,439
Product Clusters	123,932	139,777	133,595	121,380	105,341	91,985	132,844
Date Clusters	41	41	41	41	41	41	41

Notes.

TABLE G10: Estimated effect of hidden badges on orders across rank bins

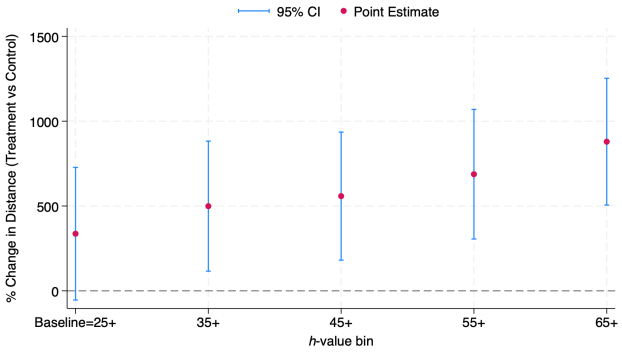
	orders						
	(1) bin: 1-48	(2) bin: 49-96	(3) bin: 97-144	(4) bin: 145-192	(5) bin: 193-240	(6) bin: 241-288	(7) bin: 289+
Hidden Badge	-0.039***	-0.010***	-0.005***	-0.004***	-0.003***	-0.002***	-0.001***
	(0.002)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Intercept	0.187***	0.050***	0.027***	0.019***	0.015***	0.012***	0.005***
	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Product FE	✓	✓	✓	✓	✓	✓	✓
Date FE	✓	✓	✓	✓	✓	✓	✓
Observations	2,489,815	2,158,302	1,368,084	961,199	691,775	518,372	3,042,439
Product Clusters	123,932	139,777	133,595	121,380	105,341	91,985	132,844
Date Clusters	41	41	41	41	41	41	41

Notes. This table reports estimates from separate regressions of log rank on a Hidden Badge indicator across seven rank bins. Data is at the product-date level for fast shipped products. All specifications include product and date fixed effects. Standard errors clustered by product and date. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

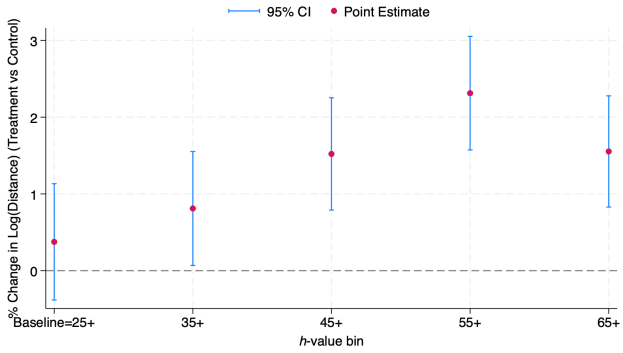
G.1. ROBUSTNESS CHECKS: DISTANCE MEASURES

FIGURE G13: Standard Distance outcomes by h-value bins (return to p. ??)

(a) Distance



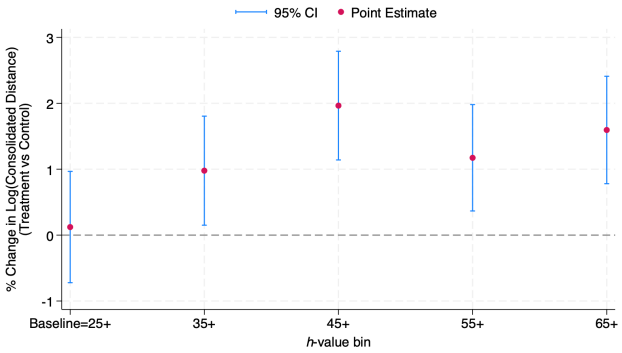
(b) Log(Distance)



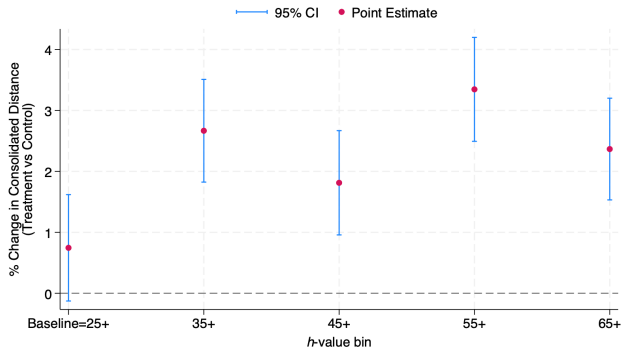
Notes. This figure shows heterogeneous distance outcomes by h-value bins. Panel (a) shows the distance changes in treatment versus control by $h-$value bins. Panel (b) shows the log distance changes in treatment versus control by $h-$value bins.

FIGURE G14: Alternative Distance outcomes by h-value bins: Upper Bound Castlegate

(a) Log(Consolidated Distance):Mean



(b) Log(Consolidated Distance):Max



Notes. This figure shows heterogeneous distance outcomes by h-value bins. Panel (a) shows the distance changes in treatment versus control by $h-$value bins. Panel (b) shows the log distance changes in treatment versus control by $h-$value bins.

G.2. ROBUSTNESS CHECKS: RANKINGS

TABLE G11: Effect of treatment and rank changes on product performance

	log(revenue)		clicked		log(clicks)		log(# orders)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treated	-0.016***	-0.016***	0.000	0.000	0.000	0.000	-0.004***	-0.004***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
% $\Delta rank$	-0.001***		0.000		-0.000***		-0.000***	
	(0.000)		(0.000)		(0.000)		(0.000)	
Treated $\times$%$\Delta rank$	0.001***		-0.000**		0.000***		0.000***	
	(0.000)		(0.000)		(0.000)		(0.000)	
$\Delta rank$		-0.000***		0.000***		0.000		-0.000***
		(0.000)		(0.000)		(0.000)		(0.000)
Treated $\times\Delta rank$		0.000***		-0.000		0.000***		0.000***
		(0.000)		(0.000)		(0.000)		(0.000)
Intercept	0.302***	0.302***	0.505***	0.505***	0.444***	0.444***	0.067***	0.067***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	45,561,809	45,561,809	45,561,809	45,561,809	45,561,809	45,561,809	45,561,809	45,561,809

Notes. This table reports OLS estimates of the effect of treatment and rank changes on product performance outcomes. Dependent variables include log(revenue) (columns 1–2), a binary indicator for receiving any clicks (columns 3–4), log(clicks) (columns 5–6), and log(number of orders) (columns 7–8). Columns alternate between using percent change in rank (% Δ rank) and absolute change in rank (Δ rank), as well as their interactions with the treatment indicator. Standard errors are reported in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE G12: Effect of treatment and rank changes on shipping speed and distance

	fast shipped		mean distance		log(mean distance)		sum distance		log(sum distance)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
treatment	-0.024***	-0.024***	5.947***	5.993***	0.013***	0.013***	35.010***	35.903***	0.022***	0.022***
	(0.003)	(0.003)	(1.034)	(1.034)	(0.002)	(0.002)	(4.117)	(4.115)	(0.002)	(0.002)
$\Delta rank$	0.000		-0.003		-0.000		-0.026		-0.000**	
	(0.000)		(0.004)		(0.000)		(0.017)		(0.000)	
Treated $\times\Delta rank$	-0.000		0.000		0.000		-0.039**		-0.000	
	(0.000)		(0.005)		(0.000)		(0.019)		(0.000)	
% $\Delta rank$		0.000		-0.063***		-0.000***		-0.270***		-0.000***
		(0.000)		(0.016)		(0.000)		(0.065)		(0.000)
Treated $\times\%\Delta rank$		-0.000		0.033*		0.000***		0.090		0.000***
		(0.000)		(0.017)		(0.000)		(0.067)		(0.000)
Intercept	-0.656***	-0.656***	922.873***	922.847***	6.389***	6.389***	1720.572***	1720.524***	6.790***	6.790***
	(0.003)	(0.003)	(0.868)	(0.868)	(0.002)	(0.002)	(3.455)	(3.454)	(0.002)	(0.002)
Observations	2,006,737	2,006,737	2,006,737	2,006,737	2,006,737	2,006,737	2,006,737	2,006,737	2,006,737	2,006,737

Notes. This table reports OLS estimates of the effect of treatment status and product rank changes on shipping outcomes. Dependent variables include the likelihood of fast shipping (columns 1–2), mean shipping distance (columns 3–6), and total shipping distance (columns 7–10), both in levels and logs. Columns alternate between using absolute (Δ rank) and percent (% Δ rank) changes in rank, as well as their interactions with the treatment indicator. Standard errors are reported in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE G13: Effect of treatment and rank changes on product returns

	returned		# returned	
	(1)	(2)	(3)	(4)
treatment	0.076***	0.077***	0.008***	0.008***
	(0.007)	(0.007)	(0.001)	(0.001)
$\Delta rank$	-0.000		-0.000	
	(0.000)		(0.000)	
Treated $\times\Delta rank$	0.000		-0.000	
	(0.000)		(0.000)	
% $\Delta rank$		-0.001***		-0.000*
		(0.000)		(0.000)
Treated $\times\%\Delta rank$		0.001**		0.000
		(0.000)		(0.000)
Intercept	-3.075***	-3.077***	0.064***	0.064***
	(0.006)	(0.006)	(0.001)	(0.001)
Observations	2,006,737	2,006,737	2,006,737	2,006,737

Notes. This table reports OLS estimates of the effect of treatment status and rank changes on product returns. Dependent variables include an indicator for whether a product was returned (columns 1–2) and the number of returned units (columns 3–4). Specifications alternate between using absolute change in rank (Δ rank) and percent change in rank (% Δ rank), along with their interactions with the treatment indicator. Standard errors are reported in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

H. Additional Experimental Results: Experiment # 2

We split the analysis based on how shipping costs are displayed and realized throughout the consumer journey (Table H14). *In-view shipping costs* refer to the shipping costs shown to consumers when a product is visible on their screen; these are subject to selection as users may scroll and skip certain items. *Pageload shipping costs* capture the shipping costs displayed at the time the product page loads, regardless of whether the consumer notices them. *Log shipping costs* represent the actual shipping costs associated with purchases—that is, those realized at checkout. In the **Substitute + Re-rank** condition, we find that consumers viewed products with 0.8% lower in-view shipping costs and loaded product pages with 0.9% lower shipping costs. Ultimately, they ordered products with 2.4% lower realized shipping costs. Note that Columns (1)–(2) report estimated—not actual—shipping costs, which may partly explain why the realized values are lower.

Table H14: Shipping costs effects

	OLS		
	(1)	(2)	(3)
	log(in-view shipping costs)	log(pageload shipping costs)	log(shipping costs)
Substitute+Re-rank	-0.008***	-0.009***	-0.024**
	(0.001)	(0.001)	(0.009)
Substitute	0.009***	0.008***	-0.008
	(0.001)	(0.001)	(0.009)
Intercept	3.307***	3.297***	3.113***
	(0.001)	(0.001)	(0.007)
Observations	7,709,473	8,144,437	39,949

Notes. This table reports the output from the estimation of equation **??** on **purchasing experimental consumers**. Data is at the device-level. Columns (1)-(4) use OLS specification results. Robust standard errors in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. s-value is taken as a mean.

In contrast, consumers in the **Substitute** condition viewed products with 0.8–0.9% *higher* shipping costs than the control group, and there was no significant change in realized shipping costs. This may be because the Substitute condition altered the product set but did not re-rank items to prioritize those with lower shipping costs. It is possible that Substitute replaced fast-shipping items with alternatives that were equally fast but less desirable, prompting users to explore other, potentially more expensive, options.

We further explore this effect by separating pageload results based on refinement and personalization. As shown in Tables H15 and H16, the **Substitute + Re-rank** condition consistently guided consumers toward lower shipping cost products, regardless of whether the pageloads were filtered or personalized. While the magnitudes vary, the direction of the effects remains uniformly negative across all specifications.

TABLE H15: Pageload shipping costs by refinement and personalization outcomes

	log(pageload shipping costs)			
	(1) filtered=0, personalization=0	(2) filtered=0, personalization=1	(3) filtered=1, personalization=1	(4) filtered=1, personalization=0
Substitute+Re-rank	-0.039***	-0.026***	-0.055***	-0.007***
	(0.003)	(0.005)	(0.003)	(0.002)
Substitute	-0.002	0.011**	-0.014***	0.002
	(0.003)	(0.005)	(0.003)	(0.002)
Intercept	3.406***	3.454***	3.485***	3.307***
	(0.002)	(0.004)	(0.002)	(0.001)
Observations	878,398	285,001	818,577	1,588,196

Notes. This table reports the output from the estimation of equation 5. Data is at the product-session-treatment group level. Columns (1)-(4) use OLS specification results. Robust standard errors in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. s-value is taken as a mean.

TABLE H16: In-view shipping costs by refinement and personalization outcomes

	log(in-view shipping costs)			
	(1) filtered=0, personalization=0	(2) filtered=0, personalization=1	(3) filtered=1, personalization=1	(4) filtered=1, personalization=0
Substitute+Re-rank	-0.037***	-0.034***	-0.051***	-0.004**
	(0.003)	(0.006)	(0.003)	(0.002)
Substitute	0.006**	0.003	-0.017***	0.002
	(0.003)	(0.006)	(0.003)	(0.002)
Intercept	3.409***	3.449***	3.511***	3.320***
	(0.002)	(0.004)	(0.002)	(0.001)
Observations	821,575	238,454	722,982	1,484,214

Notes. This table reports the output from the estimation of equation 5. Data is at the product-session-treatment group level. Columns (1)-(4) use OLS specification results. Robust standard errors in parentheses. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. s-value is taken as a mean.

TABLE H17: Experimental impact on prices and quantity of clicks

	OLS			
	(1) price	(2) log(price)	(3) quantity	(4) log(quantity)
Substitute+Re-rank	0.490	0.003	0.004	0.002
	(1.927)	(0.007)	(0.008)	(0.002)
Substitute	-1.657	-0.008	0.001	0.001
	(1.935)	(0.007)	(0.008)	(0.002)
Intercept	190.465***	4.746***	1.249***	0.771***
	(1.368)	(0.005)	(0.006)	(0.001)
Observations	112,970	112,970	112,970	112,970

Notes. This table reports that there were no statistically significant differences in prices or click quantity in experiment #2. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.