Three Essays on Data Science for Healthcare and Retail Operations Management

Zahra Jalali
Doctor of Philosophy

Desautels Faculty of Management
McGill University
Montreal, Quebec, Canada

April 23, 2024

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Abstract

This thesis comprises three essays that delve into the realm of data analytics in healthcare and retail operations management. The first essay explores the impact of boarding congestion on patients' treatment time in emergency departments (EDs) through a multimethodology approach. Empirical analyses using data from eight EDs show an inverted U-shaped relationship between boarding congestion and ED treatment time. Two mechanisms contributing to this relationship are identified: increased workload on ED resources due to boarding patients and scheduling of hospitalist visits triggered by boarding congestion. Building on the empirical findings, an analytical framework is devised, proposing two operational interventions to mitigate the impact of boarding congestion on treatment time. Simulation results demonstrate that these interventions can collectively reduce the impact of boarding congestion by 68% in a typical ED, offering valuable insights for making sound operational decisions in ED management.

In the second essay, an exploration of the increasingly popular retail practice of pickup partnerships is studied. These partnerships enable online retailers to offer in-store pickup services by collaborating with physical stores. The study examines two common policies adopted by online retailers in these partnerships: the fixed fee policy and the coupon policy. A stylized model is developed to capture the essential features of pickup partnerships. Findings reveal that while the coupon policy can expand the online retailer's market coverage, it may not always lead to increased profits. The research identifies the circumstances under which an online retailer should opt for the fixed fee

or coupon policies in a pickup partnership. The analysis also show that these two policies may entail inefficiencies when the incentives of the two parties are not aligned. To alleviate such inefficiencies, a new policy that aims to align both parties' incentives is proposed. By proposing a new policy, it also strives to make pickup partnerships more efficient than current practices.

In the third essay, the impact of sales is investigated concerning simultaneous changes in the vertical locations of multiple products, and whether this effect varies across products. To measure these effects, a novel field experiment is designed and deployed in six retail stores for 20 weeks. The results indicate that the effect of vertical location change is (i) contingent on how other products are reorganized as a result of that change and (ii) heterogeneous across products. On average, sales are generated at 13.8% and 8.5% higher levels relative to the stoop and stretch levels, respectively, when products are placed at the eye-level. The product profiles that benefit the most from being displayed at the eye-level are characterized. It is also conveyed that the eye-level shelf boost in sales comes at the expense of a sales loss for products moved to other shelves. Specifically, a pure substitution pattern is observed, leading to no change in the overall sales of the shelving unit. Given the differences in product margins and interplay among products (arising from simultaneously changing vertical locations), the careful selection of products to display at the eye-level can benefit retailers. A counterfactual analysis shows that profits can be increased by 2.2% by optimizing planograms while considering the average interplay among products. Moreover, profits can be boosted by up to 3% by incorporating product heterogeneity into planogram optimization.

These essays contribute to the fields of data analytics in healthcare and retail operations management, providing valuable insights for improving decision-making processes and optimizing operational strategies in their respective domains.

Abrégé

Cette thèse comprend trois essais qui plongent dans le domaine de l'analyse des données dans les soins de santé et la gestion des opérations de vente au détail. Le premier essai explore l'impact de la congestion de l'embarquement sur le temps de traitement des patients dans les services d'urgence par le biais d'une approche multiméthodologique. Des analyses empiriques utilisant des données provenant de huit services d'urgence montrent une relation en forme de U inversé entre la congestion de l'embarquement et le temps de traitement du service d'urgence. Deux mécanismes contribuant à cette relation sont identifiés : l'augmentation de la charge de travail sur les ressources des urgences due à l'embarquement des patients et la programmation des visites des hospitalistes déclenchée par la congestion de l'embarquement. Sur la base des résultats empiriques, un cadre analytique est élaboré, proposant deux interventions opérationnelles pour atténuer l'impact de la congestion de l'embarquement sur le temps de traitement. Les résultats de la simulation démontrent que ces interventions peuvent collectivement réduire l'impact de l'embarquement de 68% dans un service d'urgence typique, offrant des informations précieuses pour prendre des décisions opérationnelles judicieuses dans la gestion des services d'urgence.

Le deuxième essai étudie la pratique de plus en plus populaire des partenariats de ramassage dans le commerce de détail. Ces partenariats permettent aux détaillants en ligne d'offrir des services de ramassage en magasin en collaborant avec des magasins physiques. L'étude examine deux politiques communes adoptées par les détaillants en ligne dans le cadre de ces parte-

nariats : la politique des frais fixes et la politique des coupons. Un modèle stylisé est développé pour saisir les caractéristiques essentielles des partenariats de ramassage. Les résultats révèlent que si la politique des coupons peut étendre la couverture du marché du détaillant en ligne, elle n'entraîne pas toujours une augmentation des bénéfices. L'étude identifie les circonstances dans lesquelles un détaillant en ligne devrait opter pour une politique de frais fixes ou de coupons dans le cadre d'un partenariat de ramassage. L'analyse montre également que ces deux politiques peuvent entraîner des inefficacités lorsque les incitations des deux parties ne sont pas alignées. Afin d'atténuer ces inefficacités, une nouvelle politique visant à aligner les incitations des deux parties est proposée. En proposant une nouvelle politique, elle s'efforce également de rendre les partenariats de ramassage plus efficaces que les pratiques actuelles.

Dans le troisième essai, l'impact des ventes est étudié en ce qui concerne les changements simultanés dans les emplacements verticaux de plusieurs produits, et si cet effet varie d'un produit à l'autre. Pour mesurer ces effets, une nouvelle expérience de terrain est conçue et déployée dans six magasins de détail pendant 20 semaines. Les résultats indiquent que l'effet d'un changement d'emplacement vertical (i) dépend de la manière dont les autres produits sont réorganisés à la suite de ce changement et (ii) est hétérogène d'un produit à l'autre. En moyenne, les ventes sont générées à des niveaux supérieurs de 13.8% et de 8.5% par rapport aux niveaux du perron et de l'étirement, respectivement, lorsque les produits sont placés à la hauteur des yeux. Les profils de produits qui bénéficient le plus d'une présentation à hauteur des yeux sont caractérisés. Il est également précisé que l'augmentation des ventes des produits placés au niveau des yeux s'accompagne d'une perte de ventes pour les produits déplacés vers d'autres rayonnages. Plus précisément, on observe un modèle de substitution pure, qui n'entraîne aucun changement dans les ventes globales de l'unité de rayonnage. Compte tenu des différences de marges sur les produits et de l'interaction entre les produits (résultant d'un changement simultané d'emplacement vertical), les détaillants peuvent tirer profit d'une sélection minutieuse des produits à exposer au niveau des yeux. Une analyse contrefactuelle montre que les bénéfices peuvent être augmentés de 2.2% en optimisant les planogrammes tout en tenant compte de l'interaction moyenne entre les produits. En outre, les bénéfices peuvent être augmentés de 3% en intégrant l'hétérogénéité des produits dans l'optimisation des planogrammes.

Ces essais contribuent aux domaines de l'analyse des données dans les soins de santé et de la gestion des opérations de vente au détail, en fournissant des informations précieuses pour améliorer les processus de prise de décision et optimiser les stratégies opérationnelles dans leurs domaines respectifs.

Contribution

Statement of the co-authorship

The following manuscripts are based on this thesis:

- **Jalali Z.**, Kucukyazici B., Gumus M., 2023 "On the Effects of Boarding Congestion on Treatment Time in Emergency Departments." Under Revise & Resubmit at *Production and Operations Management.*

  The primary author (student) performed all data analyses, econometric modeling, and constructed the MDP and simulation model. Throughout the study, the primary author assumed the primary responsibility for research execution, manuscript writing, and the development of the computer models essential for problem-solving. The second and third co-authors participated during the modeling and computational analysis phases, offering guidance to the primary author in interpreting the results. Additionally, they reviewed the manuscript and provided editing suggestions.

- **Jalali Z**, Cohen M., Ertekin N., Gumus M., 2022 "Offline-Online Retail Collaboration via Pickup Partnership." Under Reject & Resubmit at *Manufacturing & Service Operations Management.*

  The main author (a student) carried out all data analyses and modeling. Throughout the study, the primary author led the research execution, manuscript writing, and the development of the analytical model. The second, third, and fourth co-authors participated in the modeling phase, offering guidance to the primary author on the model. Furthermore, they reviewed the manuscript and provided editing suggestions.

- **Jalali Z**, Cohen M., Ertekin N., Gumus M., 2023 "Vertical Product Location Effect on Sales: A Field Experiment in Convenience Stores" In preparation for *Manufacturing & Service Operations Management.*

  The primary author (student) conducted a field experiment to gather data sets and performed data cleaning. Throughout the study, the primary author took the lead in designing the field experiment, collecting data, conducting research, writing the manuscript, and developing econometric models, as well as performing analysis and creating an optimization model. The second, third, and fourth co-authors contributed to designing the experiment, and engaging in modeling and econometric analysis. Additionally, they guided the primary author in interpreting the results. Finally, they reviewed the manuscript and offered editing suggestions.

Acknowledgements

Embarking on a doctoral degree is a challenging journey filled with obstacles and uncertainties. I consider myself exceptionally fortunate to have received support, guidance, and love throughout this endeavor. Without the backing of my mentors, colleagues, friends, and family, completing this thesis would not have been possible.

Firstly, I express my gratitude to my supervisors, Prof. Mehmet Gumus and Prof. Maxime Cohen, for their constant guidance and support. I will always treasure our discussions and meetings, as they have profoundly influenced my perspective. Most importantly, I want to thank them for believing in me and fostering an environment conducive to pursuing my dreams. Today, I have a dissertation on a topic that I am genuinely passionate about, and it is all thanks to you.

I am grateful to my co-authors and members of the advisory committee — Dr. Necati Ertekin, Dr. Beste Kucukyazici, and Prof. Georges Zaccour — for their comments and support, which undeniably enhanced this thesis in various aspects. Specifically, I owe a debt of gratitude to Dr. Necati Ertekin for his meticulous feedback from the early stages and, more significantly, for his unwavering support and encouragement throughout these projects.

I would also like to express my sincere gratitude to Fonds de recherche du Québec (FRQNT), Bensadoun Retail School, and Retail Innovation Lab for providing financial support for these projects. Additionally, I extend my appreciation to St. Mary's Hospital in Montreal, Quebec, for supplying me with healthcare data, with special thanks to Eric Belzile for patiently explaining the data. I am grateful to Couch-Tard company for permitting me to conduct field experiments in their stores, with special acknowledgment to Eve Fontaine, Sylvain Normandeau, and Magnus Tagtstrom. Finally, my thanks go to my assistant team, including Saiyan, Tianyi, Huzbah, Anoosh, Jacob, Cheng, and Anna, for their invaluable assistance in collecting data and conducting experiments in physical stores.

A heartfelt appreciation is extended to my friends, Dr. Parnian Afshar and Hadi Pourhaji, for consistently being there for me. The most memorable moments of my years in Montreal and Vancouver involve both of you.

I want to express my deepest gratitude to my mother and father, who have taught me perseverance and dedication and provided unconditional love and support throughout my extensive studies. Special thanks go to my sisters, Somayeh, Azam, and Negar, for their endless support and love.

Lastly, I extend my heartfelt appreciation to Mohammad, my better-half and rock. Thank you for our late-night walks on the McGill Campus and for your consultations and advice during these moments. I appreciate your unwavering belief in my dreams, guiding me through dark times and showing me the light. Thank you for your continual support and patience during the stressful periods of my life.

Contents

List of Figures

List of Tables

1

Introduction

In the dynamic landscape of today's global economy, the integration of data science has emerged as a transformative force, reshaping traditional paradigms in industries ranging from healthcare to retail operations management. This revolution is particularly evident in the realms of healthcare and retail, where the power of data-driven insights has become a cornerstone for informed decision-making, efficiency optimization, and enhanced customer experiences.

1.1 Data Science in Healthcare

The healthcare sector, a cornerstone of societal well-being, has witnessed a remarkable evolution with the advent of data science. From patient care and diagnostics to research and administrative processes, the integration of advanced analytics and machine learning has unleashed a new era of possibilities. Electronic Health Records (EHRs) have become a treasure trove of valuable information, allowing healthcare professionals to analyze patient histories, predict disease trends, and personalize treatment plans.

Furthermore, predictive analytics plays a pivotal role in proactive healthcare management. By leveraging historical data, machine learning algorithms can forecast patient outcomes, enabling healthcare providers to intervene early, reduce hospital readmissions, and optimize resource allocation. This predictive capability is not only enhancing patient care but is also proving instrumental in managing healthcare costs efficiently.

The amalgamation of data science and healthcare extends beyond individual patient interactions. Population health management, a key focus area, involves the analysis of large datasets to identify and address health trends across communities. This proactive approach aids in preventive care initiatives, policy planning, and the overall improvement of public health outcomes.

Data also helps healthcare systems improve their operations by enhancing processes and identifying bottlenecks. One notable bottleneck, particularly in Canada, is within emergency departments. Emergency department (ED) crowding is a widely acknowledged issue, denoting a scenario where the demand for ED services surpasses the capacity to provide timely care.

ED crowding has been a persistent challenge in Quebec for over 40 years. Despite increased attention from political, administrative, and public spheres, instances of ED overcrowding persist, escalating in both frequency and severity. Comparative international studies reveal that the Quebec population not only has the highest rate of ED visits but also experiences the lengthiest waiting times for care in ED settings.

Successfully addressing the overcrowding problem involves three key phases: 1) assessing delays in the patient flow within ED and determining their impact on care quality, 2) investigating factors contributing to these delays, and 3) developing and testing interventions to alleviate ED crowding. This comprehensive approach relies on access to a detailed dataset encompassing emergency patients' trajectories and characteristics.

I have access to a rich and unique dataset, comprising over 500,000 ED visits to eight hospitals in the Greater Montreal region during the 2015–2016 fiscal year. This dataset includes three types of information: (i) patient-flow measures, (ii) patient-level details, and (iii) organizational and operational characteristics of the EDs and associated hospitals. The availability of such a comprehensive dataset enables me to undertake a thorough examination of delays impacting the flow of ED patients through the continuum of care, addressing all three key phases mentioned above. Further details on this research can be found in Chapter 2 of this thesis.

1.2 Data Science in Retail Operations Management

In the fast-paced world of retail, where customer preferences shift rapidly, data science is a game-changer for operations management. Retailers are leveraging data to understand consumer behavior, optimize supply chain operations, and create personalized shopping experiences.

Customer-centricity is at the forefront of retail strategies, and data science enables the collection and analysis of customer data to understand preferences, predict trends, and tailor marketing strategies. Recommendation systems, powered by machine learning algorithms, provide customers with personalized product suggestions, enhancing the overall shopping experience and driving sales.

Supply chain management is another domain where data science is revolutionizing retail operations. Predictive analytics helps in demand forecasting, inventory management, and supply chain optimization, ensuring that retailers can meet customer demands efficiently while minimizing costs. This not only improves operational efficiency but also contributes to sustainability efforts by reducing waste.

In this thesis, I undertook two projects aimed at enhancing retail operations. Firstly, I delved into the concept of "Pickup Partnership," a business model extensively utilized by companies such as Amazon and Wish in the era of omnichannel retailing. This model strategically involves online stores collaborating with physical establishments, like convenience stores, to offer customers an alternative pickup option instead of home delivery.

The rationale behind this business model is multifaceted. Primarily, it allows companies to diminish shipping costs and delivery times while providing customers with a convenient alternative. This option proves advantageous for individuals who may face challenges receiving orders at their homes. The Pickup Partnership model is mutually beneficial, enhancing customer satisfaction for the online store and attracting more customers who seek quicker deliveries without additional fees.

From the perspective of the location partner, engaging in a pickup partnership proves advantageous as well. Firstly, it generates additional foot traffic within the physical store, potentially leading to cross-selling opportunities as visitors may make additional purchases (e.g., coffee or snacks). Secondly, the location partner may have a revenue-sharing or fixed-fee arrangement with the online store for processing pickup orders, establishing a potential win-win scenario for both parties.

In the third chapter of this thesis, I focus on quantifying the benefits of such partnerships and explore various iterations to determine the optimal implementation of this model. The research aims to shed light on the most effective ways to capitalize on the Pickup Partnership business model, offering valuable insights for both online stores and their physical location partners.

Many retailers have limited shelf space available, leading to the necessity of strategically allocating this scarce resource among stocked items. The position of products on the shelf plays a pivotal role in influencing product sales, impacting consumers' choices and overall shopping convenience. Surveys examining supermarket shopping behavior consistently reveal that approximately one-third of purchases are impulse buys triggered by visual elements. This presents a significant opportunity for retailers to enhance incremental revenues and profits through an effective shelf arrangement that heightens shopper awareness and impulsiveness.

Understanding the impact of each shelf space on customer choices empowers retailers in negotiations with suppliers, allowing them to secure reservation fees for high-value locations within the store. While some suppliers currently pay location reservation fees based on agreements, the profitability for retailers to deviate from the optimal shelf arrangement for the agreed reservation fee remains uncertain. Therefore, comprehending the effect of shelf layout also aids retailers in estimating the optimal reservation fee for each space in the store.

In Chapter Four, I present findings from a field experiment conducted in convenience stores, where I systematically measure the vertical location of products on sales shelves. This research aims to provide practical insights into the implications of shelf positioning, offering valuable information for retailers seeking to optimize their shelf arrangements for increased sales and profitability.

2

On the Effects of Boarding Congestion on Treatment Time in Emergency Departments

2.1 Introduction

Emergency department (ED) boarding refers to the temporary holding of patients in the ED after making the decision to admit them (American College of Emergency Physicians 2018). Various factors can cause long boarding times in EDs such as high inpatient services occupancy (Powell et al. 2012), patient-flow-related issues (Feizi et al. 2023), and inefficiencies in the admission process (Mohr et al. 2020). According to the (Agency for Healthcare Research and Quality 2018) survey, 9 out of 10 US hospitals reported holding admitted patients in the ED, with boarding patients accounting for 10-20%

of all ED census (McKenna et al. 2019). ED boarding has been associated with numerous negative consequences for the boarding patients, including lower-quality care (Rabin et al. 2012), decreased patient safety (Armony et al. 2015), delays in receiving timely care (Sills et al. 2011), prolonged in-hospital stays (Chan et al. 2017), and reduced patient satisfaction (Viccellio et al. 2013).

In addition to its detrimental effect on patients' health conditions, ED boarding can have adverse effects on ED performance, affecting all stages of ED patient flow, from the waiting room and treatment section to the boarding area (Batt and Terwiesch 2017). Several studies have investigated the impact of boarding congestion on the length of stay (LOS)[1] in EDs (e.g., White et al. (2013), Khare et al. (2009)). These studies demonstrate that boarding congestion increases the LOS for both admitted and discharged EDs. Some studies have also investigated the effect of boarding congestion on ED waiting time[2] (e.g., Hoot and Aronsky (2008), Saghafian et al. (2015)). These studies reveal that boarding congestion can prolong waiting times in the waiting room as boarding patients continue to occupy ED beds, thereby impeding the admission of new patients to the treatment section. However, to the best of our knowledge, prior studies have not investigated how overcrowding caused by boarding patients in emergency departments affects the time it takes to provide treatment to patients, nor have they explored the potential mechanisms underlying this relationship. Understanding this impact and the underlying mechanisms can provide valuable insights into the effects of boarding congestion and facilitate the development of effective solutions to mitigate its potential negative impact. Therefore, this study aims to address the following questions:

[1]LOS refers to the duration of time a patient spends in the ED from the moment they registered until they are either discharged or admitted to the hospital for further treatment (Yoon et al. 2003).

[2]The waiting time refers to the duration from when the customer is registered in the ED until they are transferred to the treatment section (Batt and Terwiesch 2017).

- How does boarding congestion impact treatment time in EDs?
- What are the key mechanisms that contribute to the impact of boarding congestion on treatment times?
- How can we effectively mitigate the potential adverse impact of boarding congestion on treatment times?

ED boarding may influence the treatment time in multiple ways. First, boarding patients may increase the workload and multitasking level of ED staff. While admitted patients are technically under the care of the hospitalists, there are cases where ED staff are required to provide medical attention to boarding patients (Armony et al. 2015). It is not surprising that, when boarding times are prolonged, these patients may require interim medical attention; thus, ED staff (e.g., physicians and nurses) may often need to provide inpatient services to these patients (Mohr et al. 2020). Additionally, some boarding patients may require diagnostic tests while waiting to be admitted to the hospital (Coil et al. 2016). Prior literature shows that such an increase in care providers' workload and multitasking level can impact service throughput. For instance, Kc (2014) illustrates that while initial multitasking improves productivity, excessive multitasking decreases productivity due to increased non-value-added activities and task switching. Moreover, multitasking often results in increased interruptions (Chisholm et al. 2001), further deteriorating the service rate. All of these factors can contribute to longer treatment times for patients in the ED.

Second, boarding congestion might affect the discretionary behavior of care providers. Previous studies show that crowding in EDs increases physicians' "cognitive load" (Pines 2017), which leads to certain alterations in their behaviors related to diagnostic test ordering and patient prioritization. Specifically, in the face of crowding, the ED physician may order more diagnostic tests to temporarily reduce their workload while waiting for other test results (Berry Jaeker and Tucker 2019). Furthermore, as the level of boarding congestion increases, physicians may order more diagnostic tests to prevent unnecessary admissions, particularly for the “gray area” patients for whom the correct disposition decision remains ambiguous (Soltani et al. 2022). Additionally, under conditions of increased workload, care providers may change their prioritization schemes to manage their workload between boarding patients and emergency patients[3] (Armony et al. 2015). Particularly, higher levels of congestion may lead care providers to prioritize easier tasks, driven by the behavior known as “task completion preference”(Kc et al. 2020).

Lastly, the congestion in the boarding section can result in an increase in hospitalist visits to the ED, as hospitalists–who specialize in the care of hospitalized patients–may be called upon to provide care in the boarding section (Apker et al. 2007). Therefore, when evaluating the impact of boarding congestion on treatment times, it is important to consider its effect on the added workload for ED resources and the frequency of hospitalist visits.

In this paper, we adopt a multimethodology approach to evaluate the impact of boarding congestion on treatment time and propose interventions to mitigate its adverse effects. First, we investigate the impact of boarding congestion on treatment time using a large dataset collected from eight hospitals, including five tertiary and three secondary hospitals, encompassing over 470,000 ED visits. Our analyses reveal an inverted U-shared relationship between boarding congestion and treatment time. More specifically, after controlling for hospital occupancy, we find the treatment time increases from an average of 318 minutes to nearly 420 minutes when boarding congestion increases from 1 to 20 patients. This translates to a 5-minute increase in treatment time for every additional patient in boarding congestion. However, once congestion increases from 25 to 40 patients per hour, treatment time decreases to 378 minutes.

Second, we explore two mechanisms to explain this phenomenon: (i) the

[3] We refer to patients as ”emergency patients” while they are in the waiting room or treatment section and ”boarding patients” while they are in the boarding section.

additional workload imposed on ED resources by boarding patients and (ii) the hospitalist's visits triggered by boarding congestion. Regarding the first mechanism, our findings suggest that boarding congestion increases the demand for ED resources. Regarding the second mechanism, our analysis demonstrates that as boarding congestion increases, the likelihood of hospitalist visits to the boarding section also increases. Furthermore, we observe a reduction in the utilization of ED resources, including ED physicians and diagnostic tests, for boarding patients after being visited by hospitalists.

Finally, based on our empirical findings, we propose two interventions to mitigate the impact of boarding congestion on treatment time: (i) prioritizing patients in the treatment section and (ii) implementing streamlined operational policies for hospitalists' visits to the boarding section.

Regarding the first intervention, the prioritization of patients in the treatment section becomes crucial when both emergency and boarding patients require ED resources. Currently, our partner hospitals prioritize boarding patients without standardized protocols or guidelines. In response, we propose a new prioritization policy based on insights derived from an analytical model. Initially, we develop a Markov decision process (MDP) with reentries to capture the impact of boarding congestion on patient flow in the treatment section. We demonstrate that the optimal course of action for selecting the next patient to serve depends on the treatment census, challenging the current practice and providing evidence for the need to optimize prioritization. Next, we numerically extract the optimal policy properties of our MDP model. Based on these insights, we present a practical policy to control patient flow in the treatment section. To validate the effectiveness of our approach, we calibrate a simulation model using our data. Simulation results indicate that implementing our proposed policy in a tertiary hospital can reduce the treatment census by 4 patients and decrease the total treatment time by 2 patient-years compared to the current practice.

Regarding the second intervention, although hospitalist visits are an ongoing practice in EDs to overcome ED boarding issue (Howell et al. 2010), no prior work has directly addressed the operational challenges pertinent to this role, particularly in providing effective care to boarding patients without increasing their workload in the inpatient ward. Discussions with staff members at partner hospitals revealed that hospitalists often visit boarding patients who are going to be admitted to the inpatient ward where they work. However, these visit plans vary from one hospital or inpatient ward to another, even within the same hospital, due to the absence of standardized protocols. To address this issue, we propose time-based and census-based visiting policies for hospitalists, aiming to reduce the extra workload of ED resources caused by boarding patients without increasing the existing workload of hospitalists (Zhu 2018). The simulation results demonstrate that the census-based policy outperforms the time-based policy, mitigating the adverse impact of boarding patients on the treatment section by 48% with 2.7 fewer patient-years of total treatment time and 4.6 fewer patients in the treatment section on average. Furthermore, our results illustrate that implementing both proposed interventions together can alleviate the impact of boarding congestion on treatment time by up to 68%.

2.2 Literature Review

Our work is related to two areas of research: empirical works that study the impact of crowding on service times; and prior studies on boarding patients, patient flow control, and the role of hospitalists in EDs' performance.

2.2.1 The Impact of Crowding on Service Time

Fundamental queuing theory traditionally assumes that service times remain fixed in system congestion (Wolff 1989). However, recent empirical studies in operations management have demonstrated that crowding can have an impact on service time across various sectors, including banking (Staats and Gino 2012), hospitality (Tan and Netessine 2014), and retail (Wang and Zhou

2018). In the healthcare literature, the impact of crowding on service time has received considerable attention, likely due to the criticality of service time in healthcare systems and the limited control of overcrowding within these systems. Prior literature has investigated the effects of crowding measures on service time in different healthcare settings, including intensive care units (ICUs) (Kim et al. 2017), hospital inpatient wards (Berry Jaeker and Tucker 2017), patient transport services and cardiothoracic surgeries (Kc and Terwiesch 2009).

In the context of EDs, several studies examine the effects of crowding at different stages of ED on the treatment time (e.g., Batt and Terwiesch (2017), Kc (2014)). For instance, Batt and Terwiesch (2017) investigate the impact of waiting room census on the treatment time and identifies an inverted U-shaped relationship. They attribute this relationship to early task initiation, delay in medication delivery, and nurse rushing. Moreover, they show that the treatment time increases with a higher treatment census, which can be explained by ED staff multitasking caused by overcrowding in the treatment section (Kc 2014). Additionally, various medical studies evaluate the impact of ED crowding (including crowding in the waiting room, treatment section, and boarding section) on LOS and waiting time (e.g., McCarthy et al. (2009), Timm et al. (2008), Lucas et al. (2009)). These studies, employing statistical analysis, consistently report that ED crowding leads to increased LOS and waiting time (e.g., Hoot and Aronsky (2008), Morley et al. (2018)).

There is a limited number of medical studies investigating the effects of ED boarding on LOS and waiting times (e.g., White et al. (2013), Khare et al. (2009), Carmen et al. (2018)). These studies consistently demonstrate that boarding congestion contributes to prolonged LOS and waiting times, as patients awaiting admission occupy valuable ED beds and hinder the smooth movement of new patients to the treatment section. Despite focusing on the impact of ED boarding on waiting times, the influence of boarding congestion on treatment time has been largely overlooked in both operations management and medical literature. In operations management literature, there are only two exceptions: First, Carmen et al. (2018) analytically show that boarding patients adversely affect ED congestion and maximum throughput. Second, based on data from a large Israeli hospital, Armony et al. (2015) observe that a boarding patient requires an average of 1.5 minutes of the ED physician's time every 15 minutes. However, this study did not examine any causal effect of boarding patients on ED performance.

Similarly, in medical literature, there are only a few studies demonstrating that hospital occupancy prolongs LOS in EDs and negatively impacts ED performance (e.g., Hillier et al. (2009)). However, it is essential to examine the impact of boarding congestion separately from hospital occupancy. Although hospital occupancy is recognized as a critical factor contributing to boarding congestion (Powell et al. 2012), other factors such as inefficient admission processes and patient-flow issues also play a role in this problem (Luo et al. 2013, Feizi et al. 2023). Moreover, it is important to acknowledge that the effects of boarding congestion and hospital occupancy on the ED staff may differ since hospital occupancy is not directly observable to the ED staff.

Therefore, our research contributes as the first empirical evidence of the impact of congestion caused by boarding patients on ED treatment time. We provide empirical evidence demonstrating that the utilization of ED resources by boarding patients is one of the mechanisms underlying the relationship between boarding congestion and treatment time. Building upon these findings, we propose two interventions to manage boarding patients and mitigate the negative impact of boarding congestion on ED performance: (i) patient flow control in the treatment section and (ii) scheduling of hospitalist visits. In the following sections, we provide a review of the existing literature on these two interventions.

2.2.2 Patient Flow Control and Hospitalists

The first intervention is to control boarding patient flow in EDs by allocating ED resources judiciously between boarding patients requiring the attention of the ED staff and other ED patients. Most studies focusing on prioritizing patient problems in EDs aim to balance the work of ED physicians based on two classes of patients: new patients after triage and in-progress patients (e.g., He et al. (2019)). Other classifications of patients in EDs for the patient flow control problem are based on predicting patients' dispositions (e.g., Saghafian et al. (2012)) and the complexity of treatment procedures (e.g., Saghafian et al. (2014)) (More details on customer classifications in hospital and other healthcare systems can be found in (Helm et al. 2011, Ranjan et al. 2017)). All existing research on prioritizing patients in EDs assumes that patients do not need the ED staff's attention during boarding time. One exception is the study conducted by De Boeck et al. (2019), which compares four ad-hoc priority policies to balance the ED physician's workload on boarding and other ED patients using simulation. They assume that boarding patients waiting to be checked up again by ED physicians cannot move to the inpatient ward before being evaluated by the ED physician. We relax this assumption in our MDP and simulation model, which is consistent with our observations in the study EDs.

The second intervention proposes a planning framework for hospitalist visits in the boarding section. Using hospital staff in EDs to provide service to boarding patients, often referred to as the "ED hospitalist role," represents one of the the operational solutions proposed by medical studies to combat the negative impacts of boarding congestion (Kathuria et al. 2010). Some medical studies report the benefits of employing the ED hospitalist role to mitigate the negative impacts of prolonged boarding time and improve ED performance in practice (e.g., Hrycko et al. (2019)). However, there is a lack of prior OM/OR literature focusing on this role in EDs, particularly regarding scheduling policies. Our study is the first to provide a scheduling framework for ED hospitalist visits.

2.3 Empirical Analyses

In this section, we begin by providing an overview of the clinical setting and providing a description of our data. Next, we examine the effect of the boarding congestion on the treatment time. Finally, we deploy several robustness tests to validate our results.

2.3.1 Clinical Context and Data Description

Our data come from eight hospitals located in a metropolitan area in Canada. These hospitals consist of three secondary hospitals and five tertiary hospitals. Tertiary hospitals are typically larger facilities with specialized equipment and inpatient wards compared to secondary hospitals (Flegel 2015). However, the patient flow in the EDs of both secondary and tertiary hospitals is similar. Figure 2.1 illustrates a typical patient flow in these emergency departments.

Upon arrival, patients are assessed by a triage nurse who measures vital signs and assigns a triage code based on the Canadian Triage and Acuity Scale (CTAS), indicating their priority. After triage, patients wait in a waiting area until they are called into a treatment room. In the treatment room, an ED physician examines the patient, determines the necessary diagnostic and treatment procedures, and may request additional tests or specialist consultations. At the end of the treatment stage, a decision is made regarding whether the patient should be discharged or admitted.

For patients requiring admission, there is a boarding stage where they wait to be assigned a bed in the requested inpatient ward. However, boarding times in the ED can become prolonged. In such cases, boarding patients may require the utilization of ED resources, including ED physicians and diagnostic tests, in addition to continuous monitoring by ED nurses (Armony et al. 2015). While admitted patients are technically under the care of hos-

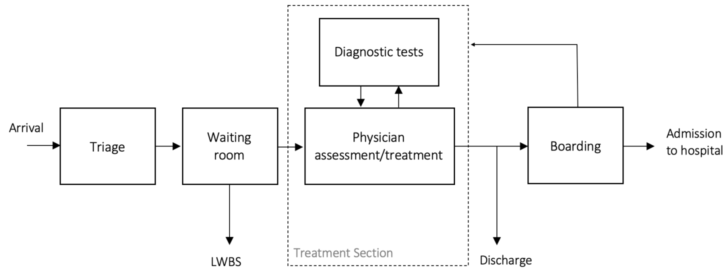


Figure 2.1: Typical Patient Flow.

pitalists, they cannot provide continuous monitoring for boarding patients in the ED. Therefore, the responsibility of providing medical attention to these patients falls upon the ED staff. Hospitalists may visit the boarding patients who are scheduled for admission to their respective inpatient wards. However, it is important to mention that there are no standardized protocols for hospitalist visits in the study hospitals. The hospitalists are informed about the number of boarding patients awaiting admission to their ward, and they choose when to visit them among their other tasks in the hospital.

In our study hospitals, long boarding times have been identified as an issue, with an average boarding time of approximately eight hours. This provides an opportunity to assess the impact of boarding congestion on ED performance. Our dataset consists of nearly 566,000 ED visits recorded over a span of 18 consecutive months. The observations in our dataset are recorded at the patient-visit level, meaning that each observation corresponds to a single patient visit to one of the eight EDs. For each visit, the dataset includes essential patient information, such as the mode of arrival (e.g., walk-ins or ambulance), demographics (e.g., age and gender), the emergency severity index measured by CTAS, diagnostic code, and major timestamps capturing critical events like triage, physician assignment, diagnostic tests, and admission. We apply a set of exclusion criteria that involve excluding incomplete ED visits, patients with triage levels 1 and 5, visits with unusual arrival modes, visits with missing information, and visits that occurred during the first and last week of the study period (for details, see Section A1.1). By applying these exclusion criteria, we obtain a final dataset comprising more than 470,000 ED visits across eight EDs.

2.3.2 Outcome and Explanatory Variables

In our analysis, the outcome variable is the treatment time, which measures the duration from when the patient is called to a treatment section until the physician makes the disposition decision. This definition aligns with previous studies in the field (e.g., Batt and Terwiesch (2017)).

To construct our explanatory variables, we leverage patient-visit level and operational data. Our primary explanatory variable is the boarding congestion experienced by each patient during their treatment time. Drawing from previous studies (Batt and Terwiesch 2017, Chan et al. 2017), we define the hourly average boarding census during the patient's treatment time as a proxy for boarding congestion. To ensure robustness, we define two alternative proxies: the boarding census and the total remaining boarding time of boarding patients at the start of treatment.

We also consider several patient-visit-specific covariates, including age, gender, mode of arrival, CTAS (triage level), assignment to a stretcher (ED bed), hospital admission, and diagnostic code (indicating the category of diagnosed diseases). Leveraging the panel structure of data, we add time and hospital fixed effects to control various dimensions of heterogeneity. For example, the staffing level plays a key role in our analyses because the impact of boarding congestion may differ depending on the number of care providers (Batt and Terwiesch 2017, Kim et al. 2015). In discussions with hospital administrators, it became evident that all EDs adhere to a specific scheduling plan. This plan entails a fixed staffing arrangement for each shift on any given day of the week and month. Consequently, to account for these consistent staffing patterns, our model incorporates time fixed effects. These effects include variables such as year, month, day of the visit, a weekend indicator, hour, and the interaction between weekend and hour variables, providing us with a way to control for variations in staff scheduling.

Considering the shared resources between the hospital and ED (McCarthy et al. 2009, Allon et al. 2013), we include the hospital occupancy level at the beginning of the patient's treatment as a control variable in our model. This variable helps account for the potential impact of hospital occupancy on both boarding congestion and treatment time. Descriptive statistics of the outcome and explanatory variables can be found in Table A1.1 in the Appendix.

2.3.3 Model Formulation

We utilize a duration regression model to examine the relationship between treatment time (TRT_TIME) and boarding congestion ($BOARD\_CGSTN$). Duration regression is a statistical modelling technique employed to predict the time or duration of an event or process based on a set of input variables or features (Greene 2012). Specifically, we employ a parametric accelerated-failure-time (AFT) model, which is a type of duration regression model that relates the logarithm of the duration to a vector of covariates and a random error term ϵ through a linear equation (Greene 2012). The general form of the model is as follows:

$$\begin{aligned} ln(TRT\_TIME_{i,h}) = \beta_1 BOARD\_CGSTN_{i,h} + \beta_2 BOARD\_CGSTN^2_{i,h} \\ + \mathbf{W}_{i,h}\theta + \mathbf{Z}_{i,h}\phi + \eta HOSP\_OCC_{i,h} + \alpha + \alpha_h + \epsilon_{i,h}, \end{aligned} \tag{2.1}$$

where the subscripts i and h denote the patient–visit pair and the ED, respectively. α is a constant term and α_h is the hospital fixed effect. The vectors $\mathbf{W}_{i,h}$ and $\mathbf{Z}_{i,h}$ contain patient-visit and time covariates, respectively. $HOSP\_OCC_{i,h}$ captures the hospital occupancy level at the start of patient i's treatment. Both the linear and squared terms of boarding congestion ($BOARD\_CGSTN$) are considered to account for the potential nonlinear effect of boarding congestion on treatment time.

In AFT models, the error term $\epsilon_{i,h}$ is typically defined as the natural logarithm of the time-to-event variable $\tau_{i,h}$ (i.e., $\epsilon_{i,h} = ln(\tau_{i,h})$). The distributional assumption of $\tau_{i,h}$ determines the underlying hazard function characteristics. After testing various distributional assumptions, we find that the Weibull distribution provides the best fit based on the Bayesian information criterion (BIC) ($BIC = 1,422,164$). We also assess variance inflation factors (VIFs) and the condition number to ensure multicollinearity is not a concern in our analysis. It is important to note that boarding congestion does not exhibit a perfect correlation with the hospital's occupancy level. This is due to the influence of various other factors, including ED admission policy, patient flow, and the occupancy level in the corresponding inpatient wards where boarding patients are expected to be admitted (McCarthy et al. 2009, Song et al. 2020, Kim et al. 2015, Mohr et al. 2020).

Despite the rich dataset, there might be unobservable factors that could simultaneously influence both treatment time and boarding congestion. For example, the skill and experience levels of the ED staff can affect both the treatment time and the boarding congestion (Hoot and Aronsky 2008). These unobservable factors introduce the possibility of bias in our estimation of β_1 and β_2 (Heckman 1998). To address this potential omitted variable bias, we enhance our previous specification by employing an instrumental variable (IV) approach, specifically the control function (CF) method. This approach has also been utilized in similar studies (e.g., Soltani et al. (2022), Kim et al. (2015)).

To establish a valid IV, two conditions must be met: (i) the IV should have a significant impact on the endogenous variable ($BOARD\_CGSTN$) (relevance condition), and (ii) the IV should be uncorrelated with unobserved factors that influence the dependent variable (TRT_TIME) (exclusion restriction) (Wooldridge 2015). We define the average boarding time of ED patients transferred to the hospital in the previous shift before the patient's arrival to the ED (LAG_BOARD_TIME) as the IV. This choice is motivated by the notion that the boarding time of recently transferred patients can serve as a reliable predictor of the current boarding time, which, in turn, drives the boarding congestion (Khare et al. 2009). Consequently, LAG_BOARD_TIME satisfies the relevance condition. Furthermore, the boarding time of patients admitted in the previous shift is unlikely to directly impact the treatment time of patients in the current shift. Given that some ED staff have changed as the shift has been changed, it is less probable that the boarding time of admitted patients during the previous period influences the behaviour of the current ED staff and subsequently affects the treatment time of current patients. However, there might be mechanisms that invalidate the exclusion restriction. We discuss these mechanisms below.

First, it is essential to investigate whether lengthy boarding times and boarding congestion in EDs can influence patient arrival patterns and potentially introduce selection bias. We explore two potential factors that may affect patient arrival patterns. Firstly, patients may strategically choose to arrive when boarding congestion is low. This is unlikely as information regarding ED and boarding congestion is not publicly available, and there is no common practice of ambulance diversion among the studied EDs. However, we evaluate this possibility by examining the relationship between ED boarding time and the arrival rate of patients with different characteristics, such as high, medium, and low triage levels. Our analysis reveals no evidence of a correlation between boarding time and patient arrival rate, indicating the absence of systematic selection bias in our IV (see Appendix A1.2 for details). Secondly, the presence of patients who leave the waiting room without being seen by an ED physician (referred to as LWBS patients) can impact the patient mix in the treatment section of the ED. If this filtering varies with boarding congestion, the patient composition during periods of high and low boarding congestion may differ. We find that 8% of patients with triage levels 2, 3, and 4 leave the ED before being called into the treatment section. However, after controlling for patient-visit, time-visit, and waiting room census, we observe that the coefficient of boarding congestion is not significantly different from zero when assessing its potential impact on abandonment probability (see Appendix A1.2 for further details). As a result, we do not need to be overly concerned about patient abandonment leading to a selection bias.

Second, we address the concern that LAG_BOARD_TIME might affect the patient's waiting time, which, in turn, can influence their treatment time and potentially compromise the validity of our IV (Wooldridge 2015). To examine the effect of LAG_BOARD_TIME on the waiting time, we use an AFT model while controlling for patient-visit and time covariates. Our analysis reveals no significant evidence of a potential effect of LAG_BOARD_TIME on waiting time (see Appendix A1.2 for more details).

Third, we consider the possibility that treatment congestion may impact boarding congestion through admission probability. To address this concern, we conduct preliminary analyses to explore the relationship between treatment and boarding congestion and the probability of admission decision. However, we do not find any evidence supporting such relationships. This observation may be attributed to the fact that, in the studied EDs, the admission decision is made by specialists rather than ED physicians. For further details, see Appendix A1.2.

Based on our analyses, we find that LAG_BOARD_TIME can be considered a valid IV in our setting. By incorporating this IV into the CF approach, we can use the following first- and second-stage equations to investigate the impact of boarding congestion on treatment time:

The first-stage specification

$$\begin{aligned} BOARD\_CGSTN_{i,h} = \alpha + \beta LAG\_BOARD\_TIME_{i,h} + \mathbf{W}_{i,h}\theta + \mathbf{Z}_{i,h}\phi \\ + \eta HOSP\_OCC_{i,h} + \alpha_h + v_{i,h} \end{aligned} \tag{2.2}$$

The second-stage specification

$$\ln(TRT\_TIME_{i,h}) = \alpha + \beta_1 BOARD\_CGSTN_{i,h} + \beta_2 BOARD\_CGSTN^2_{i,h} + \mathbf{W}_{i,h}\theta + \mathbf{Z}_{i,h}\phi + \eta HOSP\_OCC_{i,h} + \gamma\hat{v}_{i,h} + \alpha_h + \epsilon_{i,h}. \quad (2.3)$$

where $\hat{v}_{i,h}$ is the predicted residual from the first-stage specification. Note that it is sufficient to estimate the first stage of CF only for $BOARD\_CGSTN_{i,h}$ because the residuals from the first stage also adjust for the endogeneity of the quadratic term (Wooldridge 2015).

2.3.4 Results

This section provides our main empirical findings about the impact of boarding congestion on treatment time. The first- and second-stage estimation results are presented in Table 2.1. The positive and significant coefficient of the instrument from the first-stage estimation shows that the higher the boarding time of the patients admitted to the hospital before the patient arrival, the higher the degree of boarding congestion.

Table 2.1: Effects of Boarding Congestion on Treatment Time

	with IV				without IV	
	first stage		second stage			
	$BOARD\_CGSTN$		$ln(TRT\_TIME)$		$ln(TRT\_TIME)$	
LAG_BOARD_TIME	0.062***	(0.017)				
$BOARD\_CGSTN$			0.143***	(0.038)	0.097***	(0.027)
$BOARD\_CGSTN^2$			−0.023**	(0.007)	−0.025**	(0.008)
γ			−0.073***	(0.020)		
Observations	470,173		470,173		470,173	

Notes. All models include all controls, including hospital fixed effect, time fixed effects, patient covariates, and hospital occupancy. Clustered robust standard errors in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

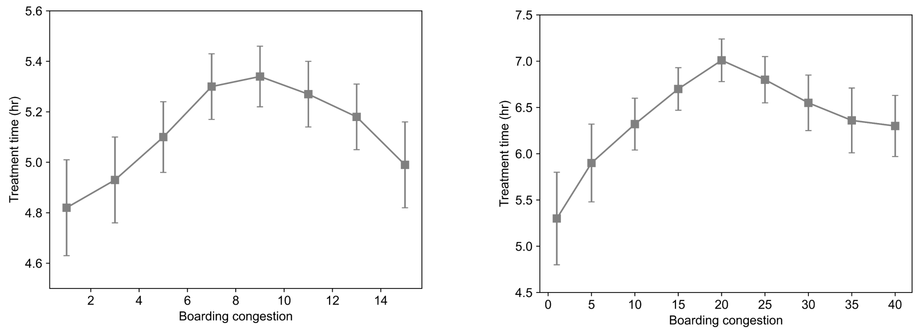
Our results from the second-stage estimation show that an inverted U-shaped relationship exists between boarding congestion and ED treatment time. The treatment time initially increases with boarding congestion ($\beta_1 = 0.143$ ($p < 0.01$)), and then decreases ($\beta_2 = -0.023$ ($p < 0.05$)). We verify the inverted U-shaped relationship by testing that the slopes at the beginning and end of the curve are indeed positive and negative, respectively (Lind and Mehlum 2010). Note that the comparison of results from estimating the baseline specification with and without IV in Table 2.1 suggests that accounting for the potential endogeneity of boarding congestion does not change the nature of the relationship between boarding congestion and treatment time.

The model's nonlinear nature and inclusion of the quadratic boarding congestion term make direct interpretation of the estimated coefficients difficult. Therefore, we depict the estimated treatment time over the boarding congestion for a tertiary and a secondary hospital ED in Figures 2.2(a) and 2.2(b), respectively, where the bars correspond to 95% confidence intervals. Both Figures 2.2(a) and 2.2(b) confirm the inverted U-shaped relationship between treatment time and boarding congestion. Figure 2.2(b) illustrates that in a large tertiary ED, as boarding congestion ranges from 1 to 20 patients per hour, the mean treatment time increases from 318 minutes to almost 420 minutes, i.e., an increase of nearly one patient per hour in boarding congestion leads to a 5-minute increase in treatment time. However, once congestion increases from 25 to 40 patients per hour, the mean treatment time decreases from 420 to 378 minutes. Our results show that this inverted U-shaped relationship between boarding congestion and treatment time exists in both tertiary and secondary EDs, suggesting that our findings can be generalizable to both types of hospitals.

2.3.5 Robustness Checks

We conduct several additional analyses to examine the robustness of our findings. In this section, we provide a summary of the results while directing readers to Section A1.3 of the Appendix for detailed information and further discussions.

First, we explore two alternative proxies to capture boarding congestion. The



(a) Secondary hospital (b) Tertiary hospital

Figure 2.2: Effects of Boarding Congestion on Treatment Time

first proxy is defined as the boarding census at the beginning of a patient's treatment. This proxy is inspired by the work of Batt and Terwiesch (2017). They utilize the waiting room census at treatment initiation to capture the workload associated with waiting patients. The second proxy is a time-based metric based on the findings of Carmen et al. (2018), which analytically show the impact of both boarding census and boarding time on ED performance. Specifically, we define the remaining boarding time of boarding patients at the start of a patient's treatment as the observed boarding congestion for that patient (similar to the measure used in queue theory by (Niu 1988)). The results obtained from these alternative proxies, as presented in Table A1.5 of the Appendix, are consistent with our main findings reported in Table 2.1, demonstrating the robustness of our conclusions under these alternative measures.

Batt and Terwiesch (2017) show that waiting room and treatment census affect treatment time. To examine the possible effect of congestion in the waiting room and treatment section, we add the linear and quadratic terms of the number of patients in these two sections at the beginning of treatment to Model 2.1, similar to the model used by Batt and Terwiesch (2017). As seen in Table A1.6 in the Appendix, the results confirm that the main conclusion regarding the impact of boarding congestion is consistent with the baseline results provided in Table 2.1.

To address the concern regarding heterogeneity across the diagnostic codes and EDs, we conduct subsampling analyses. The results of our analyses indicate that grouping observations by EDs or diagnostic codes has no discernible effect on our findings (see Table A1.7 and A1.8 in the Appendix for the results of subsampling analyses over diagnostic codes and EDs, respectively).

Finally, in the baseline analyses, we include both linear and quadratic terms of boarding congestion to allow for a nonmonotonic response to boarding congestion. To provide for further robustness test, following prior literature (Tan and Netessine 2019, Soltani et al. 2022, Kesavan et al. 2014), we conduct spline regression with two and three knots. The results confirm the inverted U-shaped relationship from thebaseline estimations (see Table A1.9 in the Appendix).

2.4 Potential Drivers of Boarding Congestion Effect: Models and Results

Building upon the patterns observed in Figure 2.2, we now explore potential mechanisms that explain the inverted U-shaped impact of the boarding congestion on the treatment time. Relying on existing literature and our discussions with clinicians at the partner hospitals, we particularly focus on (i) the additional workload on ED resources imposed by boarding patients and (ii) the hospitalists' visits triggered by boarding congestion.

2.4.1 Mechanism 1: Workload Imposed by Boarding Patients

Contrary to popular opinion, it is common for ED staff to continue providing treatment to patients in cases of long boarding times after the admission decision (Armony et al. 2015). While most of these routine checks are carried out by nurses, there are instances where boarding patients may need to return to the treatment area for reassessment (i.e., checkup) or to undergo additional diagnostic tests (Liu et al. 2011). It is important to note that "return" in this context refers to the patient's need to utilize ED resources, including ED physicians, laboratory, and imaging facilities, rather than physically returning to the treatment room. These revisits impose an additional workload on ED resources and have the potential to delay treatment procedures for other patients (Armony et al. 2015).

To examine the impact of boarding patients' returns as a mechanism underlying the effects of boarding congestion on treatment time, we estimate the following model:

$$REENTRANT\_CNGST_{i,h} = \alpha + \gamma_1 BOARD\_CNGST_{i,h} + \gamma_2 BOARD\_CNGST^2_{i,h} + \mathbf{Z}_{i,h}\phi + \alpha_h + \epsilon_{i,h}, \quad (2.4)$$

where $REENTRANT\_CNGST_{i,h}$ is the hourly average number of boarding patients returning to the treatment section during the treatment time of patient i. Note that in this model, γ_1 and γ_2 capture the effects of boarding congestion on the average number of boarding patients returning to the treatment section. As shown in Table 2.2, γ_1 and γ_2 are positive and negative, respectively ($p < 0.05$), indicating that as boarding congestion increases, the number of boarding patients returning to ED initially increases and then decreases.

Table 2.2: Estimation of Model 2.4

	$REENTRANT\_CNGST$	
$BOARD\_CNGST$	0.979^{***}	(0.230)
$BOARD\_CNGST^2$	-0.100^{**}	(0.042)
N	470,173	

Notes. The model includes all controls, including hospital fixed effect, time fixed effects, patient covariates, and hospital occupancy. Clustered robust standard errors in parentheses. $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

These findings suggest that boarding congestion contributes to the workload of ED resources by increasing the number of boarding patients who need to return to the treatment area. While this additional workload can explain the association between boarding congestion and treatment time, we still need to explore a mechanism to understand why a further increase in boarding congestion leads to a decrease in both the number of boarding patients returning to the treatment section and the ED treatment time. In the next subsection, we introduce a potential mechanism to explain the declining segment.

2.4.2 Mechanism 2: Hospitalists' Visit Schedule Triggered by Boarding Patients

As discussed in Section 2.2.2, hospitalists, who are responsible for overseeing the care of hospitalized patients, occasionally visit the boarding patients who are to be admitted to the inpatient wards where these hospitalists work. These visits can help alleviate the workload caused by boarding patients on ED resources (Hrycko et al. 2019, Chadaga et al. 2012). To examine the impact of the frequency of hospitalist visits on the boarding patients' return to the treatment section, we first investigate the relationship between boarding congestion and the likelihood of hospitalist visits. Then, we assess the relationship between hospitalist visits and the boarding patients' likelihood of returning to the treatment section.

We estimate the relationship between boarding congestion and hospitalist visits using the following model:

$$\begin{aligned} HOSP\_VISIT_{i,h} = \alpha + \gamma_1 BOARD\_CNGST_{i,h} + \gamma_2 BOARD\_CNGST_{i,h}^2 \\ + \mathbf{Z}_{i,h}\phi + \alpha_h + \epsilon_{i,h}, \end{aligned} \tag{2.5}$$

Here, $HOSP\_VISIT_{i,h}$ is a binary variable indicating whether at least one hospitalist visit occurred during patient i's treatment time in ED h[4].

Because our focus is on estimating marginal effects rather than making predictions, we opt to estimate a linear probability model instead of a binary outcome model. The results presented in Table 2.3 suggests that as the boarding congestion increases, as expected, the likelihood of hospitalist visit to the boarding section also increases.

Table 2.3: Estimation of Model 2.5

	$HOSP\_VISIT$	
$BOARD\_CNGST$	0.127***	(0.031)
$BOARD\_CNGST^2$	0.002*	(0.001)
N	470,173	

Notes. The model includes all controls, including hospital fixed effects and time fixed effects. Clustered robust standard errors in parentheses. $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Next, in order to assess the possible effect of a hospitalist's visits on $REENTRANT\_CNGST_{i,h}$, we estimate the following model:

$$\begin{aligned} REENTRANT\_CNGST_{i,h} = \alpha + \gamma_1 BOARD\_CNGST_{i,h} + \gamma_2 BOARD\_CNGST^2_{i,h} \\ + \gamma_3 HOSP\_VISIT_{i,h} + \mathbf{Z}_{i,h}\phi + \alpha_h + \epsilon_{i,h}, \end{aligned} \tag{2.6}$$

As shown in Table 2.4, the negative and significant coefficient γ_3 confirms that hospitalist visits reduce the number of boarding patients returning to the treatment area during the patient i's treatment period. This decrease in patient returns lessens the additional workload imposed on ED resources.

Boarding patients typically revisit the treatment section for two primary

[4]"Ideally, we would prefer to utilize the total count of hospitalist visits during patient i's treatment time. However, we possess only the timestamp for when a boarding patient is visited by a hospitalist, and we lack detailed information about the exact timing of each hospitalist's visit to the boarding section. This absence of precise timing data hinders our ability to calculate the frequency of visits to the boarding section accurately. Consequently, we must rely on the binary indicator variable to record the occurrence of any hospitalist visit during a patient's treatment time.

Table 2.4: Estimation of Model 2.6

	$REENTRANT\_CNGST$	
$BOARD\_CNGST$	0.764**	(0.322)
$BOARD\_CNGST^2$	−0.054*	(0.025)
$HOSP\_VISIT$	−0.021***	(0.004)
N	470,173	

Notes. The model includes all controls, including hospital fixed effects and time fixed effects. Clustered robust standard errors are provided in parentheses. $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

reasons: (i) to receive a checkup from an ED physician or (ii) to undergo additional diagnostic tests (Liu et al. 2011). To investigate the impact of hospitalist visits on each of these types of returns, we conduct an analysis. We provide the technical details behind this analysis in Appendix A1.4. Here, we summarize the key insights of this analysis. Our findings indicate that when boarding patients are visited by hospitalists, their probability of needing to be seen by ED physicians decreases. Additionally, we observe that boarding patients who are visited by ED physicians have a higher probability of experiencing diagnostic tests during their boarding time compared to those visited by hospitalists[5]. This observation can be attributed to the increased work pressure experienced by ED physicians during periods of high boarding congestion, leading to an increased cognitive load and a more risk-averse mindset (Soltani et al. 2022). Furthermore, the awareness that their work during the boarding period will be reviewed by hospitalists may further enhance the physician's risk aversion, resulting in a higher likelihood of ordering diagnostic tests for boarding patients (Deck and Jahedi 2015). ED physicians may also rely on ordering diagnostic tests as an alternative to direct patient contact when they have limited time available to spend with

[5]We also investigated whether boarding congestion influences the diagnostic test-ordering behaviour of ED physicians for patients in the treatment section. However, we did not find any evidence of changes in physicians' diagnostic test-ordering behaviour for patients in the treatment section in response to changes in boarding congestion.

boarding patients (Batt and Terwiesch 2017).

In contrast, when hospitalists visit boarding patients who are admitted to their respective inpatient wards, they typically have a higher level of expertise and familiarity with the patient's health issues. As a result, hospitalists may order fewer tests, which aligns with previous studies indicating that experienced physicians often demonstrate a more selective approach in ordering diagnostic tests (Ma et al. 2005).

In summary, an increase in boarding congestion triggers hospitalist visits, which helps relieve ED resources from the extra workload imposed by the boarding patients, leading to a decrease in the average treatment time in the ED. These findings suggest that managing boarding patients' return to the treatment section and scheduling hospitalists' visits are important operational levers to mitigate the adverse impact of boarding congestion. We operationalize these strategies in the next section.

2.5 Operational Interventions

The results obtained from empirical analyses suggest two operational interventions to mitigate the impact of boarding congestion: (i) prioritizing patients in the treatment section and (ii) streamlining operational policies for hospitalists' visits to the boarding section. Regarding the first intervention, in Section 2.5.1, we introduce an MDP model and extract the optimal policy's properties through numerical analysis. Leveraging insights from the MDP model, we propose a practical heuristic policy to control patient flow in the treatment section. Moreover, by employing a simulation model based on real data, we demonstrate the benefits of our policy. Regarding the latter intervention, we employ our simulation model to investigate alternative hospitalist visit policies and illustrate the advantages of different policies.

2.5.1 Patient Flow Control in Treatment Section

This section presents our MDP model for addressing the patient flow control problem in the treatment area, along with the simulation model we develop to evaluate the efficacy of the proposed heuristic policy in a real-world context.

Model Framework.

We formulate the problem of patient flow control as a continuous-time, infinite-horizon average cost MDP. Our MDP model aims to strike a balance between the waiting time cost of boarding patients and emergency patients when both require the same resources within the ED. In this section, we provide a brief overview of the problem setting and direct readers to Section A1.6 for a detailed specification of the system state, action, transition dynamics, cost, and objective function.

EDs are inherently complex, involving multiple stages and services, which can present challenges when trying to model the precise patient flow using an MDP. Hence, we begin by developing a simplified model and then leverage insights from the model to propose a heuristic policy. Then, we evaluate the effectiveness of our proposed policy in a realistic scenario through simulation analysis. Figure 2.3 illustrates the simplified patient flow model in the treatment section of EDs. Specifically, our system comprises two distinct queues: a treatment queue for emergency patients who have already started their treatment and await its completion and a checkup queue for boarding patients waiting to return to the treatment section.

In our model, we assume that emergency patients arrive according to a Poisson process with the rate λ_r per unit time, and their treatment time is exponentially distributed with the rate μ_r. After treatment completion, the service provider decides whether to admit the patient to the hospital (with a probability p) or discharge her (with probability $1-p$). If admitted, the patient is transferred to the boarding section, and we assume her waiting time for a hospital bed follows an exponential distribution with a rate of μ_b. While

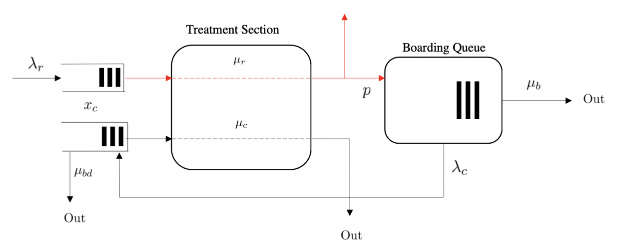


Figure 2.3: Boarding Patient Flow Control System.

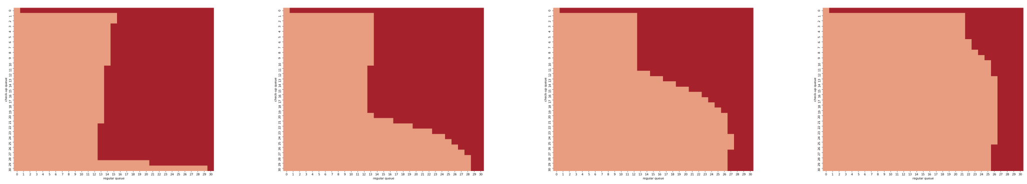
waiting for admission, boarding patients may need to revisit the treatment section. Such requests arrive according to a Poisson process with rate λ_c, and the service time for returned boarding patients follows an exponential distribution with rate μ_c. If a hospital bed becomes available, a patient in the checkup queue may leave. Thus, we assume that patients in the checkup queue depart according to an exponential distribution with a rate of μ_{bd}. In this setting, once a service is completed, a decision is made regarding which queue is chosen to serve the next patient.

Structure of Optimal Policy.

Finding the optimal policy form for the MDP model described above is challenging due to its inherent complexity, which arises from the combination of a tandem and a reentrant queue. While the structural properties of tandem queue systems have been theoretically analyzed in related literature (Koole 2004), studies focusing on queue systems with reentrance often rely on numerical methods to evaluate their structural properties (Koole and Pot 2006). Therefore, in alignment with previous research, we use numerical methods to determine the optimal policy for the MDP model.

To achieve this, we construct a comprehensive test suite encompassing a wide range of parameter values (further details of the test suite are provided in Appendix A1.6). For each instance in our test suite, we use the value iteration algorithm to solve the MDP and record the optimal actions at each state.

Figure 2.4 exhibits switching curves as a function of x_r (i.e., the treatment



(a) $x_b = 0$ (b) $x_b = 10$ (c) $x_b = 20$ (d) $x_b = 30$

Figure 2.4: Optimal Actions - Base Case

queue in the x-axis) and x_c (i.e., the checkup queue in the y-axis) for four levels of the boarding census, x_b. Note that the top left corner of each figure represents the case of no patients in both queues, and dark red regions indicate the optimal policy that prioritizes emergency patients over boarding patients. These results highlight that the optimal actions are influenced by both the number of emergency and boarding patients awaiting treatment or checkups, as well as the presence of boarding patients who are not in the checkup queue.

Although the complexity of the optimal policy structure is evident in these results, they provide valuable insights and suggest specific forms of prioritization policy, as presented in Figure 2.5.

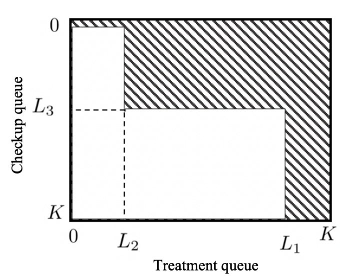


Figure 2.5: The Approximation of the Optimal Policy Structure

As one might expect, the precise form of the optimal policy is intricate and varies according to the model parameters. However, we can estimate the optimal action structure using the following threshold policy (TP), which is shown to be optimal under a special case (see Section A1.6 for more details):

$$\text{TP} = \begin{cases} \text{Choose from treatment queue} & \text{if } x_r \geq L_1 \\ \text{Choose from treatment queue} & \text{if } L_2 \leq x_r < L_1, \quad x_c \leq L_3 \\ \text{Choose from checkup queue} & \text{Otherwise.} \end{cases} \tag{2.7}$$

where L_1, L_2, and L_3 show thresholds that are functions of the system parameters. While this policy provides an approximation of the optimal patient flow control, we demonstrate in the subsequent section, through a simulation model, that it can be significantly beneficial in practice.

Simulation.

We now focus on assessing the performance of our proposed policy in a real-world scenario using a simulation model. We use our estimation results and historical data to calibrate the simulation model. We develop two simulation models: one based on data from a secondary hospital and another one based on data from a tertiary hospital. We evaluate our policies based on three metrics: average treatment time per patient, hourly average treatment section census, and total treatment time. We aim to determine whether the proposed patient control policy outperforms the existing policy.

We simulate an ED comprising two sections: a treatment section and a boarding section (similar to Figure 1, excluding triage and waiting room). Patient arrival to the treatment section follows an empirical distribution derived from the available data. We estimate the treatment time distribution by utilizing an empirical distribution based on the data and our findings from Section 2.4.2. Therefore, the treatment time for patients in the treatment section is prolonged when a boarding patient returns to the treatment section. Upon completing the treatment process, patients are either assigned to the boarding section or exit the system. The admission rate is determined based on the available data. Boarding patients' checkup request times and boarding times are determined using empirical distributions. Subsequently, we utilize the data to calculate the probability of a checkup request. For a checkup request, three possible events can occur based on the control policies: (i) the request is accepted, and the patient reenters the treatment section; (ii) the request is rejected, and the patient remains in the boarding section; or (iii) the patient is placed in a virtual queue (referred to as the checkup queue) to reenter the treatment section at the appropriate time.

The timestamp at which a boarding patient returns to the boarding section was not recorded in our available data. Therefore, the checkup duration cannot be directly derived from the data. To address this, we compare the simulation results of several potential distribution functions with the data to select the best distribution function for the checkup time.

Once the checkup is completed, patients return to the boarding section and wait until their admission time. Upon reaching their admission time, they are admitted to the hospital and leave the ED system. Boarding patients whose admission time has arrived when they are in the treatment section are admitted to the hospital immediately after returning to the boarding section, just as they would in reality. A patient can leave the checkup queue before visiting the treatment section if her admission time arrives while she is waiting in the checkup queue.

We verify and validate our simulation model's accuracy using the strategies proposed by Sargent (2010), Shechter (2010), Werker et al. (2009). We record the time of each patient's main events, which helps us verify that (i) patients follow a reasonable path in the system and (ii) the correct number of patients are flowing through the system. We validate the simulation model by comparing it with an actual "real-world" ED system. By examining key system features such as the average (hourly) census in both the treatment and boarding sections and the volume of boarding patients, we demonstrate the alignment between our simulation results and the observations in the selected hospital. For detailed information, refer to Table A1.20 and Figure A1.4.

When formulating our proposed TP, we estimate its parameters L1, L2, and

L3 using a simulation-based optimization technique. The objective of this optimization is to minimize the combined average treatment time and waiting time of boarding patients in the checkup queue. Alternatively, we can define the objective function as minimizing the total average census of both the treatment section and checkup queues.

Considering the potential challenges in implementing boarding-congestion-dependent values for L1, L2, and L3 in practical settings, we propose an alternative approach. In this approach, we calculate L-values for low, moderate, and high congestion scenarios in the boarding section. Firstly, we determine the first and third quartiles of the boarding congestion distribution. Subsequently, we compute the optimal values of L1, L2, and L3 for three different scenarios: boarding congestion below the first quartile, between the first and third quartiles, and above the third quartile. This approach allows us to adapt the TP based on the different levels of congestion in the boarding section, providing a more flexible and practical solution.

In order to assess the effectiveness of TP, we consider two benchmark scenarios. Firstly, we examine the existing situation in which boarding patients are given priority over emergency patients in utilizing ED resources. We refer to this scenario as the "status quo." Secondly, we evaluate a situation in which boarding patients are redirected to a completely separate system, effectively eliminating any additional burden they may impose on ED resources. This scenario represents a lower bound for the treatment time. We refer to this scenario as the "lower bound." By comparing TP with these benchmark scenarios, we can evaluate its performance and understand its potential for improvement.

Results: To evaluate the different policies, we conduct a simulation over the course of one full year, with a warm-up period of one month and 1,000 iterations. The results for the tertiary hospital are summarized in Table 2.5 (refer to Table A1.22 for the results of the secondary hospital).

The first row of the table displays the outcomes obtained under the current policy, which align well with the observed data. A comparison between the status quo and the lower bound reveals that boarding patients can potentially increase the treatment time by approximately 15%, highlighting the need for ED administrators to consider the additional workload caused by boarding patients when making operational decisions, such as staffing levels.

Our analysis strongly supports the effectiveness of the proposed TP in mitigating the negative impact of boarding congestion on treatment time. Specifically, when implementing TP, we observe a reduction of four patients in the treatment section (corresponding to a 66.67% improvement in treatment time congestion) and a decrease of two patient years on average in total treatment time (representing a 45.32% improvement) compared to the current policy.

Table 2.5: Simulation Results of Boarding Patient Flow Control Policies

Policy	Average treatment time (minute)	Hourly average treatment section census	Total treatment time* (patient-years)
Status quo (SQ)	378.84 (0.48)	40 (0.14)	44.64 (0.05)
Lower Bound (LB)	330.40 (0.39)	34 (0.13)	38.95 (0.04)
Threshold Policy (TP)	361.92 (0.42)	36 (0.13)	42.63 (0.04)
	34.92%	66.67%	45.32%

Notes. Pct. improvement compared to the status quo and lower bound presented in the last row (i.e., (SQ-TP)/(SQ-LB)·100). Standard errors in parentheses for 1000 iterations. * Total treatment time is calculates as (average treatment time · total annual patient volume)/(60 · 24 · 365)

2.5.2 Hospitalist Visit Scheduling (HVS)

In Section 2.4, we demonstrate the beneficial impact of hospitalist visits on relieving ED resources from the increased workload caused by boarding patients. Typically, hospitalists are called to the ED based on the number of boarding patients requiring admission to their inpatient wards. Nevertheless, the timing of these visits is often influenced by the hospitalists' other responsibilities within the wards. To address this issue, we propose and evaluate two policies for hospitalist visits in this section (for scheduling policies used in other healthcare systems, refer to e.g., Robinson and Chen (2003, 2010)).

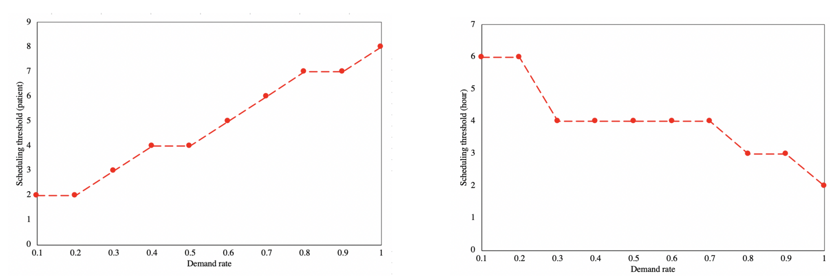
One of the challenges faced by operational policies in this context is the heterogeneity of hospitalists across different inpatient wards. The number of patients requiring hospitalist visits varies significantly depending on the type or specialty of the ward. Additionally, hospitalists may have different preferences when it comes to scheduling. They often have busy schedules at their respective facilities, and checking on boarding patients may disrupt their rounds and other activities. Moreover, some hospitalists may prefer fixed visit times rather than irregular visits based on the census in the boarding section (Zhu 2018).

To address these challenges, we consider two types of policies for hospitalist visits: census-based and time-based policy. The census-based policy suggests that hospitalists should visit the boarding section when a certain number of boarding patients waiting to be admitted to their ward is reached. This number is referred to as the "census-based scheduling threshold." On the other hand, in the time-based policy, hospitalists visit the boarding section at predetermined intervals (e.g., every 6 hours), regardless of the census. This predetermined interval is called the "time-based scheduling threshold." Using simulation-based optimization, we determined the optimal level for census-based scheduling threshold and time-based scheduling threshold for census-based and time-based policy, respectively. Our objective function considers two primary goals: (i) to minimize the additional workload on the treatment section imposed by boarding patients (i.e., to maximize the number of boarding patients visited by a hospitalist in the early hours of boarding time before the checkup request), and (ii) to minimize the frequency of hospitalist visits. Therefore, we introduce a new key performance metric alongside the metrics discussed in the previous section: the daily average number of hospitalist visits.

To account for the heterogeneity of patient volume across different inpatient wards, we determine the optimal scheduling threshold for each policy based on varying demand rates. This approach empowers hospitalists to select the most suitable policy, whether it is time-based or census-based scheduling, taking into consideration their preferences and the specific demand rate of their respective inpatient ward. Through this framework, we effectively facilitate policy selection, streamlining the decision-making process.

Results.

This section presents the results using the historical data of a tertiary hospital. The findings for the secondary hospital, provided in A1.7, align closely with those presented here. Our proposed framework is visually summarized in Figures 2.6(a) and 2.6(b). With this framework, EDs can seamlessly choose the optimal policy by taking into account the demand rate and hospitalists' preferences regarding time- and census-based scheduling.



(a) Census-based scheduling framework (b) Time-based scheduling framework

Figure 2.6: A Tertiary ED's Scheduling Framework for an ED-hospitalist.

Table 2.6 shows the results for the "status quo" and "threshold policy (TP).[6]" A comparison between time-based and census-based policies reveals that, despite hospitalists' preference for fixed visit times in the time-based policy, the census-based policy yields fewer hospitalist trips to the boarding section and outperforms the time-based policy in terms of average treatment time and census.

Implementing such an operational policy can mitigate the adverse impact of

[6]Threshold policy refers to the patient flow control intervention discussed before.

boarding patients on the treatment section by 48% with two fewer patient-years of total treatment time and five fewer patients in the treatment section on average. By combining both interventions, the improvement can reach 68%, translating to 3.89 fewer patient-years and six fewer patients in the estimated current system, in terms of total treatment time and census, respectively. It is important to highlight that these improvements can be achieved while keeping the expected daily number of hospitalist visits to the ED almost the same (i.e., 2.97 vs. 2.91). Furthermore, while the mean waiting time of boarding patients for a checkup under TP (with the existing pattern of hospitalists' visits) is 39 minutes, it is reduced to 4 minutes with the combined approach of TP and census-based visit policy.

Table 2.6: Simulation Results of Hospitalist Visit Scheduling (HVS) Policies

	Average treatment time (minute)		Hourly average treatment section census		Daily average number of hospitalist visit per category*	
Policy	Status quo	TP	Status quo	TP	Status quo	TP
Without HVS**	378.84 (0.48)	361.92 (0.42)	40 (0.14)	37 (0.12)	2.97 (0.02)	2.97 (0.02)
Time-based HVS	363.46 (0.44)	347.08 (0.39)	35 (0.12)	34 (0.11)	3.11 (0.02)	3.19 (0.02)
Census-based HVS	361.93 (0.42)	345.77 (0.39)	35 (0.12)	34 (0.11)	2.86 (0.01)	2.91 (0.01)
Mixed HVS	362.29 (0.40)	346.11 (0.38)	35 (0.13)	34 (0.13)	3.04 (0.01)	3.17 (0.02)

Notes. * We consider four categories. Standard errors in parentheses. 1000 iterations.

Lastly, we assess the efficacy of the proposed interventions considering the characteristics of the emergency department (ED). Our primary focus revolves around the arrival rate of emergency patients, treatment time, and the probability of boarding for patients. We consider three levels for each characteristic: a low scenario (25% less than the base case), a medium scenario (base case), and a high scenario (25% more than the base case). Here, we summarize our key findings as follows.

Observation 1. Table 2.7 suggests that both census-based and time-based HVS, as well as the TP, exhibit greater improvements as the arrival rate or treatment time of patients increases. This suggests that congested EDs with higher arrival rates or longer treatment times can benefit more from the proposed interventions. This observation aligns with intuition, as the effectiveness of these policies increases with the level of congestion in the ED. By effectively preventing the overflow of boarding patients into the treatment section, these policies alleviate the additional burden on ED resources.

Observation 2. As arrival rate and/or treatment time increase, the difference between census- and time-based policies and the additional benefit from HVS policy on top of TP reduces. This phenomenon occurs due to the increased congestion in the ED when the arrival rate and treatment time are high. In such scenarios, both census-based and time-based policies recommend a higher number of visits for hospitalists. Consequently, the additional advantages offered by census-based HVS diminish in these demanding settings.

Observation 3. We observe that as the boarding probability increases, the effectiveness of both interventions (HVS and TP) decreases. However, this decrease is more pronounced when only the TP is in place. This suggests that HVS policies are more robust in mitigating the negative effects of boarding congestion across different situations. The underlying intuition behind this observation is that as the boarding probability increases, prioritizing boarding patients over emergency patients (which is the current practice) becomes the optimal policy in most instances. Therefore, the relative improvement offered by the TP over the status quo diminishes. This observation has potentially important implications for identifying the scenarios in which HVS and/or TP are most effective. It indicates that congested EDs that primarily handle less acute or complex patients, such as those in primary or secondary hospitals like community hospitals, are more likely to benefit from such policies compared to EDs with a higher proportion of acute cases, such as tertiary hospitals and trauma centers.

In summary, this simulation analysis provides two valuable managerial insights. First, by improving patient flow control and effectively managing

Table 2.7: Percentage Improvements in based on ED Characteristics

					Time-based HVS + TP			Census-based HVS + TP		
Arrival	Treatment	Boarding percentage			Boarding percentage			Boarding percentage		
Rate	Time	Low	Med	High	Low	Med	High	Low	Med	High
Low	Low	25.56%	21.36%	10.18%	51.86%	49.48%	44.47%	54.60%	52.72%	46.69%
	Med	29.71%	23.46%	13.06%	65.96%	61.36%	50.99%	68.45%	63.78%	53.48%
	High	39.42%	27.89%	18.07%	73.90%	64.42%	59.62%	77.83%	67.26%	62.22%
Med	Low	31.79%	24.40%	18.24%	55.22%	51.64%	46.66%	58.03%	53.94%	48.51%
	Med	48.84%	34.92%	24.59%	72.74%	65.26%	58.74%	76.47%	68.27%	61.01%
	High	53.73%	37.24%	27.98%	74.56%	71.16%	62.87%	78.35%	73.82%	65.01%
High	Low	41.61%	30.74%	22.06%	61.30%	58.71%	51.78%	63.76%	60.62%	53.35%
	Med	57.85%	39.12%	27.57%	73.89%	66.89%	60.13%	76.75%	69.15%	61.78%
	High	60.85%	41.12%	29.01%	76.42%	72.43%	63.57%	79.32%	74.38%	65.21%

resources such as hospitalists, the adverse effects of boarding congestion on ED treatment time can be significantly mitigated without the need for additional financial investments. Second, the characteristics of ED can influence the effectiveness of proposed interventions in alleviating the negative impact of boarding congestion on treatment time.

2.6 Discussion and Conclusion

Using a comprehensive dataset comprising eight different EDs (three secondary and five tertiary), we examine the impact of boarding congestion on treatment time. Previous studies have often assumed that boarding patients simply occupy ED beds and potentially hinder access for patients awaiting treatment, focusing on waiting time or LOS as metrics. In contrast, our study specifically examines treatment time and reveals a nuanced relationship between boarding congestion and treatment duration, characterized by an inverted U-shape pattern. For instance, in a typical secondary ED, increasing boarding congestion from one patient to nine patients per hour results in an average treatment time increase of 10%. However, this positive relationship between boarding congestion and treatment time only persists up to a certain threshold, beyond which treatment time starts to decrease with further increases in boarding congestion. We demonstrate the robustness of our findings through a range of robustness tests, including the use of instrumental variables, alternative proxy definitions, and subsampling analyses.

We also investigate the potential mechanisms underlying the relationship between boarding congestion and treatment time. Our analysis reveals that boarding congestion generally places an additional workload on ED resources, as boarding patients may require further checkups by ED physicians or undergo additional diagnostic tests. This finding aligns with the observations made by Armony et al. (2015), who estimated that approximately 10% of ED physicians' workload is attributed to boarding patients. Furthermore, we show that the likelihood of hospitalist visits to the boarding section increases as boarding congestion rises. This, in turn, leads to a decrease in the number of visits made by boarding patients back to the treatment section. This explains the decreasing segment of the inverted U-shaped relationship between boarding congestion and treatment time in the ED.

Based on empirical analyses, we propose two interventions to mitigate the impact of boarding congestion on treatment time: boarding patient flow control and hospitalist visit scheduling policies. The simulation results illustrate that implementing these two interventions together can reduce the effects of boarding congestion by 68% for a tertiary hospital.

2.6.1 Implications for Theory

Our study has several implications for healthcare operations management theory. The first implication is for staff levels in EDs. Our study corroborates the descriptive findings in Armony et al. (2015) regarding the additional workload on the ED staff resulting from congestion caused by boarding patients. Although patient boarding is a common issue in EDs globally (Khare et al. 2009), we do not find any OM/OR study quantifying the impact of the congestion caused by boarding patients on treatment time and considering it in ED staffing decisions. Previous studies have generally assumed that admitted patients promptly leave the ED, regardless of whether they are immediately transferred to the inpatient ward or remain in the ED awaiting admission. However, our findings exhibit the significance of the boarding congestion effect, emphasizing the need for researchers to consider the presence of boarding patients when planning for staffing levels.

The second implication relates to patient flow in EDs. Previous studies on patient flow typically categorize patients into two groups: ”new patients” waiting for treatment and ”in-process patients” already in the treatment section. However, our analysis across eight EDs reveals that approximately 60% of boarding patients require additional ED resources beyond regular nurse checkups during their boarding period. These findings highlight the importance of accounting for boarding patients in ED patient flow models in future studies.

Thirdly, our study holds implications for scholars interested in examining the role of hospitalists in EDs. While hospitalists play a crucial role in mitigating the adverse effects of boarding congestion on ED performance and patient outcomes (Chadaga et al. 2012), we are not aware of any OR/MS study that specifically focuses on this role or addresses operational questions related to it. This study proposes an easy-to-implement framework for selecting the optimal HVS policy for a given hospitalist based on their inpatient ward's demand volume. This opens up a promising avenue of research on the role of hospitalists in the ED and its impact on both ED and hospital performance.

2.6.2 Implications for Practice

Our study also has implications for practice. First, by analyzing the data of more than 470,000 patients from eight EDs, our study highlights the importance of patient boarding in EDs and shows that the adverse impact of boarding congestion on the ED's performance is far beyond what was documented previously. These findings emphasize how important it is for hospital administrators to carefully characterize boarding congestion for their ED operations. Ignoring this impact on treatment times when making operational decisions may result in persistent overcrowding and an inability to respond to patients with quality and timely care.

Financial incentives and the volatility of demand for ED services inevitably lead to the phenomenon of patient boarding in some EDs. This study proposes two easily implementable interventions to mitigate the negative impact of congestion caused by boarding patients on treatment time. These interventions are based on optimizing the use of ED and hospital resources and do not require major resource allocation or structural changes. ED managers can consider these interventions and customize them according to the size and complexity of the patient mix in their EDs to alleviate the adverse effects of boarding patient congestion on ED performance.

Our findings suggest that prioritizing boarding patients over emergency patients for using ED resources can be effective. However, this policy is not universally optimal for all EDs. Figure 2.7 summarizes the results of our sensitivity analyses (the details are provided in Section A1.6), which suggest that prioritizing boarding patients is best suited for EDs with (i) a high arrival rate of checkup requests from boarding patients, (ii) a high probability of emergency patients to be admitted to the hospital, and (iii) a longer duration of checkup time for boarding patients in the treatment section. In general, our results indicate that EDs handling a more acute and/or complex patient mix, including patients with lower triage codes, more comorbidities, or older ages, can improve their ED performance by prioritizing boarding patients. Conversely, EDs managing less acute or severe cases, such as community hospitals, may consider prioritizing emergency patients to improve their ED performance.

Regarding our second intervention, our analyses reveal that census-based HVS consistently outperforms time-based HVS across various ED characteristics. However, the advantages of this policy become more prominent as the ED experiences higher levels of congestion, while its mitigating effect diminishes as the probability of patient boarding increases. Thus, we can infer

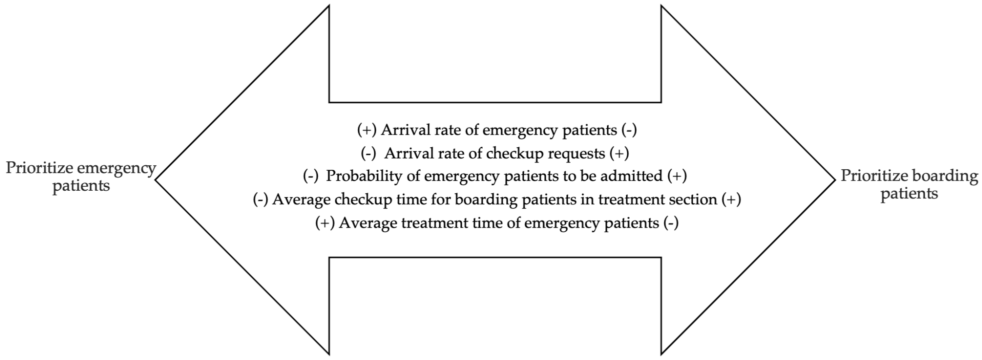


Figure 2.7: Patient Flow Control Based on Key Characteristics of the ED

that highly congested EDs in primary or secondary hospitals can benefit most from this intervention, particularly when combined with TP. Future research can delve into exploring novel interventions to address patient boarding and its impact on ED performance and patient outcomes.

Finally, the hospital admission team faces a dilemma when deciding whether to assign a patient to a nonprimary bed (off-service placement) when all primary beds are occupied or wait until a primary bed becomes available, thus leaving a nonprimary bed unused (Dai and Shi 2019). Given that off-service placement has negative impacts on patient outcomes and hospital length of stay, some studies recommend assigning patients to primary beds, even if it results in prolonged boarding time (Song et al. 2020). By quantifying the effects of boarding congestion on ED performance, our study provides valuable assistance to the hospital admission team in navigating this trade-off. However, further research is needed to develop a framework that guides practitioners in determining the optimal timing for off-service placement based on the current state of boarding and the hospital.

Bibliography

Agency for Healthcare Research and Quality (2018). The need to address emergency department crowding.

Allon, G., Deo, S. and Lin, W. (2013). The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence, *Operations Research* **61**(3): 544–562.

American College of Emergency Physicians (2018). Boarding of admitted and intensive care patients in the emergency department.

Apker, J., Mallak, L. A. and Gibson, S. C. (2007). Communicating in the "gray zone": perceptions about emergency physician–hospitalist handoffs and patient safety, *Academic Emergency Medicine* **14**(10): 884–894.

Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y. and Yom-Tov, Y. T. G. (2015). On patient flow in hospitals: A data-based queueing-science perspective, *Stochastic Systems* **5**(1): 146–194.

Batt, R. and Terwiesch, C. (2017). Early task initiation and other load-adaptive mechanisms in the emergency department., *Management Science* **63**(11): 3531–3551.

Berry Jaeker, J. and Tucker, A. (2017). Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity, *Management Science* **63**(4): 1042–1062.

Berry Jaeker, J. and Tucker, A. (2019). The value of process friction: The role of justification in reducing medical costs, *Journal of Operations Management* **66**(1-2): 12–34.

Carmen, R., Nieuwenhuyse, I. V. and Houdt, B. V. (2018). Inpatient boarding in emergency departments: Impact on patient delays and system capacity, *European Journal of Operational Research* **271**(3): 953–967.

Chadaga, S., Shockley, L., Keniston, A., Klock, N., Dyke, S. V., Davis, Q. and Chu, E. (2012). Hospitalist-led medicine emergency department team: Associations with throughput, timeliness of patient care, and satisfaction, *Journal of Hospital Medicine* **7**(7): 562–566.

Chan, C. W., Farias, V. F. and Escobar, G. J. (2017). The impact of delays on service times in the intensive care unit, *Management Science* **63**(7): 2049–2072.

Chisholm, C., Dornfeld, A., Nelson, D. and Cordell, W. (2001). Work interrupted: a comparison of workplace interruptions in emergency departments and primary care offices, *Annals of Emergency Medicine* **38**(2): 146–151.

Coil, C. J., Flood, J. D., Belyeu, B. M., Young, P., Kaji, A. H. and Lewis, R. J. (2016). The effect of emergency department boarding on order completion, *Annals of Emergency Medicine* **67**(6): 730–736.

Dai, J. and Shi, P. (2019). Inpatient overflow: An approximate dynamic programming approach, *Manufacturing & Service Operations Management* **21**(4): 894–911.

De Boeck, K., Carmen, R. and Vandaele, N. (2019). Needy boarding patients in emergency departments: An exploratory case study using discrete-event simulation, *Operations Research for Health Care* **21**: 19–31.

Deck, C. and Jahedi, S. (2015). The effect of cognitive load on economic decision making: A survey and new experiments, *European Economic Review* **78**: 97–119.

Feizi, A., Carson, A., Jaeker, J. B. and Baker, W. E. (2023). To batch or not to batch? impact of admission batching on emergency department boarding time and physician productivity, *Operations Research* **71**(3): 939–957.

Flegel, K. (2015). Tertiary hospitals must provide general care.

Freeman, M., Savva, N. and Scholtes, S. (2017). Gatekeepers at work: An empirical analysis of a maternity unit, *Management Science* **63**(10): 3147–3167.

Greene, W. H. (2012). *Econometric Analysis*, 7th edn, Prentice Hall, Upper Saddle River, NJ.

He, S., Sim, M. and Zhang, M. (2019). Data-driven patient scheduling in emergency departments: A hybrid robust-stochastic approach, *Management Science* **65**(9): 4123–4140.

Heckman, J. (1998). Dummy endogenous variables in a simultaneous equation system, *Econometrica* **46**(4): 931–959.

Helm, J. E., AhmadBeygi, S. and Van Oyen, M. P. (2011). Design and analysis of hospital admission control for operational effectiveness, *Production and Operations Management* **20**(3): 359–374.

Hillier, D. F., Parry, G. J., Shannon, M. W. and Stack, A. M. (2009). The effect of hospital bed occupancy on throughput in the pediatric emergency department, *Annals of Emergency Medicine* **53**(6): 767–776.

Hoot, N. and Aronsky, D. (2008). Systematic review of emergency department crowding: causes, effects, and solutions, *Annals of Emergency Medicine* **52**(2): 126–136.

Howell, E., Bessman, E., Marshall, R. and Wright, S. (2010). Hospitalist bed management effecting throughput from the emergency department to the intensive care unit, *Journal of Critical Care* **25**(2): 184–189.

Hrycko, A., Tiwari, V., Vemula, M., Donovan, A., Scibelli, C., Joshi, K., Visintainer, P. and Stefan, M. (2019). A hospitalist-led team to manage patient boarding in the emergency department: Impact on hospital length of stay and cost, *Southern Medical Journal* **112**(12): 599–603.

Kathuria, N., Jagoda, A., Baumlin, K., Hill, S., Mumm, L., Jervis, R., Briones, A., Dunn,

A. and Markoff, B. (2010). A model of a hospitalist role in the care of admitted patients in the emergency department, *Journal of Hospital Medicine* **5**(6).

Kc, D. (2014). Does multitasking improve performance? evidence from the emergency department., *Manufacturing & Service Operations Management* **16**(2): 168–183.

Kc, D., Staats, B., Kouchaki, M. and Gino, F. (2020). Task selection and workload: A focus on completing easy tasks hurts performance, *Management Science* **66**(10): 4397–4416.

Kc, D. and Terwiesch, C. (2009). Impact of workload on service time and patient safety: An econometric analysis of hospital operations, *Management Science* **55**(9): 1486–1498.

Kesavan, S., Staats, B. and Gilland, W. (2014). Volume flexibility in services: The costs and benefits of flexible labor resources, *Management Science* **60**(8): 1884–1906.

Khare, R. K., Powell, E. S., Reinhardt, G. and Lucenti, M. (2009). Adding more beds to the emergency department or reducing admitted patient boarding times: which has a more significant influence on emergency department congestion?, *Annals of Emergency Medicine* **53**(5): 575–585.

Kim, S., Chan, C., Olivares, M. and Escobar, G. (2015). Icu admission control: An empirical study of capacity allocation and its implication for patient outcomes, *Management Science* **61**(1): 19–38.

Kim, S., Pinker, E., Rimar, J. and Bradley, E. (2017). Refining workload measure in hospital units: From census to acuity-adjusted census in intensive care units, *USC Marshall School of Business, Working Paper* .

Koole, G. (1998). Structural results for the control of queueing systems using event-based dynamic programming, *Queueing Systems* **30**(3-4): 323–339.

Koole, G. (2004). Convexity in tandem queues, *Probability in the Engineering and Informational Sciences* **18**(1): 13–31.

Koole, G. and Pot, A. (2006). Workload minimization in re-entrant lines, *European Journal of Operational Research* **174**(1): 216–233.

Lind, J. and Mehlum, H. (2010). With or without u? the appropriate test for a u-shaped relationship, *Oxford Bulletin of Economics and Statistics* **72**(1): 109–118.

Lippman, S. (1975). Applying a new device in the optimization of exponential queuing systems, *Operations Research* **23**(4): 687–710.

Liu, S. W., Singer, S. J., Sun, B. C. and Camargo Jr, C. A. (2011). A conceptual model for assessing quality of care for patients boarding in the emergency department: structure–process–outcome, *Academic Emergency Medicine* **18**(4): 430–435.

Lucas, R., Farley, H., Twanmoh, J., Urumov, A., Olsen, N., Evans, B. and Kabiri, H. (2009). Emergency department patient flow: the influence of hospital census variables on emergency department length of stay, *Academic Emergency Medicine* **16**(7): 597–602.

Luo, W., Cao, J., Gallagher, M. and Wiles, J. (2013). Estimating the intensity of ward admission and its effect on emergency department access block, *Statistics in Medicine* **32**(15): 2681–2694.

Ma, O. J., Gaddis, G., Steele, M. T., Cowan, D. and Kaltenbronn, K. (2005). Prospective analysis of the effect of physician experience with the fast examination in reducing the use of ct scans, *Emergency Medicine Australasia* **17**(1): 24–30.

McCarthy, M., Zeger, S., Ding, R., Levin, S., Desmond, J., Lee, J. and Aronsky, D. (2009). Crowding delays treatment and lengthens emergency department length of stay, even among high-acuity patients, *Annals of Emergency Medicine* **54**(4): 492–503.

McKenna, P., Heslin, S. M., Viccellio, P., Mallon, W. K., Hernandezand, C. and Morley, E. J. (2019). Emergency department and hospital crowding: causes, consequences, and cures, *Clinical and Experimental Emergency Medicine* **6**(3): 189–195.

Mohr, N., Wessman, B., Bassin, B., Elie-Turenne, M. C., Ellender, T., Emlet, L., Ginsberg, Z., Gunnerson, K., Jones, K. M., Kram, B., Marcolini, E. and Rudy, S. (2020). Boarding of critically ill patients in the emergency department, *Critical Care Medicine* **48**(8): 1180–1187.

Morley, C., Unwin, M., Peterson, G. M., Stankovich, J. and Kinsman, L. (2018). Emergency department crowding: a systematic review of causes, consequences and solutions, *PloS one* **13**(8): e0203316.

Niu, S. (1988). Representing workloads in gi/g/1 queues through the preemptive-resume lifo queue discipline, *Queueing Systems* **3**(2): 157–178.

Pines, J. (2017). What cognitive psychology tells us about emergency department physician decision-making and how to improve it, *Academic Emergency Medicine* **24**(1): 117–119.

Porteus, E. (2002). Foundations of stochastic inventory theory, *Stanford University Press* .

Powell, E. S., Khare, R. K., Venkatesh, A. K., Van Roo, B. D., Adams, J. G. and Reinhardt, G. (2012). The relationship between inpatient discharge timing and emergency department boarding, *Journal of Emergency Medicine* **42**(2): 186–196.

Puterman, M. . (1994). Markov decision processes: discrete stochastic dynamic programming, *John Wiley & Sons* .

Rabin, E., Kocher, K., McClelland, M., Pines, J., Hwang, U., Rathlev, N., Asplin, B., Trueger, N. and Weber, E. (2012). Solutions to emergency department 'boarding' and crowding are underused and may need to be legislated, *Health Affairs* **31**(8): 1757–1766.

Ranjan, C., Paynabar, K., Helm, J. E. and Pan, J. (2017). The impact of estimation: A new method for clustering and trajectory estimation in patient flow modeling, *Production and Operations Management* **26**(10): 1893–1914.

Robinson, L. W. and Chen, R. R. (2003). Scheduling doctors' appointments: optimal and empirically-based heuristic policies, *Iie Transactions* **35**(3): 295–307.

Robinson, L. W. and Chen, R. R. (2010). A comparison of traditional and open-access policies for appointment scheduling, *Manufacturing & Service Operations Management* **12**(2): 330–346.

Saghafian, S., Austin, G. and Traub, S. (2015). Operations research/management contributions to emergency department patient flow optimization: Review and research prospects, *IIE Transactions on Healthcare Systems Engineering* **5**(2): 101–123.

Saghafian, S., Hopp, W., Oyen, M. V., Desmond, J. and Kronick, S. (2012). Patient streaming as a mechanism for improving responsiveness in emergency departments, *Operations Research* **60**(5): 1080–1097.

Saghafian, S., Hopp, W., Oyen, M. V., Desmond, J. and Kronick, S. (2014). Complexity-augmented triage: A tool for improving patient safety and operational efficiency, *Manufacturing & Service Operations Management* **16**(3): 329–345.

Sargent, R. (2010). Verification and validation of simulation models, *In Proceedings of the 2010 winter simulation conference 2010, IEEE* p. 166:183.

Shechter, S. (2010). Monte carlo simulation as an aid for deciding among treatment options, *Wiley Encyclopedia of Operations Research and Management Science* .

Sills, M. R., Fairclough, D., Ranade, D. and Kahn, M. G. (2011). Emergency department crowding is associated with decreased quality of care for children with acute asthma, *Annals of Emergency Medicine* **57**(3): 191–200.

Soltani, M., Batt, R., Bavafa, H. and Patterson, B. (2022). Does what happens in the ed stay in the ed? the effects of emergency department physician workload on post-ed care use, *Manufacturing & Service Operations Management* .

Song, H., Tucker, A., Graue, R., Moravick, S. and Yang, J. (2020). Capacity pooling in hospitals: The hidden consequences of off-service placement, *Management Science* **66**(9): 3825–3842.

Staats, B. and Gino, F. (2012). Specialization and variety in repetitive tasks: Evidence from a japanese bank, *Management Science* **58**(6): 1141–1159.

Tan, T. and Netessine, S. (2014). Specialization and variety in repetitive tasks: Evidence from a japanese bank, *Management Science* **60**(6): 1574–1593.

Tan, T. and Netessine, S. (2019). When you work with a superman, will you also fly? an empirical study of the impact of coworkers on performance, *Management Science* **65**(8): 3495–3517.

Timm, N. L., Ho, M. L. and Luria, J. W. (2008). Pediatric emergency department overcrowding and impact on patient flow outcomes, *Academic Emergency Medicine* **15**(9): 832–837.

Viccellio, P., Zito, J. A., Sayage, V., Chohan, J., Garra, G., Santora, C. and Singer, A. J. (2013). Patients overwhelmingly prefer inpatient boarding to emergency department boarding, *Journal of Emergency Medicine* **45**(6): 942–946.

Wang, J. and Zhou, Y. (2018). Impact of queue configuration on service time: Evidence from a supermarket, *Management Science* **64**(7): 3055–3075.

Werker, G., Sauré, A., French, J. and Shechter, S. (2009). The use of discrete-event simulation modelling to improve radiation therapy planning processes, *Radiotherapy and Oncology* **92**(1): 76–82.

White, B. A., Biddinger, P. D., Chang, Y., Grabowski, B., Carignan, S. and Brown, D. F. (2013). Boarding inpatients in the emergency department increases discharged patient length of stay, *Journal of Emergency Medicine* **44**(1): 230–235.

Wolff, R. (1989). Stochastic modelling and the theory of queues, *Pearson College Division* .

Wooldridge, J. (2015). Control function methods in applied econometrics, *Journal of Human Resources* **50**(2): 420–445.

Yoon, P., Steiner, I. and Reinhardt, G. (2003). Analysis of factors influencing length of stay in the emergency department, *Canadian Journal of Emergency Medicine* **5**(3): 155–161.

Zhu, C. (2018). *Three essays on data-driven models in health care operations management*, Doctoral dissertation, McGill University (Canada).

A1 Appendix for Article 1

A1.1 Sample Selection and Descriptive Statistics

We impose a set of exclusion criteria on the data for our analyses. First, we exclude the observations in which patients leave before seeing an ED physician. They accounted for approximately 5% of ED visits. We exclude these observations because of an incomplete visit. Second, we exclude patients encounter with a CTAS code of 1 and 5, which constitute approximately 1.5% and 9.9% of total visit volume, respectively. The patients with a CTAS level of 1 are always given high priority because of their elevated risks; they are generally tracked separately from other patients. Usually, patients with CTAS 5 are also tracked separately, because their treatment needs are relatively simpler and they only have one-time interaction with the ED physician for more than 95% of cases. Next, we eliminate patients who come by helicopters or are escorted by police, because different protocols were followed for these patients. Finally, we exclude the ED visits with missing information (less than 1% of the observations). We also exclude the data of first and last weeks to avoid censored estimation of occupancy. The final dataset includes more than 470,000 ED visits over eight EDs. See Figure A1.1 for the summary of the data-extraction procedure. We provide the descriptive statistics in Table A1.1.

A1.2 Details on Main Model

Patient Characteristics and Boarding Time

To address concerns of selection bias in our instrumental variable (IV) analysis, we conduct several analyses. First, we demonstrate significant variation in the hourly rate of ED arrivals at different levels of average boarding time. As shown in Figure A1.2, no consistent pattern is found between ED arrivals and average boarding time.

Second, we assess the similarity of patients visiting the ED during different average boarding times by examining their observable characteristics including age, gender,

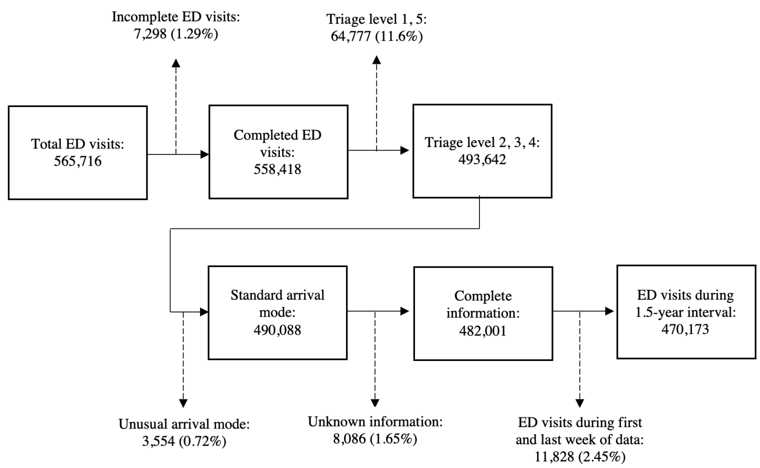


Figure A1.1: Selection of the emergency department (ED) visit sample.

triage level, diagnostic code, and arrival mode. In Figure A1.3, we present the arrival rates for different patient groups across four different levels of boarding time. Furthermore, we conduct a multivariate analysis of variance (MANOVA) to test for differences in patient characteristics and severity levels among those visiting the ED during different levels of average boarding time. The analysis reveals no significant differences in patient characteristics among the different levels of the average boarding time (Wilks' $\lambda = 0.96$, $p = 0.14$).

Boarding Time and Waiting Time

In this section, we estimate the potential relationship between the proposed instrumental variable (LAG_BOARD_TIME) on waiting time. To assess this relationship, we use a parametric accelerated-failure-time (AFT) model, which is formulated as follows:

$$\begin{aligned} ln(WAIT\_TIME_{i,h}) = \alpha + \beta_1 LAG\_BOARD\_TIME_{i,h} + \beta_2 LAG\_BOARD\_TIME^2_{i,h} \\ + \mathbf{W}_{i,h}\theta + \mathbf{Z}_{i,h}\phi + \alpha_h + \epsilon_{i,h}, \end{aligned} \tag{A1.1}$$

Table A1.2 presents the results indicating that the average boarding time of pa-

Table A1.1: Summary Statistics

	Secondary			Tertiary				
	H1	H2	H3	H4	H5	H6	H7	H8
N	50,362	39,302	23,688	72,684	78,154	111,693	50,742	43,548
Age	51.93 (0.096)	48.67 (0.114)	49.18 (0.148)	53.41 (0.085)	49.65 (0.092)	52.72 (0.064)	50.90 (0.091)	49.81 (0.105)
Female	0.37 (0.002)	0.47 (0.003)	0.46 (0.003)	0.45 (0.002)	0.45 (0.002)	0.44 (0.001)	0.45 (0.002)	0.50 (0.002)
Arrival type: Ambulant	0.77 (0.002)	0.70 (0.002)	0.80 (0.003)	0.68 (0.002)	0.71 (0.002)	0.82 (0.001)	0.75 (0.002)	0.61 (0.002)
Triage 2	0.08 (0.001)	0.15 (0.002)	0.04 (0.001)	0.28 (0.002)	0.29 (0.002)	0.24 (0.001)	0.18 (0.002)	0.25 (0.002)
Triage 3	0.45 (0.002)	0.45 (0.003)	0.31 (0.003)	0.46 (0.002)	0.43 (0.002)	0.44 (0.001)	0.45 (0.002)	0.40 (0.002)
Admission probability	0.14 (0.002)	0.13 (0.002)	0.08 (0.002)	0.23 (0.002)	0.26 (0.002)	0.16 (0.001)	0.17 (0.002)	0.18 (0.002)
Reentrance probability*	0.41 (0.002)	0.17 (0.001)	0.38 (0.002)	0.59 (0.003)	0.29 (0.001)	0.59 (0.002)	0.38 (0.002)	0.32 (0.002)
Diagnostics ordered	5.73 (0.027)	1.28 (0.009)	1.92 (0.022)	4.87 (0.017)	1.8 (0.010)	4.26 (0.013)	4.16 (0.023)	4.08 (0.023)
Treatment time (minute)	333 (1.863)	385 (2.513)	310 (3.122)	372 (1.462)	406 (1.875)	320 (1.128)	434 (2.124)	471 (2.204)
Boarding congestion	5.45 (0.013)	4.65 (0.018)	3.61 (0.012)	24.81 (0.026)	25.59 (0.032)	21.93 (0.022)	9.47 (0.017)	6.72 (0.016)
Hospital occupancy level	0.88 (0.007)	0.82 (0.008)	0.79 (0.01)	0.91 (0.006)	0.89 (0.008)	0.93 (0.007)	0.89 (0.01)	0.90 (0.008)

Notes. * Contingent on being admitted. Standard errors in parentheses

tients admitted to the hospital during the previous shift prior to patient arrival does not have a significant impact on the patient's waiting time.

Table A1.2: Effect of Boarding Time on Waiting Time

	(1)	
LAG_BOARD_TIME	0.002	(0.001)
$LAG\_BOARD\_TIME^2$	0.001	(0.001)
Observations	470,173	

Notes. The model includes all controls including hospital fixed effect, time fixed effects, and patient covariates. Clustered robust standard errors in parentheses. $*p < 0.1$; $**p < 0.05$; $***p < 0.01$.

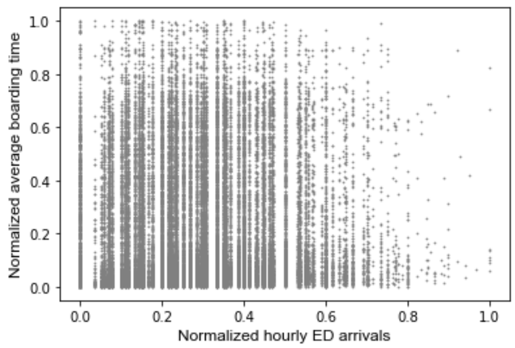


Figure A1.2: The Relationship between ED Arrivals and Average Boarding Time

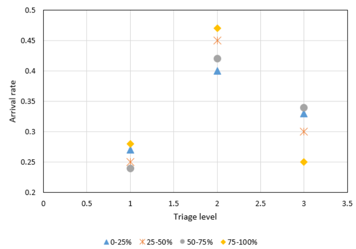
Admission Probability and ED Congestion

Here, we examine the possible relationship between the workload in the treatment section and admission probability. Specifically, we estimate the following equation:

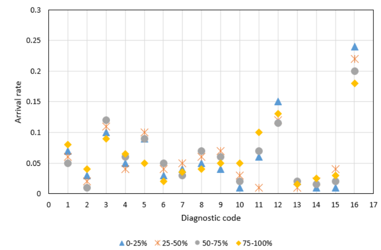
$$ADMIT_{i,h} = \alpha_h + \beta_1 WORKLOAD_{i,h} + \beta_2 WORKLOAD_{i,h}^2 + \mathbf{W}_{i,h}\theta + \mathbf{Z}_{i,h}\phi + \epsilon_{i,h} \tag{A1.2}$$

where $WORKLOAD_{i,h}$ is defined as the treatment section census divided by the number of available ED physicians when treatment begins for patient i. Since we are interested in the marginal effects of workload on hospital admission rather than the hospital admission prediction, we use a linear probability model rather than a binary outcome model. Table A1.3 presents the results indicating that the physician's workload does not have a significant effect on the admission probability. These findings differ from previous literature (e.g., Pines (2017), Freeman et al. (2017)). One possible explanation for this discrepancy is that in the eight study EDs, specialists, rather than ED physicians, possess the authority to make admission decisions.

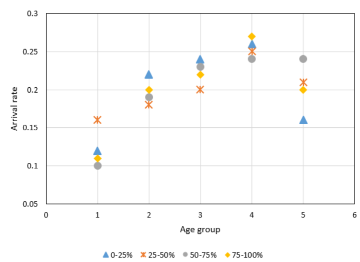
Using Model A1.2, we also examine the potential impact of boarding congestion on the admission probability (we use variable $BOARD\_CNGST$ instead of $WORKLOAD$ in Model A1.2) . However, as shown in Table A1.3, we do not find any evidence of a significant effect of boarding congestion on the admission



(a) Triage-level groups



(b) Diagnostic-code groups



(c) Age groups

Figure A1.3: Patient Arrival Rate

probability.

Boarding Congestion and Abandonment Behavior

To evaluate the possible effect of the boarding congestion on the patients' abandonment behavior, we use the following model:

$$\begin{aligned} LWBS_{i,h} = \alpha + \beta_1 BOARD\_CNGST_{i,h} + \beta_2 BOARD\_CNGST^2_{i,h} \\ + \mathbf{W}_{i,h}\theta + \mathbf{Z}_{i,h}\phi + \alpha_h + \epsilon_{i,h}, \end{aligned} \tag{A1.3}$$

The vectors $\mathbf{W}_{i,h}$ and $\mathbf{Z}_{i,h}$ contain patient-visit and time covariates, respectively. The vectors $\mathbf{W}_{i,h}$ also includes waiting time and waiting room census observed by i in ED h. Because we are interested in marginal effects rather than predictions, we choose to estimate a linear probability model rather than a binary outcome model. The results presented in Table A1.4 indicate that boarding congestion does not

Table A1.3: Effects of Boarding Congestion on Treatment Time

	(1) $ADMIT$		(2) $ADMIT$	
$WORKLOAD$	0.0034	(0.0035)		
$WORKLOAD^2$	-0.0002	(0.0003)		
$BOARD\_CNGST$			-0.0004	(0.0003)
$BOARD\_CNGST^2$			-0.0000	(0.0000)
N	470,173			
R^2	0.294		0.294	

Notes. the models includes all controls including hospital fixed effect, time fixed effects, and patient covariates. Clustered robust standard errors in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

have a significant impact on patients' abandonment behavior.

Table A1.4: Estimation of Model A1.3

	$LWBS$	
$BOARD\_CNGST$	0.015	(0.011)
$BOARD\_CNGST^2$	0.001	(0.001)
N	470,173	

Notes. The model includes all controls including hospital fixed effect and time fixed effects. Clustered robust standard errors in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

A1.3 Robustness Check and Sensitivity Analysis

Alternative Proxies for Boarding Congestion

Table A1.5: Boarding Congestion's Effect on Treatment Time (Alternative Proxies for Boarding Congestion).

	(1)		(2)	
Boarding congestion proxy	Boarding census upon start of treatment		Total remaining boarding time upon start of treatment	
$BOARD\_CNGSTN$	0.043^{***}	(0.007)	0.021^{**}	(0.008)
$BOARD\_CNGSTN^2$	-0.006^{**}	(0.002)	-0.003^{**}	(0.000)
Observations	470,173		470,173	

Notes. The model includes all controls including hospital fixed effect, time fixed effects, and patient covariates. Clustered robust standard errors in parentheses. $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Boarding Congestion's Effects on Treatment Time by Considering Other Census Variables

We investigate the impact of boarding congestion on treatment time by incorporating the census of the waiting room and treatment section[7]. Specifically, we introduce the linear and quadratic terms of the number of patients in the waiting room and treatment section once treatment begins for patient i into Model 2.1, following a similar approach to Batt and Terwiesch (2017). As presented in Table A1.6, the effect of boarding congestion on treatment time remains robust even after accounting for the number of patients in the waiting room and treatment section.

[7]Note that the treatment census does not include boarding patients who return to the treatment section for a checkup.

Table A1.6: Boarding Congestion's Effects on Treatment Time by Considering ED Census.

	(1)		(2)		(3)	
$BOARD\_CNGSTN$	0.143^{***}	(0.038)	0.139^{***}	(0.036)	0.139^{***}	(0.035)
$BOARD\_CNGSTN^2$	-0.023^{**}	(0.007)	-0.024^{**}	(0.007)	-0.024^{**}	(0.007)
$WAIT\_CENSUS$			-0.016	(0.026)	-0.016	(0.025)
$WAIT\_CENSUS^2$			0.001	(0.005)	0.001	(0.005)
$TRTMT\_CENSUS$					0.013	(0.017)
$TRTMT\_CENSUS^2$					-0.000	(0.001)
Observations	470,173		470,173		470,173	

Notes. the models include all controls including hospital fixed effect, time fixed effects, patient covariates, and hospital occupancy level. Clustered robust standard errors in parentheses. $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Subsample Analysis

We conduct subsample analyses based on diagnostic codes and individual hospitals. The estimated coefficients from these subsamples are presented in Table A1.7, demonstrating that the effect of boarding congestion on treatment time remains consistent with our baseline estimation results discussed in Section 2.3.

A1.4 Impact of Hospitalist Visit on Boarding Patients

To examine the impact of hospitalist visits on boarding patients' need to be visited by ED physicians, we estimate the following model:

$$CHECKUP_{j,h} = \alpha + \beta_1 HOSPITALIST\_VISIT_{j,h} + \mathbf{V}_{j,h}\gamma + \mathbf{W}_{j,h}\theta + \mathbf{Z}_{j,h}\phi + \alpha h + \epsilon_{j,h}, \tag{A1.4}$$

Here, the subscripts j and h indicate the boarding patient j and hospital h, respectively. $CHECKUP_{j,h}$ is a binary variable indicating whether boarding patient j has returned to the treatment section to be visited by an ED physician. $HOSPITALIST\_VISIT_{j,h}$ is a binary variable indicating whether patient j has

Table A1.7: Boarding Congestion's Effect on Treatment Time based on Diagnostic Codes.

	(1) Most common diagnostic code		(2) Least common diagnostic code	
$BOARD\_CNGSTN$	0.104^{***}	(0.026)	0.207^{**}	(0.061)
$BOARD\_CNGSTN^2$	-0.026^{**}	(0.008)	-0.041^{**}	(0.012)
Observations	138,964		12,985	

Notes. the models include all controls including hospital fixed effect, time fixed effects, patient covariates, and hospital occupancy level. Clustered robust standard errors in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

been visited by a hospitalist before being seen by an ED physician. The vector $\mathbf{V}_{j,h}$ contains census covariates, including the boarding census and treatment census at the beginning of the boarding time for patient j. The vectors $\mathbf{W}_{j,h}$ and $\mathbf{Z}_{j,h}$ contain patient-visit and time covariates, respectively.

Because our interest lies in examining the marginal effects rather than making predictions, we opt to estimate a linear probability model instead of a binary outcome model. The results shown in Table A1.10 reveal that the estimated coefficient β_1 is negative and statistically significant. This indicates that boarding patients who are visited by a hospitalist during their boarding time are less likely to require a visit by an ED physician during their boarding time.

We examine whether there is a correlation between boarding patients who are visited by an ED physician and their likelihood of receiving diagnostic tests, as opposed to those who are visited by a hospitalist. To explore this association, we use the following model for estimation:

$$\begin{aligned} TEST_{j,h} = \alpha + \beta_1 PHYSICIAN\_VISIT_{j,h} + \beta_2 HOSPITALIST\_VISIT_{j,h} \\ + \mathbf{V}_{j,h}\gamma + \mathbf{W}_{j,h}\theta + \mathbf{Z}_{j,h}\phi + \alpha h + \epsilon_{j,h}, \end{aligned} \tag{A1.5}$$

Table A1.8: Boarding Congestion's Effect on Treatment Time in Each Hospital.

Hospital	Observations	$BOARD\_CNGSTN$		$BOARD\_CNGSTN^2$	
1	50,362	0.082***	(0.027)	−0.038***	(0.010)
2	72,684	0.103***	(0.031)	−0.054***	(0.018)
3	78,154	0.093**	(0.032)	−0.028**	(0.014)
4	111,693	0.128***	(0.042)	−0.025***	(0.007)
5	39,302	0.124**	(0.051)	−0.085**	(0.037)
6	50,742	0.085***	(0.027)	−0.036***	(0.011)
7	43,548	0.096**	(0.033)	−0.012**	(0.005)
8	23,688	0.056***	(0.014)	−0.016**	(0.006)

Notes. the models include all controls including time fixed effects, patient covariates, and hospital occupancy level. Robust standard errors are clustered by diagnostic codes in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Where $TEST_{j,h}$ is a binary variable indicating whether boarding patient j experienced a diagnostic test during her boarding time. $PHYSICIAN\_VISIT_{j,h}$ and $HOSPITALIST\_VISIT_{j,h}$ are binary variables indicating whether boarding patient j was visited by an ED physician and hospitalist, respectively. The vector $\mathbf{V}j, h$ includes census covariates and boarding time of patient j. The vectors $\mathbf{W}j, h$ and $\mathbf{Z}j, h$ contain patient-visit and time covariates, respectively. The error term is denoted as $\epsilon_{j,h}$.

The coefficients β_1 and β_2 quantify the effects of each type of visit on the likelihood of receiving diagnostic tests during the boarding time. The results, presented in Table A1.11, indicate that when boarding patients are visited by an ED physician, they have a higher probability of receiving additional diagnostic tests during their boardin time compared to those visited by a hospitalist.

A1.5 Diagnostic Tests

To explore the potential impact of boarding congestion on the number of diagnostic tests ordered during treatment time, we employ a zero-inflated negative binomial (ZINB) model. The ZINB model combines a binary logit process for inflation and

Table A1.9: Spline Regressions - Piecewise Linear Function

	(1) Two knots		(1) Three knots	
$BOARD\_CNGSTN1$	0.031^{***}	(0.008)	0.044^{**}	(0.018)
$BOARD\_CNGSTN2$	-0.083^{***}	(0.022)	-0.015	(0.010)
$BOARD\_CNGSTN3$			-0.183^{***}	(0.017)
Observations	470,173		470,173	

Notes. the models include all controls including hospital fixed effect, time fixed effects, patient covariates, and hospital occupancy level. Clustered robust standard errors in parentheses. $*p < 0.1$; $**p < 0.05$; $***p < 0.01$.

Table A1.10: Estimation of Model A1.4

	$CHECKUP$	
$HOSPITALIST\_VISIT$	-0.0203^{*}	(0.0089)
N	84,932	

Notes. The model includes all controls including hospital fixed effect and time fixed effects. Clustered robust standard errors in parentheses. $*p < 0.1$; $**p < 0.05$; $***p < 0.01$.

a negative binomial count process for the count of tests ordered, as our dependent variable is discrete and relatively small. Let $f_{inf}(.)$ represent the probability density of the logit process and $f_{count}(.)$ denote the probability density of the negative binomial count process. The ZINB model defines the density function as follows:

$$f(y|\mathbf{x}) = \begin{cases} f_{inf}\left(1|\mathbf{x}\right) + \left\{1 - f_{inf}\left(1|\mathbf{x}\right)\right\} f_{count}\left(0|\mathbf{x}\right) & \text{if } y = 0 \\ \{1 - f_{inf}\left(1|\mathbf{x}\right)\} f_{count}\left(y|\mathbf{x}\right) & \text{if } y \geq 1 \end{cases} \tag{A1.6}$$

Hence, the mean of the model is given by:

$$E[DIAG\_TEST_i|\mathbf{x}_i] = \frac{1}{1 + \exp\left(\mathbf{x}_i \eta_{inf}\right)} \times \exp\left(\mathbf{x}_i \eta_{count}\right), \tag{A1.7}$$

Here, $\mathbf{x}_i \eta_{inf}$ and $\mathbf{x}_i \eta_{count}$ represent the predicted values for the inflation and count processes, respectively, using the same set of explanatory variables as in Equation

Table A1.11: Estimation of Model A1.5

	$TEST$	
$PHYSICIAN\_VISIT$	0.0233^{***}	(0.0021)
$HOSPITALIST\_VISIT$	-0.0184^{**}	(0.0071)
N	84,932	

Notes. The model includes all controls including hospital fixed effect and time fixed effects. Clustered robust standard errors in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

2.1. The equation can be written as:

$$\begin{aligned} \mathbf{x}_{i,h}\eta_j &= \alpha_j + \eta_{j,1}BOARD\_CGSTN_{i,h} + \eta_{j,2}BOARD\_CGSTN_{i,h}^2 \\ &\quad + \beta HOSP\_OCC_{i,h} + \mathbf{W}_{i,h,j}\theta_j + \mathbf{Z}_{i,h,j}\phi_j + \alpha_h \quad \text{for } j = inf, count. \end{aligned} \tag{A1.8}$$

Table A1.12 provides a summary of the results from Model A1.8. Due to the ZINB model's nonlinear and two-part nature, direct interpretation of the coefficients is challenging. Therefore, the bottom panel of Table A1.12 presents the mean marginal effect of the variables of interest. Our findings indicate that the number of diagnostic tests ordered remains unaffected by boarding congestion.

A1.6 Patient Flow Control in Treatment Section

Model Formulation

We formulate the patient flow control as a continuous-time dynamic program, which let us limit our attention only to times of a change in the state of the system (Puterman 1994). The time horizon is considered infinite, which is consistent with the idea of running an ED.

State Variable: The system's state includes information about the number of patients in the treatment queue, boarding section, and checkup queue. Let $\mathbf{x} = (x_r, x_b, x_c)^T$ represent the system's state, where x_r, x_b, and x_c correspond to the number of patients in the treatment queue, boarding section, and checkup queue,

Table A1.12: Impact of Boarding Congestion on the Number of Diagnostic Tests Ordered

	(1)	
Count process		
$BOARD\_CNGSTN$	0.001	(0.007)
$BOARD\_CNGSTN^2$	-0.000	(0.000)
Inflation process		
$BOARD\_CNGSTN$	-0.069^{**}	(0.028)
$BOARD\_CNGSTN^2$	0.001^{**}	(0.000)
Marginal effect		
$BOARD\_CNGSTN$	0.028	(0.027)
Observations	470,173	
Non-zero obs.	336,624	
Zero obs.	133,549	

Notes. the models includes all controls including hospital fixed effect, time fixed effects, patient covariates, and hospital occupancy level. Clustered robust standard errors in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

respectively. We assume that the capacity of each queue or section is limited to K, and thus $\mathbf{x} \in \mathcal{S}(\mathbf{x}) = (x_r, x_b, x_c)|x_r, x_b, x_c \leq K$. Consequently, the state space is finite.

Actions: The possible actions are to serve the next patient from either the treatment or checkup queue. Therefore, we can define actions as follows:

$$\mathcal{A}(\mathbf{x}) = \{a \in \{0, 1\}|a \leq x_r\} \tag{A1.9}$$

The variable a is a binary variable, where its value of "1" represents choosing a patient from the treatment queue. Clearly, this can happen only where there is at least one patient in this queue. Thus, the constraint $a \leq x_r$ forces a to be equal to 0 when there is no patient waiting for treatment.

Transition Probabilities: Let T denote the random time between two decision points. Because all the events follow Poisson processes, T is exponentially distributed (Porteus 2002). The distribution rate is the sum of all rates, i.e., $v(\mathbf{x}) = \lambda_r + \max\{\mu_r, \mu_c\} + x_b\lambda_c + x_b\mu_b + x_c\mu_{bd}$. Also, when a transition occurs at time t, the probability that the transition is caused by a specific event is the rate of the event divided by the sum of all rates. This probability is independent of the time that has passed. Because the system's state changes over time, the transition rate in each period is not constant. To transform such a system into a Markov chain with a uniform transition rate, we apply the *uniformization* technique (Lippman 1975). While using this technique, we note that an upper bound for the transition rate is $v^{\max} = \lambda_r + \max\{\mu_r, \mu_c\} + K\lambda_c + K\mu_b + K\mu_{bd}$. We define the discrete time parameters after uniformization corresponding to the transition probabilities in the embedded discrete time Markov chain (DTMC) as follows: $\lambda'_r = \lambda_r/v^{\max}$, $\mu'_r = \mu_r/v^{\max}$, $\mu'_c = \mu_c/v^{\max}$, $\lambda'_c = \lambda_c/v^{max}$, $\lambda'_{db} = \lambda_{db}/v^{max}$and $\mu'_b = \mu_b/v^{max}$.

Let A, D_r, T_{rb}, T_{bc}, D_c, and D_b be events representing an emergency patient arrival at the treatment queue, a discharge of an emergency patient, a patient's transfer from the treatment section to the boarding section, a patient's transfer from the boarding section to the checkup queue, a discharge/leave from the checkup queue, and finally a discharge from the boarding section. These operators are defined as follows:

$$\begin{aligned}
A\mathbf{x} &= (x_r + 1, x_b, x_c) \\
D_r\mathbf{x} &= (\max\{x_r - 1, 0\}, x_b, x_c) \\
T_{rb}\mathbf{x} &= \begin{cases} (x_r - 1, x_b + 1, x_c) & \text{if } x_r \geq 1 \\ \mathbf{x} & \text{if } x_r = 0 \end{cases} \\
T_{bc}\mathbf{x} &= \begin{cases} (x_r, x_b - 1, x_c + 1) & \text{if } x_b \geq 1 \\ \mathbf{x} & \text{if } x_b = 0 \end{cases} \\
D_c\mathbf{x} &= (x_r, x_b, \max\{x_c - 1, 0\}) \\
D_b\mathbf{x} &= (x_r, \max\{x_b - 1, 0\}, x_c)
\end{aligned} \tag{A1.10}$$

With this notation, the one-step state transition probabilities after uniformization are given by the following:

$$p(\mathbf{x}_{k+1}|\mathbf{x}_k, a_k) = \begin{cases} \lambda'_r, & \text{if } \mathbf{x}_{k+1} = A\mathbf{x}_k \\ a(1-p)\mu'_r, & \text{if } \mathbf{x}_{k+1} = D_r\mathbf{x}_k \\ ap\mu'_r, & \text{if } \mathbf{x}_{k+1} = T_{rb}\mathbf{x}_k \\ x_b\lambda'_c, & \text{if } \mathbf{x}_{k+1} = T_{bc}\mathbf{x}_k \\ x_c\mu'_{db} + (1-a)\mu'_c, & \text{if } \mathbf{x}_{k+1} = D_c\mathbf{x}_k \\ x_b\mu'_b, & \text{if } \mathbf{x}_{k+1} = D_b\mathbf{x}_k \\ 1 - \left(\lambda'_r + a\mu'_{rt} + x_b\lambda'_c + x_c\mu'_{db} + (1-a)\mu'_c + x_b\mu'_b\right), & \text{if } \mathbf{x}_{k+1} = \mathbf{x}_k \end{cases} \tag{A1.11}$$

Bellman Optimality Equation: Let α be the continuous-time discount factor in the original problem and ξ be an exponential random variable with rate v_{max} (the length of time for one transition in the discrete chain). Then, the equivalent discrete time discount factor for the uniformized model becomes the following:

$$\beta = E[e^{\alpha\xi}] = \int_0^\infty \left(e^{-\alpha t}\right)\left(v_{max}e^{-v_{max}t}\right) dt = \frac{v_{max}}{v_{max} + \alpha}. \tag{A1.12}$$

Let γ_r be the per unit time cost associated with holding a patient in the treatment queue, and γ_c be the per unit time cost associated with holding a patient in the checkup queue. In addition, assume that h_r and h_c are the costs associated with reaching the treatment and checkup queues' capacity, respectively. When the treatment queue and checkup queue reach capacity, the ED cannot accept any patient in the treatment section; thus, h_r and h_c show the costs that ensue from blocking the treatment section. As with the original problem, consider α the continuous-time discount factor and let ξ be an exponential random variable with rate v_{max}. We can write the discount time instantaneous one-stage cost as follows:

$$\begin{aligned} C(\mathbf{x}) &= E\bigg[\int_0^{\xi}(\gamma_r x_r + \gamma_c x_c + h_r \mathbb{I}_{\{x_r=K\}} + h_c \mathbb{I}_{\{x_c=K\}})e^{-\alpha t}dt\bigg] \\ &= \frac{1-\beta}{\alpha}(\gamma_r x_r + \gamma_c x_c + h_r \mathbb{I}_{\{x_r=K\}} + h_c \mathbb{I}_{\{x_c=K\}}) \\ &= \frac{\gamma_r}{v_{max}+\alpha}x_r + \frac{\gamma_c}{v_{max}+\alpha}x_c + \frac{h_r}{v_{max}+\alpha}\mathbb{I}_{\{x_r=K\}} + \frac{h_c}{v_{max}+\alpha}\mathbb{I}_{\{x_c=K\}} \\ &= \gamma_r^{'} x_r + \gamma_c^{'} x_c + h_r^{'} \mathbb{I}_{\{x_r=K\}} + h_c^{'} \mathbb{I}_{\{x_c=K\}}, \end{aligned} \tag{A1.13}$$

where $\gamma_r^{'}$, $\gamma_c^{'}$, $h_r^{'}$, and $h_c^{'}$ are continuous-time costs.

Now, we can formulate a recursive optimality equation that considers the finite-horizon optimal expected discounted cost:

$$\begin{aligned} V_{n+1,\beta}(\mathbf{x}) =& C(\mathbf{x}) + \beta\Big[\mathbb{I}_{\{x_r<K\}}\lambda_r^{'}\big(V_{n,\beta}(\mathbf{x}+e_1) - V_{n,\beta}(\mathbf{x})\big) \\ & + \mathbb{I}_{\{x_b>0,x_c<K\}}x_b\lambda_c^{'}\big(V_{n,\beta}(\mathbf{x}-e_2+e_3) - V_{n,\beta}(\mathbf{x})\big) + \mathbb{I}_{\{x_b>0\}}x_b\mu_b^{'}\big(V_{n,\beta}(\mathbf{x}-e_2) - V_{n,\beta}(\mathbf{x})\big) \\ & + \mathbb{I}_{\{x_c>0\}}x_c\mu_{db}^{'}\big(V_{n,\beta}(\mathbf{x}-e_3) - V_{n,\beta}(\mathbf{x})\big) + \mathbb{I}_{\{x_r>0,x_c>0,x_b<K\}}\min\big\{\mu_r^{'}\big((1-p)V_{n,\beta}(\mathbf{x}-e_1) \\ & + pV_{n,\beta}(\mathbf{x}-e_1+e_2) - V_{n,\beta}(\mathbf{x})\big), \mu_c^{'}\big(V_{n,\beta}(\mathbf{x}-e_3) - V_{n,\beta}(\mathbf{x})\big)\big\} \\ & + \mathbb{I}_{\{x_r>0,x_c=0,x_b<K\}}\mu_r^{'}\big((1-p)V_{n,\beta}(\mathbf{x}-e_1) + pV_{n,\beta}(\mathbf{x}-e_1+e_2) - V_{n,\beta}(\mathbf{x})\big) \\ & + \mathbb{I}_{\{x_r=0,x_c>0\}}\mu_c^{'}\big(V_{n,\beta}(\mathbf{x}-e_3) - V_{n,\beta}(\mathbf{x})\big) + V_{n,\beta}(\mathbf{x})\Big], \end{aligned} \tag{A1.14}$$

where $V_{n,\beta}(\mathbf{x})$ represents the optimal cost of the n-period β-discounted problem starting in state $\mathbf{x} = (x_r, x_b, x_c)$, and $\mathbb{I}_{\{.\}}$ is the indicator function. Here, e_i represents the ith unit vector. In other words, e_i is a vector that contains all zeros except for a 1 in the ith position. The initial condition, $V_{0,\beta} \equiv 0$ is assumed for mathematical convenience and has no effect on the infinite-horizon problem's results. In Equation A1.14's last line, the negative term $V_{n,\beta}(\mathbf{x})$ as well as the last term are associated with no state change. These are added for uniformity.

Exploring Treatment Census' Effect on Patient Flow Control.

In this model, we focus on the impact of the treatment census alone on the decision-making process. Therefore, we exclude the boarding section and assume that some patients need to return to the treatment section after completing their treatment. The system must make a decision regarding whether to accept these patients' requests. Patients whose requests are accepted join the treatment queue, while those whose requests are denied incur a cost of c due to the potential impact on the boarding patient's health outcome. Although this MDP model simplifies the problem, it offers valuable insights into the role of the treatment census and facilitates the development of practical policies for complex systems.

In this model, the state variable $x \in \bullet^+$ represents only the patients in the treatment queue, as we do not explicitly include the boarding section. Consequently, $C(x)$ is adjusted accordingly. Therefore, the value function in Equation A1.15represents the uniform version of this model.

$$\begin{aligned} V_{n+1,\beta}(x) =& C(x) + \beta\big[\lambda_r^{'} V_{n,\beta}(x+1) + \mu_r^{'}[\mathbb{I}_{\{x>0\}}((1-p)V_{n,\beta}(x-1) + p\min\{V_{n,\beta}(x-1)+c, V_{n,\beta}(x)\}) \\ & + \mathbb{I}_{\{x=0\}} V_{n,\beta}(x)]\big]. \end{aligned} \tag{A1.15}$$

We can reformulate this model using *event-based dynamic programming* operators (see Koole (1998)) and easily show that the value function in Equation A1.15 is convex by the event operators' closure under convexity.

theoremhe value function as defined in Equation A1.15 is convex.

W e can define the following operators as the arrival, discharge, and checkup decision operators.

$$T_A f(x) = f(x+1). \tag{A1.16}$$

$$T_D f(x) = \begin{cases} f(x-1) & \text{if } x > 0 \\ f(x) & \text{if } x = 0 \end{cases} \tag{A1.17}$$

$$T_C f(x) = \begin{cases} min\{f(x-1)+c, f(x)\} & \text{if } x > 0 \\ f(x) & \text{if } x = 0 \end{cases} \tag{A1.18}$$

By defining the cost operator as $T_{CO} f(x) = C(x) + f(x)$ and uniformization operator as $T_U[f_1(x), f_2(x), f_3(x)] = \beta\lambda_r^{'} f_1(x) + (1-p)\beta\mu_r^{'} f_2(x) + p\beta\mu_r^{'} f_3(x)$, we can rewrite Equation A1.15 using these operators:

$$V_{n+1,\beta}(x) = T_{CO} T_U [T_A V_{n,\beta}(x), T_D V_{n,\beta}(x), T_C V_{n,\beta}(x)] \tag{A1.19}$$

Thus, we can see that the value function is convex by demonstrating the closure if the these operators are under convexity, because $V_{n,\beta}$ is only a combination of these operators with the initial convex value function V_0. $V_{0,\beta} \equiv 0$ is trivially convex. The closure under other operators' convexity is shown in Koole (1998).

proposition 1. *The optimal policy for the value function as defined in Equation A1.15 is a threshold policy.*

Proof of Proposition 1. The optimal policy's threshold structure follows directly from the value function's convexity.

These results illustrate that the optimal policy is a function of the treatment census. Consequently, the current practice of prioritizing boarding patients over emergency patients cannot be an optimal policy unless the cost of rejecting checkup requests is so high that $c > f(x) - f(x-1)$ for any x.

Numerical Results for MDP

We examine the structure of the optimal policy for the MDP presented in Equation A1.14 using numerical analysis. To thoroughly investigate the existence of a robust optimal policy, we create a test suite including a wide range of parameters (refer to Table A1.13). The first column in Table A1.13 provides details about a typical ED setting. The second column indicates the specific parameter ranges we test while keeping the values of other parameters based on Case 1. Lastly, the last column represents a randomized test suite. We solve 1,000 instances from the randomized test suite. It is important to note that in all cases, we assume that

$\lambda_r < \min \mu_r, \mu_c$. Additionally, to maintain realism, we assume that the average boarding time exceeds the average time before requesting a checkup (i.e., $\mu_b < \lambda_c$). For each parameter set provided in Table A1.13, we solve the MDP using the value iteration algorithm and record the optimal actions for each state. To visualize the complexity of the optimal structure, we plot the minimizing action against the number of patients in the regular queue (x_r, on the horizontal axis) and the number of patients in the checkup queue (x_c, on the vertical axis). However, it is important to note that the optimal action is also influenced by the boarding census. Hence, we generate these plots for four levels of the boarding census $(0, \lfloor \frac{K}{3} \rfloor, \lfloor \frac{2K}{3} \rfloor, K)$. While we conduct these experiments for over 1,000 instances, we only present results that demonstrate significant differences in the optimal structure to gain insights into how the optimal structure varies across different parameter levels.

Figure 2.4 in the main body of the paper displays switching curves that illustrate the relationship between the regular queue (x_r) and the checkup queue (x_c) for three levels of the boarding census (x_b). It is important to note that the optimal actions are dependent on the boarding census. When the boarding section is empty, the optimal policy closely resembles a threshold policy, where a patient from the regular queue is selected when $x_r \geq L$, and a patient from the checkup queue is chosen when $x_r < L$. Here, L^* represents the threshold value. As the census in the boarding section increases, the optimal policy increasingly favors the checkup queue. This can be attributed to the fact that an increase in the boarding census results in a higher arrival rate of boarding patients, while the arrival rate of regular patients remains constant and unaffected by the boarding census.

Furthermore, our analysis reveals that the optimal policy continues to prioritize the checkup queue as the census in the boarding section increases. We also observe that when the arrival rate of patients in the checkup queue (λ_c) is relatively low (e.g., $\lambda_c = 0.5$), the optimal policy gives preference to patients in the treatment queue, unless the boarding census is high. More detailed results can be found in Figures A1.15 and A1.16.

Another crucial parameter that has a significant impact on the optimal policy is the service rate, represented by μ_c and μ_r. In our base case, we assume that the service time for boarding patients is shorter than that for emergency patients, as is typically observed in practice (i.e., $\mu_r > \mu_c$). However, if the service time for checkups is assumed to be equal to the service time for emergency patients, the optimal policy consistently prioritizes emergency patients when the census of the checkup queue falls below a certain threshold (as shown in the first row of Table A1.14). Conversely, when the checkup time is shorter than expected, the optimal policy primarily favors boarding patients, particularly if the boarding census is high, in order to alleviate congestion in the overall system.

The admission probability (represented by p) exhibits similar effects to the boarding arrival rate. As the admission probability decreases, the number of patients in the boarding section diminishes, resulting in a decline in the arrival rate of boarding patients. Consequently, the optimal policy tends to accept regular patients more frequently in order to prevent congestion in the regular queue (refer to Table A1.18).

In Case 1, we assume that the waiting time cost per unit is the same for both regular and boarding patients. The waiting time cost plays a critical role in determining the optimal policy. As expected, when we increase the ratio of the waiting time cost for a "regular patient" to that of a "boarding patient," the optimal policy predominantly gives priority to regular patients (see Table A1.17).

While other parameters also influence the optimal policy, our analysis of over 1,000 instances reveals that most of the variations can be attributed to changes in the following parameters: arrival rates (λ_r and λ_c), admission probability (p), service rates (μ_r and μ_c), and waiting time cost ratios ($\frac{\gamma_r}{\gamma_c}$). We present additional results in the tables below.

Table A1.13: Parameters Used for the Test Suite

	Case 1	Case 2 (tested values)	Case 3 (randomized test suite)
λ_r	3	[1, 2, 3.5]	U[0.5, 3]
λ_c	2	[0.5, 1, 3, 4]	U[0.5, 6]
μ_r	4	[3.5, 6, 8, 10]	U[4, 10]
μ_c	6	[4, 8, 10]	U[4, 10]
μ_b	0.25	[0.1, 0.5, 1, 1.5]	U[0.1, 2] & $< \lambda_c$
μ_{db}	$(\mu_b^{-1} - \lambda_c^{-1})^{-1}$	-	-
p	0.1	[0.3, 0.5, 0.7]	U[0.1, 0.9]
K	30	[10, 20, 50]	30
$\frac{\gamma_r}{\gamma_c}$	1	[0.5, 0.75, 1.5, 2]	U[0.25, 4]
$\frac{h_r}{h_c}$	1	[0.5, 0.75, 1.5, 2]	U[0.25, 4]
$\frac{h_r}{\gamma_r}$	1	[1.5, 2, 3]	U[1, 4]
α	0.5	[0.1, 0.9]	0.5

Table A1.14: Optimal action over various μ_c values.

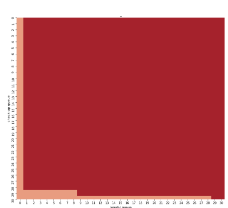
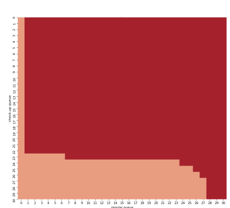
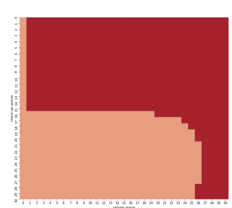
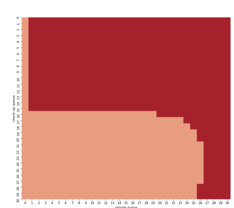
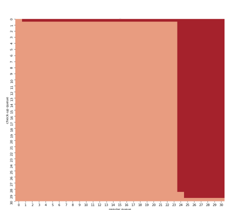
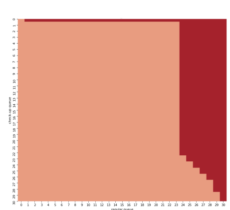
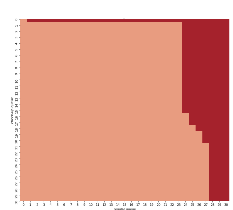
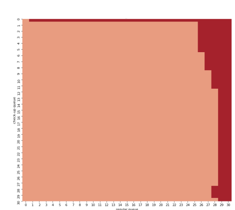
	$x_b = 0$	$x_b = 10$	$x_b = 20$	$x_b = 30$
$\mu_c = 4$				
$\mu_c = 8$				

Table A1.15: Optimal action over various λ_r values.

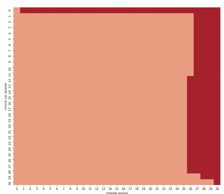
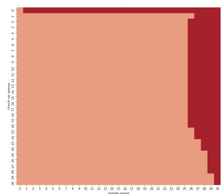
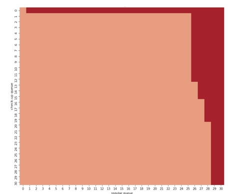
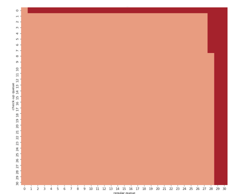
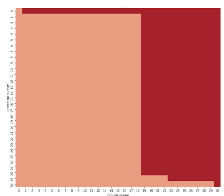
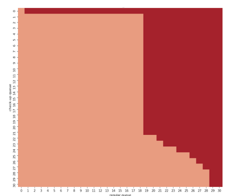
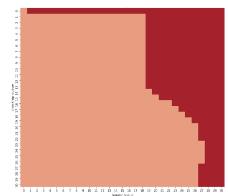
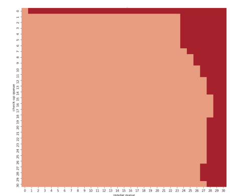
	$x_b = 0$	$x_b = 10$	$x_b = 20$	$x_b = 30$
$\lambda_r = 1$				
$\lambda_r = 3.5$				

Table A1.16: Optimal action over various λ_c values.

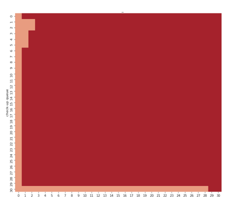
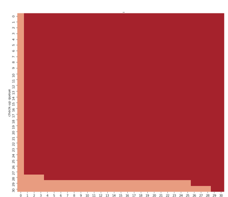
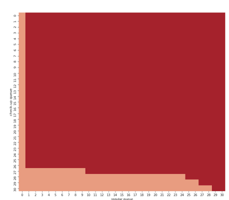
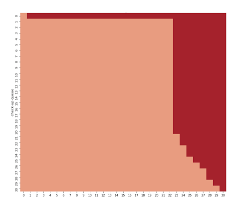
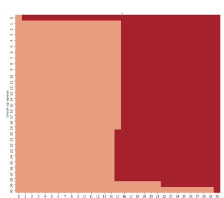
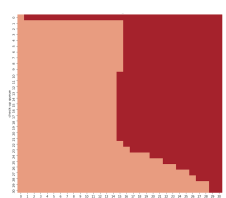
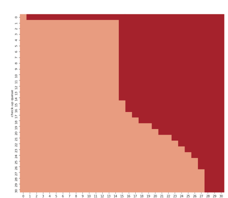
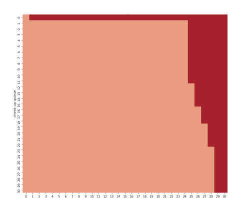
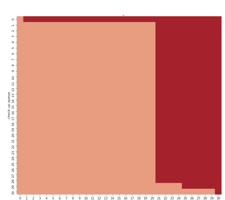
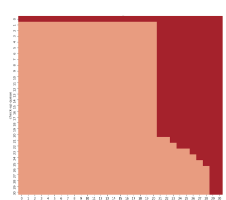
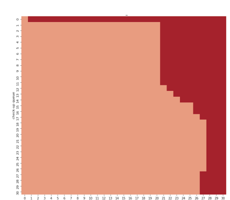
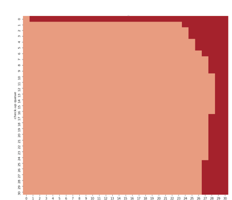
	$x_b = 0$	$x_b = 10$	$x_b = 20$	$x_b = 30$
$\lambda_c = 0.5$				
$\lambda_c = 1$				
$\lambda_c = 3$				

Table A1.17: Optimal action over various $\frac{\gamma_r}{\gamma_c}$ values.

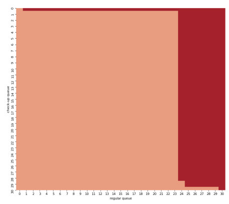
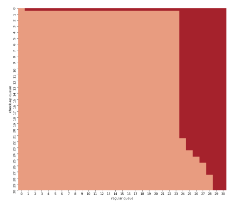
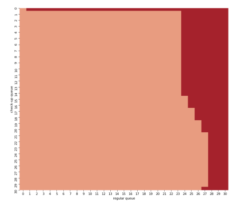
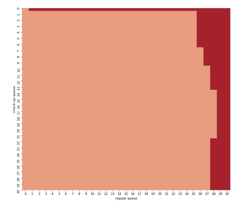
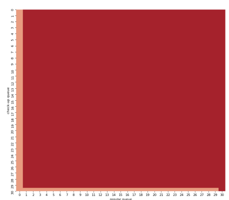
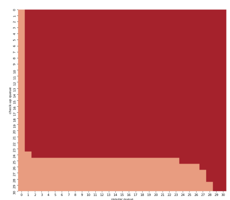
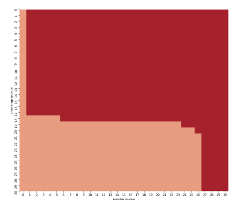
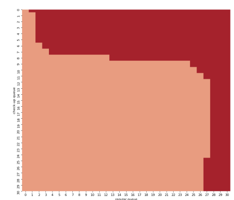
	$x_b = 0$	$x_b = 10$	$x_b = 20$	$x_b = 30$
$\frac{\gamma_r}{\gamma_c} = 0.5$				
$\frac{\gamma_r}{\gamma_c} = 2$				

Table A1.18: Optimal action over various p values.

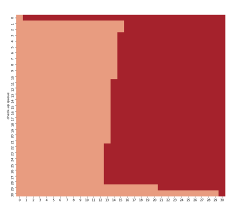
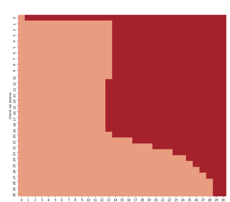
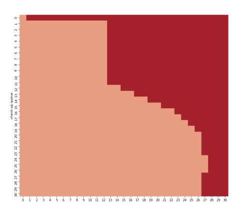
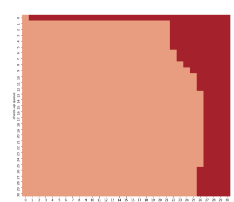
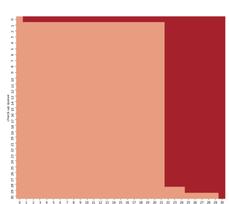
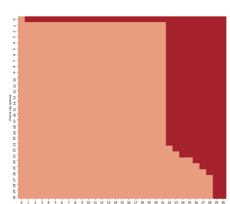
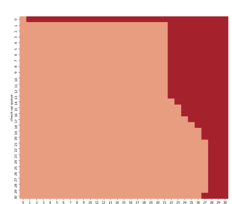
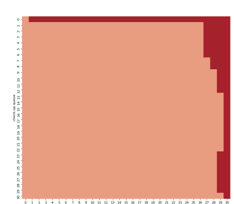
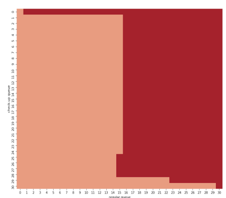
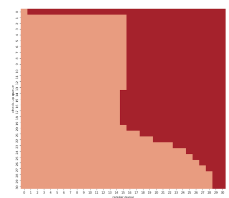
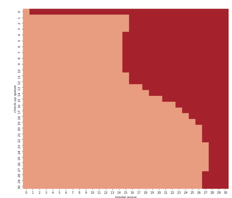
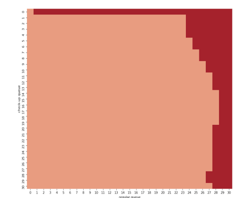
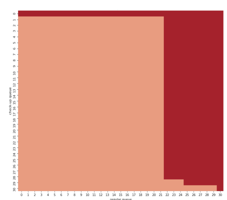
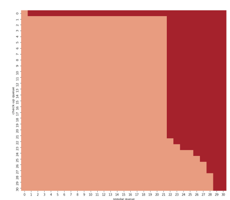
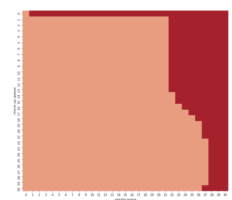
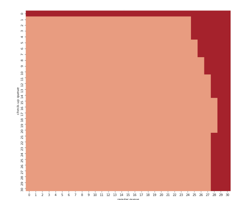
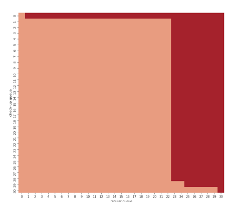
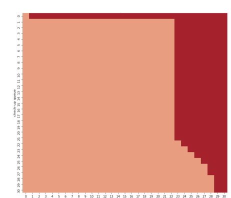
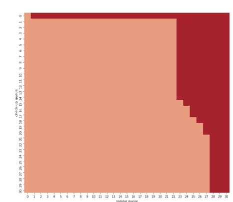
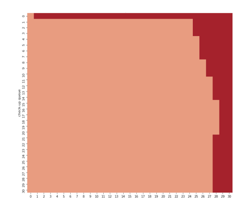
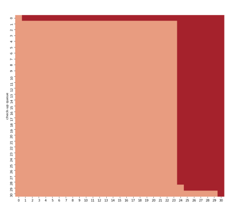
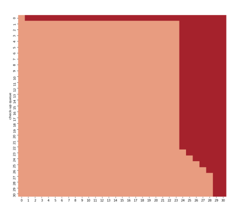
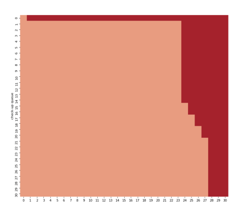
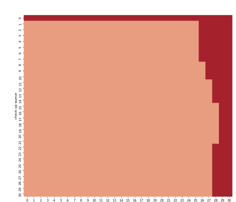
	$x_b = 0$	$x_b = 10$	$x_b = 20$	$x_b = 30$
$p = 0.1$				
$p = 0.5$				

Table A1.19: Optimal action over various α values.

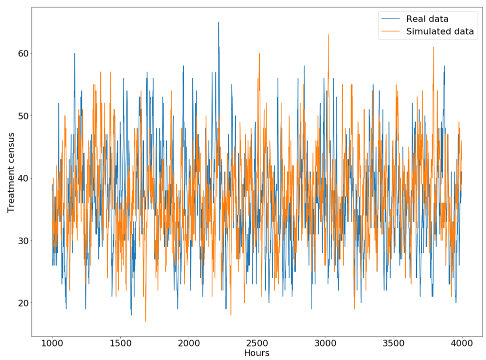
	$x_b = 0$	$x_b = 10$	$x_b = 20$	$x_b = 30$
$\alpha = 0.6$				
$\alpha = 0.7$				
$\alpha = 0.8$				
$\alpha = 0.9$				

A1.7 Simulation Model

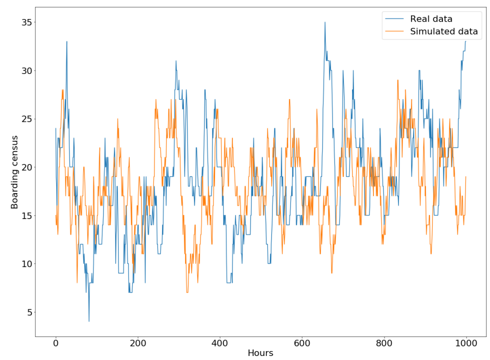
Table A1.20: Comparing the simulation model and real data.

	Treatment time		Treatment census		Boarding time		Boarding census		checkup request probability	Admission Probability
	Mean	SD	Mean	SD	Mean	SD	Mean	SD		
Real data	377.96	283.75	39.88	8.14	976.96	741.41	28.01	6.77	0.58	0.23
Simulation	376.26	281.75	39.65	7.81	987.79	735.07	27.54	5.82	0.60	0.22

Notes. We performed 1,000 iterations.



(a) Hourly treatment census (in a randomly selected 3,000-hours interval)



(b) Hourly boarding census (in a randomly selected 1,000-hours interval)

Figure A1.4: Comparing the simulation model output and real data.

Simulation Results for a Secondary ED

Table A1.21: Comparing a secondary hospital's simulation model and real data.

	Treatment time		Treatment census		Boarding time		Boarding census		checkup request probability	Admission Probability
	Mean	SD	Mean	SD	Mean	SD	Mean	SD		
Real data	329.87	267.35	34.17	7.26	613.13	580.65	6.18	2.80	0.41	0.13
Simulation	330.17	250.16	34.21	7.15	615.53	577.95	6.37	2.59	0.40	0.13

Notes. We performed 1,000 iterations.

Table A1.22: Simulation results of a secondary ED's boarding patient flow control policies.

Policy	Average treatment time	Hourly average treatment section census	Total treatment time (patient years)
Status quo	330.17 (0.35)	34 (0.12)	25.13 (0.02)
Threshold Policy	317.82 (0.35)	33 (0.13)	24.19 (0.02)
Lower Bound	302.74 (0.37)	31 (0.11)	23.04 (0.02)

Notes. We performed 1,000 iterations.

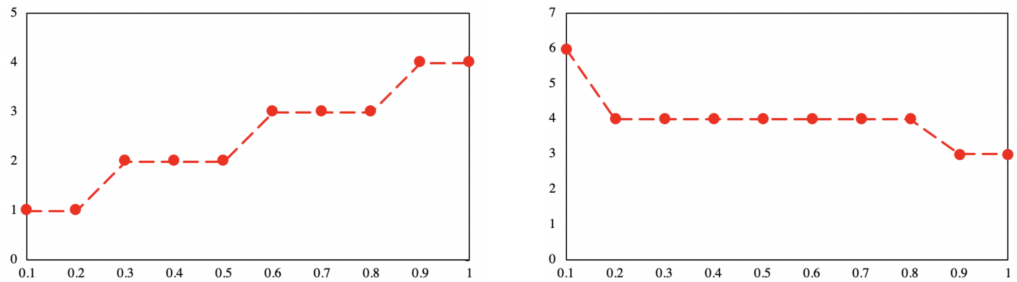


Figure A1.5: A Secondary ED's Scheduling Framework for an ED-hospitalist.

Table A1.23: Simulation Results of hospitalist Visit Scheduling Policies (A Secondary ED)

	Average treatment time		Hourly average treatment section census		Daily average number of hospitalist visit per category*	
Patient flow control policy	Status quo	TP***	Status quo	TP	Status quo	TP
Without HS**	330.17 (0.36)	317.82 (0.33)	34 (0.12)	33 (0.12)	-	-
Time-based HS	320.14 (0.34)	308.09 (0.30)	31 (0.11)	30 (0.11)	2.67 (0.01)	2.95 (0.01)
Census-based HS	317.83 (0.34)	306.81 (0.31)	31 (0.11)	30 (0.10)	2.38 (0.01)	2.58 (0.01)
Mixed HS	319.04 (0.34)	307.11 (0.31)	31 (0.11)	30 (0.11)	2.46 (0.01)	2.72 (0.01)

Notes. * We consider four categories. 1000 iterations.

3

Offline-Online Retail Collaboration via Pickup Partnership

In the previous chapter, we demonstrated the profound role of data science in tackling complex healthcare challenges, particularly emergency department overcrowding. Now, shifting focus to the retail domain, the subsequent chapters will provide examples of how data science can revolutionize operations and address potential issues within the retail sector. Specifically, we will explore how leveraging data-driven insights can enhance the efficiency of retail operations, optimizing processes and maximizing outcomes.

In an upcoming chapter, we will introduce a stylized model aimed at improving partnership dynamics among retailers in the context of pickup partnerships, fostering collaboration and synergy to create mutually beneficial outcomes for all involved parties—a true win-win situation.

3.1 Introduction

Customer demand for increasingly convenient shopping (Boston Retail Partners 2021) has exacerbated the competition between pure online retailers and multichannel retailers. To stay ahead in this competition, taking advantage of their physical presence to enhance the customer experience (Chen et al. 2021), many multichannel retailers have started offering in-store pickup services that allow customers to pick up online orders in physical stores(Gao and Su 2017). Due to their convenience, in-store pickup services have rapidly become popular among customers. In 2020, in-store pickup sales in the United States doubled the previous year's total, and this trend is expected to continue to grow at an annual rate of at least 15% until 2024 (Chevalier 2021). To remain competitive and respond to the growing customer demand for in-store pickup services, pure online retailers have recently started forming strategic partnerships with brick-and-mortar retailers (referred to as offline partners in the rest of this paper), which is termed *pickup partnership*.

A pickup partnership enables an online retailer to use an offline partner's stores as pickup locations for customers who prefer to pick up their online orders from a nearby store at no additional shipping cost. Under this fulfillment option, the online retailer ships the customer order to the offline partner store chosen by the customer and notifies the customer when the order is ready for pickup. When the customer arrives for pickup, the offline partner handles the process using its own staff and informs the online retailer at the end of the process. Amazon Hub Counter (AHC) is a widely known example of such pickup partnerships. Through AHC, physical retailers collaborate with Amazon to make their stores assisted-pickup locations for Amazon orders. Examples include Rite Aid, GNC, Health Mart, and Stage Stores in the United States, NEXT in the U.K., Librerie Giunti al Punto, Fermopoint, and SisalPay in Italy, and ParcelPoint in Australia. Similarly, Uniqlo, a Japanese fashion retailer, has partnered with 5,700 7-Eleven convenience stores in Tokyo to offer in-store pickup services.[1]

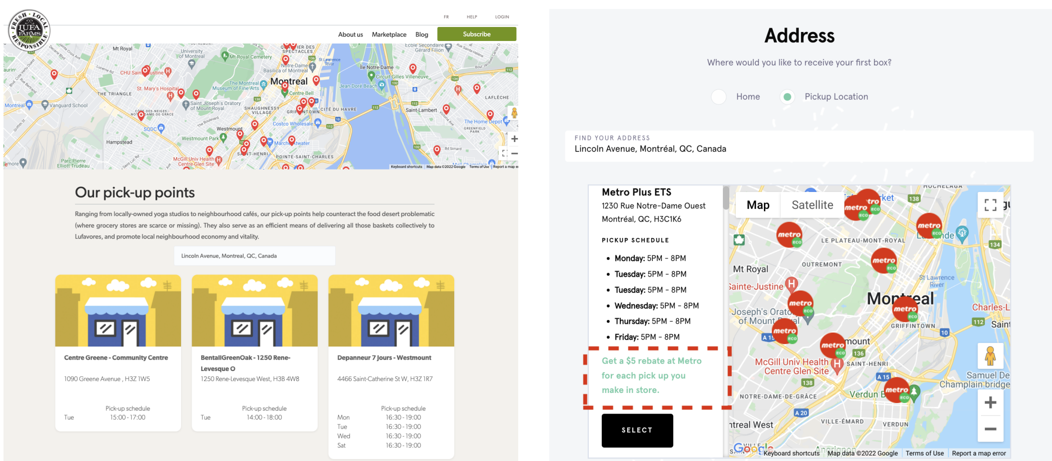
Pickup partnerships promise several benefits to both online retailers and offline partners.For online retailers, offering a convenient in-store pickup option should translate into increased sales. In addition, compared to direct-delivery shipments, in-store pickup fulfillment is likely to reduce online retailers' freight costs due to the possibility of pooling several orders and hence decreasing (expensive) last-mile delivery costs (Morganti et al. 2014). For offline partners, in-store pickups induce additional foot traffic that can generate increased revenue through cross-selling. It has been reported that 45% of customers who used in-store pickup services made an in-store purchase during the pickup visit (UPS 2015). In addition, online retailers may compensate offline partners for handling each pickup order, creating an additional revenue stream for offline partners. These benefits suggest that the trend of forming pickup partnerships is likely to persist or even expand in the future. This trend raises the question of the best way to establish pickup partnerships between online retailers and offline partners. In practice, we observe two different policies (see Figure 3.1.1 for illustrative examples). The first is a *fixed fee policy* under which the online retailer pays a fixed commission to the offline partner for each in-store pickup order (Morganti et al. 2014, Fang et al. 2019).[2] For example, this type of pickup partnership has been established between Maturin and LOCO, between Lufa Farms (a Canadian online grocery retailer) and local stores as illustrated in Figure 3.1.1(a), and between Amazon and thousands of AHC retailers.[3] The second is a *coupon policy* under which the online retailer issues a coupon to the customers who select the in-store pickup delivery option. Customers can redeem this coupon

[1] `https://ww.fashionnetwork.com/news/uniqlo-japan-launches-in-store-pickup-service-with-7-eleven,625015.html`

[2] `https://join.healthmart.com/pharmacy-marketing-and-promotions/becoming-an-amazon-hub-counter/`

[3] `https://www.pudoinc.com/member-benefits/pudopoint-counter/`

to make a discounted purchase at the offline partner's store. If the coupon is redeemed, the online retailer will then reimburse the offline partner for the discounted amount. The pickup partnership between Cookit (a meal-kit retailer) and Metro (a Canadian grocery retailer) is based on the coupon policy, as illustrated in Figure 3.1.1(b). It is worth noting that even though the compensation payment for each in-store pickup order is similar between the two policies, the recipient of the payment depends on the policy. Under the fixed fee policy, the online retailer pays the offline partner, whereas under the coupon policy, the payment is made to the customer, making the two policies structurally different.



(a) Lufa Farm and Local Stores

(b) Cookit and Metro

Figure 3.1.1: Illustration of the Fixed Fee Policy (Left) and the Coupon Policy (Right)

When establishing a pickup partnership, should online retailers pay a fixed fee directly to the offline partner or incentivize customers by offering a coupon? We are not aware of any academic study on pickup partnerships. From a practitioner perspective, the fixed fee policy may be more desirable as it generates a readily visible revenue stream for offline partners through the fixed commission paid for each in-store pickup order. At the same time, the coupon policy may also be desirable since incentivizing customers will likely drive more traffic to offline partners' stores, resulting in more cross-selling opportunities. Overall, it is not clear which of the two policies is more beneficial. This leads to our first two research questions: *How should an online retailer choose between fixed fee and coupon policies when establishing a pickup partnership? What type of online retailer is more suitable for the fixed fee policy versus the coupon policy?*

The presence of fixed fee and coupon policies in practice does not necessarily imply that they always help retailers establish an efficient pickup partnership. After all, a pickup partnership is tantamount to a contract between two parties. The supply chain contract literature (see Tsay et al. 1999) has established that inefficient contracts are common in practice (Loch and Wu 2008) and arise due to misaligned incentives between the two parties (Pavlov et al. 2022). From this perspective, just like inefficient contracts, pickup partnerships may also entail inefficiencies when the incentives between the online retailer and the offline partner are misaligned. This leads to our next two research questions: *Do fixed fee and coupon policies result in an inefficient pickup partnership? If so, can we propose an alternative policy that mitigates such inefficiencies?*

To answer our research questions, we develop a stylized model that captures the key features of a pickup partnership. Specifically, we consider an online retailer who sells a product only via an online channel. The online retailer contemplates the opportunity to offer an in-store pickup service to her customers through a partnership with an offline partner. The online retailer's objective is to maximize her profit while ensuring that the proposed partnership is also beneficial to the offline partner. When the pickup partnership is established, customers strategically decide between the direct-delivery and in-store pickup options to maximize their utility. We first analytically examine the effect of the partnership on the demand and profits of both partners under each policy. We then identify conditions under which each policy is beneficial; that is, both partners earn higher profits (in a non-strict sense) under the pickup partnership than they would without that. Armed with these results, we compare the two policies and characterize when each policy is optimal for the two parties. We then infer what type of online retailers are more suitable for the fixed fee policy versus the coupon policy when establishing a pickup partnership. Lastly, we examine the conditions under which the two policies generate inefficiencies and prescribe an alternative policy, termed the *hybrid policy*, to mitigate those inefficiencies.

Our study makes several contributions. First, we find that establishing a pickup partnership (regardless of policy) affects the demand for the online retailer's product in two ways: (i) the pickup partnership enables the online retailer to expand its market coverage due to the increased convenience of the direct-delivery option; (ii) the in-store pickup option incentivizes some existing customers to switch their delivery mode to the in-store pickup option. While the former effect (i.e., market expansion) on demand increases the online retailer's profit, the latter effect (i.e., demand shift) can actually hurt the online retailer's profit if the profit margin from in-store pickup orders is lower than the profit margin from direct-delivery orders. Hence, the pickup partnership will be beneficial as long as these two demand streams result in a net profit gain. Second, our results reveal that the two demand streams induced by the partnership are more substantial under the coupon policy. Indeed, beyond offering the convenience of the in-store pickup option (as in the fixed fee policy), the coupon policy also incentivizes additional customers to use the in-store pickup option to take advantage of the coupon. These additional customers do not necessarily increase the online retailer's profit, especially when the in-store pickup fulfillment is less profitable, on a per item basis, than the direct-delivery fulfillment. Thus, choosing between the fixed fee and coupon policies when establishing a pickup partnership requires a careful assessment of both partners' cost structures. Third, when examining the pickup partnership with respect to the offline partner's cost structure, we find that the online retailer should use the coupon policy if the offline partner can manage in-store pickups at a low handling cost. A low handling cost allows the online retailer to pay a low compensation per in-store pickup order, hence increasing the profit margin of in-store pickup orders. Consequently, the additional customers attracted by the coupon policy (relative to the fixed fee policy) will increase the online retailer's profit. When the offline retailer's handling cost is moderate, we show that the fixed fee policy is the best option for the partnership, but when the offline partner's handling cost is high, a partnership becomes plausible only with high compensation, so that the online retailer is better off not establishing a pickup partnership. When examining the pickup partnership with respect to the online retailer's cost structure, we find that a pickup partnership is not profitable for online retailers with low direct-delivery fulfillment costs or high in-store pickup fulfillment costs. Otherwise, the fixed fee policy is more suitable for online retailers with moderate direct-delivery fulfillment costs, moderate in-store pickup fulfillment costs, or low-priced products, whereas the coupon policy is more beneficial for online retailers with high direct-delivery fulfillment costs, low in-store pickup fulfillment costs, or high-priced products.

Finally, while our model suggests that the two policies used in practice can ensure a profitable partnership, it remains unclear whether these policies allow online retailers to fully unlock the potential benefits of such partnerships. In fact, we find that both policies can be inefficient in the sense that selecting the optimal policy (either fixed fee or coupon) for the partnership comes with an opportunity cost for the online retailer. In such cases, the online retailer has to establish the partnership under a suboptimal compensation value. To mitigate such inefficiencies, we propose a new hybrid policy that leverages the features of both the fixed fee and the coupon policies. In particular, the hybrid policy allows the online retailer to split the compensation such that a portion is paid to the offline partner as a fixed fee commission, with the remainder offered as a coupon to customers. We find numerically that inefficient pickup partnerships occur frequently and that the profit improvement generated from the hybrid policy can be substantial.

The rest of the paper is organized as follows. In Section 2, we review the related literature, and in Section 3, we formalize our model. We then analyze the model and derive several analytical results in Section 4. We consider various extensions in Section 5. In Section 6, we identify inefficiencies induced by the current policies and propose a new policy. Finally, we conclude and outline the managerial implications of our results in Section 7.

3.2 Literature Review

This paper is related to three streams of literature: in-store pickup services, retail partnerships, and coupon promotions.

In-Store Pickup Services.

The recent growth of in-store pickup services (e.g., click-and-collect and ship-to-store) has led to an increase in research on that topic. The literature has examined two types of in-store pickup services: buy-online-pickup-in-store (BOPIS) and ship-to-store (STS). The major difference between the two is the order fulfillment point. BOPIS orders are fulfilled using store inventory and can thus only be placed for products available in a store (Gao and Su 2017), whereas STS orders are fulfilled using the distribution center inventory and can be placed for any product available online, regardless of whether it is stocked in any store (Ertekin et al. 2021).

Gallino and Moreno (2014) empirically show that even though using BOPIS can reduce online sales, the sales generated from the additional store traffic can make retailers better off when offering such services. Focusing on individual products, Gallino et al. (2017) find that STS services may shift sales from high-selling products to low-selling products. In the same vein, Ertekin et al. (2021) find that STS has a heterogenous effect on sales of online-only products versus products available both online and offline. The authors conclude that considering the STS effect when choosing channel(s) to sell a product can improve the performance of in-store pickup services. Akturk and Ketzenberg (2021) evaluate the competitive impact of BOPIS. They show that both online and store sales at a focal retailer are adversely affected after the competitors' launch of a BOPIS service. Focusing on customer behavior, Song et al. (2020) find that BOPIS has a positive effect on offline purchase frequency and on online purchase amounts. Glaeser et al. (2019) demonstrate that the location of the pickup stores can have a significant effect on BOPIS profitability.

Among analytical studies, Gao and Su (2017) examine the impact of BOPIS on store operations. The authors find that despite enabling retailers to increase demand from new customers, BOPIS may not be suitable for products that sell well in stores. Hu et al. (2022) demonstrate that retailers can leverage the additional demand induced by BOPIS to improve their store fill rates. Ertekin et al. (2021) illustrate that when implementing STS, retailers should offer easy-to-substitute products only online and difficult-to-substitute products both online and in stores. Similarly, Cao et al. (2016) find that in-store pickups may not be suitable for all products. Finally, Gao et al. (2022) show that it might be optimal for retailers to reduce their physical store presence under BOPIS.

The studies in this literature primarily examine in-store pickup services when the retailer owns both the online and offline channels. In contrast, our paper investigates in-store pickup services when offered by a pure online retailer that partners with an offline store. Thus, some of the highlighted benefits of BOPIS or STS for multichannel retailers (e.g., cross-selling, additional store traffic) will not be present for online retailers under the pickup partnership. More importantly, unlike a pickup partnership, which can be implemented using different policies, traditional in-store pickup services (whether BOPIS or STS) are quite standard across retailers. Therefore, existing studies on traditional in-store pickup services cannot help identify which policy retailers should use when establishing a pickup partnership. Overall, we contribute to this literature by developing a theoretical model (i) to demonstrate when and how a pickup partnership should be established, (ii) to identify how online retailers should choose between fixed fee and coupon policies, and (iii) to propose an alternative policy to improve the pickup partnership efficiency.

Retail Partnerships.

These kinds of partnerships have been studied in several settings, ranging from supply chain contracts and coordination (see Cachon and Lariviere 2005, for a comprehensive review) to coalition and coopetition contracts (Nagarajan and Sošić 2007, Cohen and Zhang 2022, Yuan et al. 2021). Our work is closely related to a growing stream in this literature that studies offline-online partnerships to enhance omnichannel retailing offerings, such as buy online, return in-store (Hwang et al. 2021, Nageswaran et al. 2021) and search offline, buy online (i.e., showrooming) (Dan et al. 2021). In this stream, Nageswaran et al. (2021) theoretically examine the potential of a return partnership between a pure online retailer and an offline partner that serves as the in-store return location for the online retailer. The authors find that such a return partnership can be formed either when there is only a small product assortment overlap between the two parties, or when the offline partner has a small number of physical locations. Hwang et al. (2021) empirically show that such return partnerships generate additional sales for the offline partner. Most studies on offline showrooming focus on a single company (e.g., Bell et al. 2018, Gao et al. 2022). Dan et al. (2021) analytically study how an online retailer should choose between competing and non-competing offline retailers to offer a physical showrooming service and how the type of offline retailer (competing or non-competing) can affect an online retailer's pricing strategy under an exogenous commission fee. In these studies, the focus is primarily on the types of offline partners that should be selected as partners. By contrast, our study focuses on how online retailers should select partnership policies according to offline partners' characteristics. Even when we extend our review to the broader retail partnership literature, we could not find any study with guidance on how online retailers and offline partners should establish a pickup partnership. We contribute to this stream by studying the impact of the two pickup partnership policies on the decisions and payoffs of the key stakeholders.

Coupon Promotions.

There is a large literature stream on coupon redemption (e.g., Reibstein and Traver 1982, Danaher et al. 2015), coupon effects on customer behavior (e.g., Narasimhan 1984, Neslin et al. 1985, Heilman et al. 2002, Su et al. 2014), and marketing effects and optimal coupon scheduling (e.g., Sethuraman and Mittelstaedt 1992, Reimers and Xie 2019, Baardman et al. 2019). We position our work with respect to papers that consider the role of coupon promotions in channel coordination. Among those, Martin-Herran and Sigué (2015) find that manufacturers prefer coupon promotions over a cooperative pricing strategy. Li et al. (2020) evaluate how issuing coupons by either manufacturers or retailers can affect the supply chain profit. Pauwels et al. (2011) show that offering online promotions can also increase the demand for the offline channel, creating a channel synergy effect. Despite all these valuable contributions, there is no study that leverages coupons to facilitate a pickup partnership between an online retailer and an offline partner. Our study contributes to this literature by demonstrating how a coupon promotion can be used to design an effective mechanism for pickup partnerships.

3.3 Model Description

In this section, we develop a stylized model to characterize the key features of a pickup partnership between an online retailer and an offline partner. In the following subsections, we describe our modeling framework, introduce a baseline policy in which the pickup partnership does not exist, and consider two different pickup partnership scenarios by building on the baseline policy.

3.3.1 Modeling Framework

The model consists of an online retailer that sells a product through its online channel at price p. Consistent with the literature that models interactions between retailers and customers using the circular location model (e.g., Balasubramanian 1998, Shulman et al. 2009, Gao et al. 2022), we assume that the online retailer serves customers who are uniformly distributed on the circumference of a circular city with a circumference of one (Salop 1979). The online retailer's warehouse is located at the center of the circular city and is thus equidistant from all customers. Without loss of generality, we assume that the size of the market is normalized to one.

To provide customers with an in-store pickup service for online orders, the online retailer considers establishing a partnership with an offline partner. When such a partnership does not exist, the online retailer can only offer a direct-delivery option to its customers under which orders are shipped directly to customers. If a pickup partnership is established, in addition to the direct-delivery option, customers are now able to select a free in-store pickup option. With this option, the online retailer ships orders to the offline partner, and customers pick up their orders by visiting the offline partner at their convenience. We assume that the offline partner is randomly located on the circumference of the circle. We also assume that the product sold by the online retailer is not offered by the offline partner.

Customers: We assume that customers make purchasing decisions based on their utility. The valuation of product is v for all customers. If customers opt for the direct-delivery option, they incur a "hassle" cost h_o that includes both the shipping cost and the inconvenience of waiting for the delivery. We assume that customers are heterogeneous with respect to h_o such that $h_o \sim U[0, 1]$. If customers opt for the in-store pickup option, they incur a hassle cost of $h_p x$ for visiting the store to pick up their order, where h_p is the hassle cost per unit distance and x is the distance between a customer's location and the offline partner's location. Since customers are uniformly

distributed on the circumference of the circular city, we have $x \sim U[0, 1/2]$.
Online Retailer: When the online retailer fulfills an order via direct delivery, she incurs a direct delivery fulfillment cost of $c_o < p$ due to the logistics required to ship the order from her warehouse to the customer's doorstep. When the online retailer fulfills an order via the in-store pickup option, she incurs an in-store pickup fulfillment cost of c_p due to the logistics required to ship the order from her warehouse to the offline partner. Following the literature (Morganti et al. 2014), we assume that, compared to the direct-delivery fulfillment option, the online retailer can save on logistics costs with the in-store pickup fulfillment option by pooling multiple orders into a single delivery; that is, $c_p \leq c_o$. Without loss of generality, we assume a zero procurement cost.
Offline Partner: When the offline partner acts as an in-store pickup point, she incurs a handling cost c_s for each pickup order because she must temporarily store the order and assign staff to process in-store pickups. Customers who visit the offline partner to pick up their orders can generate cross-selling opportunities for the offline partner. To capture this effect, following the cross-selling literature (Gao and Su 2017, Ertekin et al. 2021), we assume that the offline partner earns a profit r from customers who visit the offline partner to pick up their order.

3.3.2 Baseline Policy

Under the baseline policy, the online retailer does not form a partnership with the offline partner and consequently offers only the direct-delivery option to her customers. Under this benchmark scenario, the utility of purchasing with the direct-delivery option amounts to $u_o^B = v - p - h_o$, where the superscript B denotes the *baseline policy*. We let d_o^B denote the online retailer's endogenous demand from customers who opt for the direct-delivery option under the baseline policy. Then, the online retailer's profit under the baseline policy is given by (details of the demand derivation appear in Appendix A2.2)

$$\pi_o^B = (p - c_o)d_o^B. \tag{3.3.1}$$

Since there is no partnership, the offline partner does not earn any profit from the online retailer's customers (i.e., $\pi_s^B = 0$) under the baseline policy.

3.3.3 Fixed Fee Policy

Under the fixed fee policy, the online retailer and the offline partner establish a pickup partnership under which the online retailer pays the offline partner a fixed fee α for each in-store pickup order. Subsequently, the online retailer offers both direct-delivery and in-store pickup options to her customers. Customers who opt for the direct-delivery option earn a utility $u_o^F = v - p - h_o$, where the superscript F denotes the *fixed fee policy*. Customers who opt for the in-store pickup option earn a utility $u_p^F = v - p - xh_p$. We let d_o^F and d_p^F denote the online retailer's demand for the direct-delivery and in-store pickup options, respectively, under the fixed fee policy. Then, the online retailer's profit under this policy is

$$\pi_o^F(\alpha) = (p - c_o)d_o^F + (p - c_p - \alpha)d_s^F. \tag{3.3.2}$$

In turn, the offline partner's profit from the partnership under the fixed fee policy is equal to

$$\pi_s^F(\alpha) = (r + \alpha - c_s)d_s^F. \tag{3.3.3}$$

3.3.4 Coupon Policy

The coupon policy establishes a pickup partnership between the online retailer and offline partner under which the online retailer provides a coupon with monetary value β to customers who opt for the in-store pickup option. Customers can then redeem the coupon to receive a discount on any purchase made at the offline partner. The online retailer reimburses the offline partner the amount β for any redeemed coupon.

We assume that customers who opt for the in-store pickup option will make a purchase from the offline partner to redeem the coupon with probability θ.[4] Consistent with the literature showing that customers increase purchase amounts when they have a coupon (Gupta 1988, Krishna and Shoemaker 1992), we assume that when the coupon is redeemed, the offline partner's profit from the cross-selling opportunity increases by β (i.e., the profit due to cross-selling becomes $r+\beta$ with probability θ). Under this policy, customers who opt for the direct-delivery option earn a utility $u_o^C = v - p - h_o$, where the superscript C denotes the *coupon policy*. Customers who opt for the in-store pickup option earn a utility $u_p^C = v + \theta\beta - p - xh_p$. We let d_o^C and d_p^C denote the online retailer's demand for the direct-delivery and in-store pickup options, respectively, under the coupon policy. Consequently, the online retailer's profit is equal to

$$\pi_o^C(\beta) = (p - c_o)d_o^C + (p - c_p - \theta\beta)d_s^C, \tag{3.3.4}$$

whereas the offline partner's profit amounts to

$$\pi_s^C(\beta) = (r + \theta\beta - c_s)d_s^C. \tag{3.3.5}$$

Under this modeling framework, the timeline of events shown in Figure 3.3.1 unfolds as follows:

1. The online retailer decides whether and with which policy to form a pickup partnership with the offline partner. Under the fixed fee policy, the online retailer determines the optimal parameter α to maximize her profit $\pi_o^F(\alpha)$, subject to the offline partner's rationality constraint; that is, $\pi_s^F(\alpha) \geq 0$. Under the coupon policy, the online retailer determines

[4] An alternative way to model the coupon redemption probability θ is to assume that θ increases with β, implying that customers are more likely to redeem coupons as their value increases. Our findings remain valid under this alternative modeling framework.

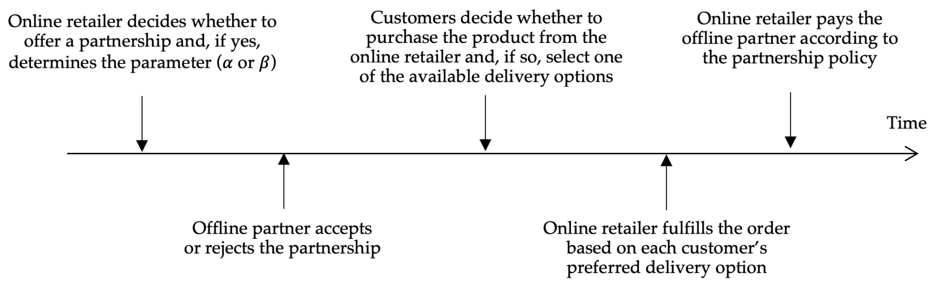


Figure 3.3.1: Timeline of Events

the optimal coupon value β that maximizes her profit $\pi_o^C(\beta)$, subject to the offline partner's rationality constraint $\pi_s^C(\beta) \geq 0$.

2. The offline partner accepts or rejects the partnership.

3. The customers decide whether to purchase the product from the online retailer and, if so, select one of the available delivery options.

4. The online retailer fulfills the order based on each customer's preferred delivery option.

5. The online retailer pays the offline partner according to the partnership policy.

Figure 3.3.2 reports the partnership policies considered by the online retailer and offline partner and the customer decision tree under each policy. Table 3.3.1 summarizes the customer utilities, online retailer's profit, and offline partner's profit under each policy. Appendix A2.1 summarizes the notation used throughout the paper.

3.4 Results

In this section, we first assess the feasibility and potential benefits of the fixed fee and coupon policies by comparing each policy to the baseline policy. We then compare all three policies to identify the most preferred policy for both

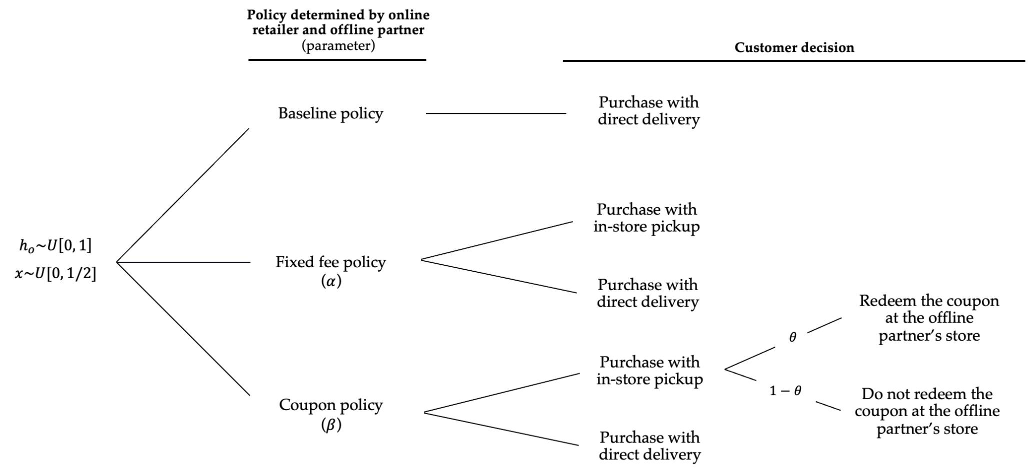


Figure 3.3.2: Partnership Policies and Customer Decision Trees

Table 3.3.1: Summary of Customer Utilities and Profit Functions

Policy	Customer Utility	Online Retailer's Profit	Offline Partner's Profit
Baseline	$u_o^B = v - p - h_o$	$(p - c_o)d_o^B$	0
Fixed fee	$u_o^F = v - p - h_o$ $u_p^F = v - p - xh_p$	$(p - c_o)d_o^F + (p - c_p - \alpha)d_s^F$	$(r + \alpha - c_s)d_s^F$
Coupon	$u_o^C = v - p - h_o$ $u_p^C = v - p - xh_p + \theta\beta$	$(p - c_o)d_o^C + (p - c_p - \theta\beta)d_s^C$	$(r + \theta\beta - c_s)d_s^C$

the online retailer and the offline partner. Finally, we examine whether a given policy is suitable for certain types of online retailers as characterized by the various model parameters.

3.4.1 Fixed Fee Policy

We start by comparing the fixed fee policy to the baseline policy. We let Δd_o^i, Δd_s^i, and Δd^i denote the differences in direct-delivery demand, in-store pickup demand, and total demand, respectively, between a pickup partnership policy, where $i \in \{F, C\}$, and the baseline policy. We thus have $\Delta d^i = \Delta d_o^i + \Delta d_s^i$. We derive analytical expressions for Δd_o^F, Δd_s^F, and Δd^F in the following proposition (details on the demand derivations appear in Appendix A2.2).

proposition 2. *Compared to the baseline policy, the fixed fee policy affects direct-delivery demand, in-store pickup demand, and total demand by*

$$\begin{aligned}
\Delta d_o^F &= -\underbrace{\frac{(v-p)^2}{h_p}}_{\substack{\textit{Demand shift due to}\\ \textit{convenience}}} \\
\Delta d_s^F &= \underbrace{\frac{2(v-p)(1-v+p)}{h_p}}_{\substack{\textit{Market expansion due to}\\ \textit{convenience}}} + \underbrace{\frac{(v-p)^2}{h_p}}_{\substack{\textit{Demand shift due to}\\ \textit{convenience}}} \\
\Delta d^F &= \underbrace{\frac{2(v-p)(1-v+p)}{h_p}}_{\substack{\textit{Market expansion due to}\\ \textit{convenience}}}
\end{aligned}$$

Proposition 2 reveals that the fixed fee policy induces two effects on the online retailer's demand relative to the baseline policy. First, some customers who were not purchasing under the baseline policy (due to the inconvenience caused by its high direct-delivery hassle cost) will now make a purchase via the in-store pickup option under the fixed fee policy. Specifically, these customers find visiting the offline partner's store to pick up their order more convenient due to the lower hassle cost. We call this effect the *market expansion effect* induced by the fixed fee policy. Second, due to the convenience of in-store pickups, some existing customers under the baseline policy will now choose this option. We call this effect the *demand shift effect* induced by the fixed fee policy. As a result, the fixed fee policy decreases direct-delivery demand due to its demand shift effect and creates a new demand stream through in-store pickups due to its market expansion and demand shift effects. Subsequently, the total demand increases only due to the market expansion effect since the demand shift effect simply transfers the existing demand from the direct-delivery option to the in-store pickup option.

A natural question that arises is how these demand changes affect the online retailer's and offline partner's profits, and whether it is beneficial to establish a pickup partnership under the fixed fee policy. We formally answer this question in Proposition 3. We let $\Delta\pi_o^i(\alpha)$ and $\Delta\pi_s^i(\alpha)$ denote the profit differences for the online retailer and the offline partner, respectively, between a pickup partnership policy, where $i \in \{F, C\}$, and the baseline policy.

proposition 3. *(a) Compared to the baseline policy, the fixed fee policy with α affects the online retailer's and offline partner's profits by*

$$\Delta\pi_o^F = \underbrace{\frac{2(v-p)(1-v+p)}{h_p}(p-c_p-\alpha)}_{\text{Increase due to market expansion}} + \underbrace{\frac{(v-p)^2}{h_p}(c_o-c_p-\alpha)}_{\text{Change due to demand shift}}$$

$$\Delta\pi_s^F = \underbrace{\frac{(v-p)(2-v+p)}{h_p}(r+\alpha-c_s)}_{\text{Change due to market expansion and demand shift}}$$

(b) It is beneficial for both parties to establish a pickup partnership under the fixed fee policy if and only if $\alpha \in [\underline{\alpha}, \bar{\alpha}]$, where $\underline{\alpha} = \max\{0, c_s - r\}$ and $\bar{\alpha} = p - c_p - \frac{v-p}{2-v+p}(p-c_o)$.

Proposition 3 shows that the two demand streams induced by the fixed fee policy are key to its profitability for both parties. For the online retailer, the market expansion effect induced by the fixed fee policy increases profit due to the additional margins obtained from the new customers. However, the impact of the demand shift effect on the profit is more intricate. When α is relatively small, the cost of direct-delivery (c_o) is higher than the cost of in-store pickup ($c_p + \alpha$). Subsequently, the demand shift effect results in another profit increase for the online retailer due to the additional margins obtained from existing customers who alter their delivery option when the pickup partnership is available. By contrast, a sufficiently high α will make the in-store pickup option more costly for the online retailer, so customers generating the demand shift effect will lower the profit. In that case, the gain from new customers will be sufficient to compensate the loss from existing customers so long as $\alpha \leq \bar{\alpha}$. Otherwise (i.e., when $\alpha > \bar{\alpha}$), the fixed fee policy will decrease the online retailer's profit. For the offline partner, the margin from each new or existing customer who opts for the in-store pickup option is equal to $r + \alpha - c_s$. When $\alpha \geq \underline{\alpha}$, this margin is positive, so that the fixed fee policy will benefit the offline partner. Otherwise (i.e., when $\alpha < \underline{\alpha}$), the offline partner will be worse off under the fixed fee policy. We note that even if there is no fixed fee compensation (i.e., $\alpha = 0$), the pickup partnership can still be beneficial for the offline partner when the profit from cross-selling purchases is high enough (i.e., $r > c_s$).

Proposition 3 also establishes that when the fixed fee compensation lies in $[\underline{\alpha}, \bar{\alpha}]$, neither party will be worse off under the fixed fee policy, resulting in a beneficial partnership.[5] Otherwise, one of the parties will always be worse off; hence, a pickup partnership will not be established under the fixed fee policy. We note that as α increases, the profit will decrease (resp., increase) for the online retailer (resp., offline partner). This implies that the optimal value of α is $\underline{\alpha}$ for the online retailer and $\bar{\alpha}$ for the offline partner. However, the optimal α that maximizes one party's profit will result in the other party

[5] The word beneficial is used in a non-strict sense throughout the paper.

earning no profit from the partnership. To avoid this situation and ensure that both parties are strictly better off under the fixed fee policy, the pickup partnership can be established by choosing a value of α such that $\underline{\alpha} < \alpha < \bar{\alpha}$. Such a well-designed partnership will yield a win-win situation, in a similar spirit as a revenue-sharing contract between a supplier and a retailer in supply chain management (Cachon and Lariviere 2005).

3.4.2 Coupon Policy

We compare the demand under the coupon and baseline policies in the following proposition.

proposition 4. *Compared to the baseline policy, the coupon policy affects direct-delivery demand, in-store pickup demand, and total demand by*

$$\Delta d_o^C = -\underbrace{\Big(\underbrace{\frac{(v-p)^2}{h_p}}_{\textit{due to convenience}} + \underbrace{\frac{(v-p)}{h_p}\hat{\beta}}_{\textit{due to promotion}}\Big)}_{\textit{Demand shift}}$$

$$\Delta d_s^C = \underbrace{\underbrace{\frac{2(v-p)(1-v+p)}{h_p}}_{\textit{due to convenience}} + \underbrace{\frac{(1-v+p)}{h_p}\hat{\hat{\beta}}}_{\textit{due to promotion}}}_{\textit{Market expansion}} + \underbrace{\underbrace{\frac{(v-p)^2}{h_p}}_{\textit{due to convenience}} + \underbrace{\frac{(v-p)}{h_p}\hat{\beta}}_{\textit{due to promotion}}}_{\textit{Demand shift}}$$

$$\Delta d^C = \underbrace{\underbrace{\frac{2(v-p)(1-v+p)}{h_p}}_{\textit{due to convenience}} + \underbrace{\frac{(1-v+p)}{h_p}\hat{\hat{\beta}}}_{\textit{due to promotion}}}_{\textit{Market expansion}}$$

where $\hat{\beta}$ and $\hat{\hat{\beta}}$ are increasing functions of β, as shown in Appendix A2.2. Proposition 4 conveys that, similar to the fixed fee policy, the coupon policy induces both market expansion and demand shift effects, albeit with higher magnitudes. This is due to the fact that the coupon policy provides two levers to influence demand. First, as in the fixed fee policy, it attracts new customers and shifts some existing customers from direct-delivery to in-store pickup due to the increased convenience of the in-store pickup option. Second, unlike the fixed fee policy, the coupon policy induces additional new and existing customers who are incentivized by the monetary value of the coupon. In particular, some customers who were not purchasing under the fixed fee policy (despite its convenience) will now make a purchase under the coupon policy to take advantage of the discount they receive at the offline partner store when picking up their online order. Consequently, the coupon policy induces a greater market expansion effect relative to the fixed fee policy. Similarly, some existing customers who use the direct-delivery option under both the baseline and the fixed fee policies will opt for the in-store pickup option under the coupon policy to take advantage of the coupon at the offline partner, resulting in a greater demand shift effect compared to that under the fixed fee policy. Thus, the pickup partnership's effects on direct-delivery demand, in-store pickup demand, and total demand are greater under the coupon policy than under the fixed fee policy.

Proposition 5 characterizes the corresponding change in profit for both partners and the condition when it is beneficial to establish a pickup partnership under the coupon policy.

proposition 5. *(a) Compared to the baseline policy, the coupon policy affects the online retailer's and offline partner's profit by*

$$\begin{aligned}
\Delta\pi_o^C &= \underbrace{\left(\frac{2(v-p)(1-v+p)}{h_p} + \frac{(1-v+p)}{h_p}\hat{\hat{\beta}}\right)(p-c_p-\theta\beta)}_{\text{Increase due to market expansion}} \\
&\quad + \underbrace{\left(\frac{(v-p)^2}{h_p} + \frac{(v-p)}{h_p}\hat{\beta}\right)(c_o-c_p-\theta\beta)}_{\text{Change due to demand shift}} \\
\Delta\pi_s^C &= \underbrace{\left(\frac{(v-p)(2-v+p)}{h_p} + \frac{(1-v+p)}{h_p}\hat{\hat{\beta}} + \frac{(v-p)}{h_p}\hat{\beta}\right)(r+\theta\beta-c_s)}_{\text{Change due to market expansion and demand shift}}
\end{aligned}$$

(b) *There exist two thresholds* $\underline{\beta}$ *and* $\bar{\beta}$ *such that it is beneficial for both parties to establish a pickup partnership under the coupon policy if and only if* $\beta \in [\underline{\beta}, \bar{\beta}]$. *Within this range, for the online retailer, the optimal coupon value is* $\beta^* = \max\{\underline{\beta}, \check{\beta}\}$.

The closed-form expression for $\check{\beta}$ is reported in Appendix A2.2. Proposition 5 shows that the profit implications of the demand change under the coupon policy are similar to those under the fixed fee policy. In short, the online retailer will be better off under the coupon policy so long as the additional profit margin earned from new customers induced by the market expansion effect offsets the loss from existing customers induced by the demand shift effect (i.e., when $\beta < \bar{\beta}$). Similarly, the offline partner will benefit from the partnership if she can collect a positive margin from customers who pick up their orders (i.e., when $\beta \geq \underline{\beta}$). Thus, when the monetary value of the coupon lies in $[\underline{\beta}, \bar{\beta}]$, no party is worse off under the coupon policy relative to the baseline policy.

As in the fixed fee policy, the offline partner prefers the highest possible value of the coupon (i.e., $\bar{\beta}$) to maximize her profit under the coupon policy, which comes at the expense of a zero gain for the online retailer. However, unlike the fixed fee policy, the optimal coupon value for the online retailer is not necessarily the minimum feasible value (i.e., $\underline{\beta}$) under the coupon policy. The rationale is that although an increase in β will decrease the profit margin from an in-store pickup order for the online retailer, it may also lead to more customers (both new and existing) opting for the in-store pickup delivery option to take advantage of the higher discount β. If the net profit from these customers offsets the decrease in profit margin per in-store pickup order, then the optimal coupon value for the online retailer would be $\beta^* > \underline{\beta}$. Such a coupon also provides a strictly positive gain from the partnership for the offline partner, resulting in a win-win situation for the two parties. Otherwise, the optimal coupon value for the online retailer is $\beta^* = \underline{\beta}$, which makes the offline partner indifferent between the baseline and coupon policies. In that case, as with the fixed fee policy, a win-win pickup partnership under the coupon policy can be established with $\underline{\beta} < \beta < \bar{\beta}$.

3.4.3 Optimal Policy

Having characterized each pickup partnership in the previous subsections, we next compare all three policies to identify the optimal policy for both the online retailer and the offline partner. To do so, we first find the optimal solution for each policy and then compare the three optimal solutions to determine the best policy. To ensure that the two pickup partnership policies are compared objectively, we impose the constraint that the average compensation per in-store pickup order is the same under both policies (i.e., $\alpha = \theta\beta$). Proposition 6 characterizes the results of this analysis conditional on c_s. We condition the analysis on c_s because it represents the offline partner's operational cost related to the partnership. Therefore, given that the process to establish a partnership starts with the online retailer selecting a partnership policy, the proposition conditioned on c_s can enable the online retailer to make that choice based on the offline partner's operational characteristics. We examine the sensitivity of the optimal policy with respect to other parameters in Section 3.4.4.

proposition 6. *There exist two thresholds $\underline{c}_s$ and $\bar{c}_s$ such that it is optimal for the online retailer and the offline partner*

- *not to establish a pickup partnership when $c_s > \bar{c}_s$,*
- *to establish a pickup partnership under the fixed fee policy with parameter $\alpha \in [c_s - r, \bar{c}_s - r]$ when $\underline{c}_s < c_s \leq \bar{c}_s$, and*
- *to establish a pickup partnership under the coupon policy with parameter $\beta \in [\max\{0, \frac{c_s - r}{\theta}\}, \frac{\underline{c}_s - r}{\theta}]$ when $c_s \leq \underline{c}_s$.*

The rationale behind Proposition 6 is as follows. When the offline partner's in-store pickup handling cost is high (i.e., $c_s > \bar{c}_s$), she finds the partnership beneficial only if the compensation for each in-store pickup order (i.e., α or β) is sufficiently high. However, such a high compensation makes the online retailer worse off with any pickup partnership (as we show in Figure 3.4.1, when $c_s > \bar{c}_s$, the online retailer is better off under the baseline policy). Therefore, a beneficial partnership does not exist, making the baseline policy the best option.

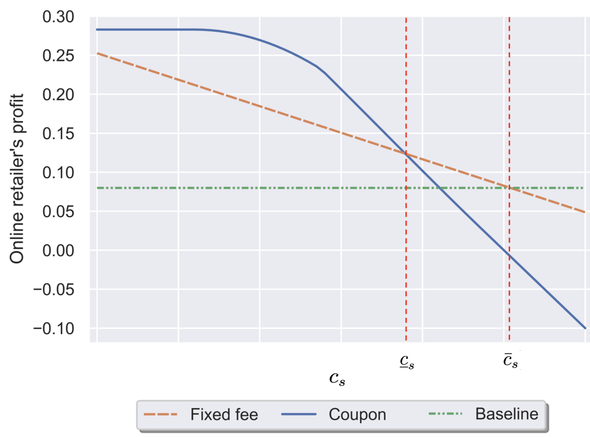


Figure 3.4.1: The Online Retailer's Profit under the Fixed Fee, Coupon, and Baseline Policies

When a beneficial partnership exists (i.e., $c_s \leq \bar{c}_s$), the optimal policy is determined by the additional customers (both new and existing) who opt for the in-store pickup delivery option only under the coupon policy (i.e., customers forming the "market expansion due to promotion" and "demand shift due to promotion" segments shown in Proposition 4). When the in-store pickup handling cost is moderate (i.e., $\underline{c}_s < c_s \leq \bar{c}_s$), a beneficial partnership (under both fixed fee and coupon policies) can be established only with a moderately high compensation (β or α). In this case, the coupon policy increases the profit by boosting the demand relative to the fixed fee policy. However, a moderately high compensation makes the in-store pickup fulfillment more costly (and thus less profitable) than direct-delivery fulfillment for the online retailer. Therefore, the coupon policy also decreases the profit by making more existing customers who would choose the direct-delivery option under the fixed fee policy switch to the more costly in-store pickup option. When the coupon value is moderately high, the increase in profit due to the demand boost effect cannot offset the loss due to the demand shift effect. As a result, the online retailer is worse off under the coupon policy with a moderately high β than under the fixed fee policy with an economically equivalent compensation (i.e., a moderately high α), making the fixed fee policy optimal. This can be seen in Figure 3.4.1, which shows that the online retailer's profit is higher under the fixed fee policy when $\underline{c}_s < c_s \leq \bar{c}_s$.

Finally, when the in-store pickup handling cost is low (i.e., $c_s \leq \underline{c}_s$), a beneficial partnership under both fixed fee and coupon policies can be established with low compensation. In this case, the net profit from the additional new and existing customers induced by the coupon policy with a low β makes the online retailer better off relative to the fixed fee policy. Thus, the coupon policy is optimal when $c_s \leq \underline{c}_s$.

3.4.4 Comparative Statics

In this section, we investigate whether a given optimal policy is more suitable for certain types of online retailers that can be characterized based on three key model parameters; namely c_o, c_p, and p. To do so, we examine the sensitivity of the optimal policy with respect to c_o, c_p, and p. All the technical details related to this analysis appear in Appendix A2.5.

We make three main observations. First, we find that the benefit of the pickup partnership increases as the direct-delivery fulfillment cost (c_o) increases and the in-store pickup fulfillment cost (c_p) decreases. Under a high c_o and a low c_p, the direct-delivery becomes a more costly (and less profitable) fulfillment method for the online retailer relative to the in-store pickup delivery. Consequently, the online retailer will earn a higher profit with the pickup partnership (under either a fixed fee or a coupon policy) due to customers choosing the relatively less costly in-store pickup delivery option. Second, between the two partnership policies, the coupon policy outperforms the fixed fee policy as c_o increases and c_p decreases. Since the coupon policy generates more in-store pickup demand than the fixed fee policy (as characterized by Proposition 4), then the increasing benefit of the pickup partnership (as c_o increases and c_p decreases) becomes more pronounced under the coupon policy. Third, as the product price (p) increases, the coupon policy becomes more profitable than the fixed fee policy. An increase in p has two competing effects. While it increases the profit margin for both direct-delivery orders and in-store pickup orders, it also leads to lower demand for both types of orders. Since the coupon policy will generate a higher demand than the fixed fee policy (due to its promotional lever), the negative impact of the price increase on demand is more mitigated under the coupon policy.

Overall, as summarized in Table 3.4.1, these results suggest that a pickup partnership is not suitable for online retailers with low direct-delivery fulfillment cost (c_o) or high in-store pickup fulfillment cost (c_p) (e.g., online jewelry and luxury fashion retailers). The fixed fee policy is suitable for retailers with moderate direct-delivery cost, moderate in-store fulfillment cost, or low-priced products (e.g., online farmer marketplaces, online supermarkets), whereas the coupon policy is suitable for retailers with high direct-delivery cost, low in-store fulfillment cost, or high-priced products (e.g., meal kit companies, cosmetics retailers).

Table 3.4.1: Optimal Policy based on Online Retailer Characteristics

		$c_o(c_p)$		
		High (Low)	Moderate (Moderate)	Low (High)
p	High	Coupon	Coupon/Fixed fee	Baseline
	Low	Coupon/Fixed fee	Fixed fee	Baseline

3.5 Extensions

Having established the fundamental features of the pickup partnership under two different policies, we now consider several extensions of our theoretical model. In particular, we examine the pickup partnership with (i) a budget constraint, (ii) multiple pickup locations, and (iii) the consideration of consumer surplus. In this section, we summarize the results from these analyses; the technical details appear in Appendix A2.6.

3.5.1 Budget Constraint

In our main model, we assumed that the online retailer was willing to pay any compensation (α or β) for each in-store pickup so long as it is profitable. In practice, however, the online retailer may have a limited budget to establish a pickup partnership. In this subsection, we evaluate how such a budget constraint affects the optimal pickup partnership policy.

We define the partnership budget K as the total amount of money allocated by the online retailer to compensate the offline partner for all in-store pickup orders. Note that the expected overall compensation to meet all potential in-store pickup orders equals αd_s^F under the fixed fee policy and $\theta\beta d_s^C$ under the coupon policy (we highlight that d_s^F and d_s^C are different). Therefore, the budget constraint imposes that the expected overall compensation for in-store pickup orders should not exceed the online retailer's budget (i.e., $\alpha d_s^F \leq K$ under the fixed fee policy and $\theta\beta d_s^C \leq K$ under the coupon policy).[6] Under a budget constraint, we make three observations.

First, when the budget is small (i.e., $K < \underline{K}$ where the expression of $\underline{K}$ is provided in Appendix A2.6), for each in-store pickup order, the online retailer can compensate the offline partner with a maximum fixed fee of $\alpha < \underline{\alpha}$ under the fixed fee policy and a maximum coupon value of $\beta < \underline{\beta}$ under the coupon policy. As Propositions 3 and 5 imply, such compensation levels are not high enough to incentivize the offline partner to accept the pickup partnership under any policy. As a result, the pickup partnership cannot be established under a small budget. Second, when the budget is moderate (i.e., $\underline{K} \geq K > \bar{K}$; the expression of $\bar{K}$ appers in Appendix A2.6), the maximum

[6] Alternatively, the budget can also be defined per in-store pickup order such that the compensation for each in-store pickup order cannot exceed a certain budget. Under this type of budget constraint, we find consistent results.

compensation that the online retailer can pay is $\alpha > \underline{\alpha}$ under the fixed fee policy and $\beta < \underline{\beta}$ under the coupon policy, making the partnership beneficial only under a fixed fee policy. In this case, it is beneficial for both parties to establish a pickup partnership under the fixed fee policy so long as $\alpha \in [\underline{\alpha}, \frac{K}{d_s^F}]$. Third, when the budget is large (i.e., $K \geq \bar{K}$), both policies are beneficial. In this case, the online retailer can choose an optimal policy as characterized in Proposition 6, along with the budget consideration. As such, an optimal fixed fee policy can be implemented with parameter $\alpha \in [\underline{\alpha}, \min\{\bar{c}_s - r, \frac{K}{d_s^F}\}]$. Similarly, an optimal coupon policy can be implemented with parameter $\beta \in [\underline{\beta}, \min\{\frac{\underline{c}_s - r}{\theta}, \frac{K}{\theta d_s^C}\}]$.

In summary, the budget allocated by the online retailer to the pickup partnership will determine the policy to be chosen. Due to the additional demand it generates for in-store pickup orders, the coupon policy requires a higher budget than the fixed fee policy. Therefore, for online retailers with a limited budget, a pickup partnership is beneficial only under a fixed fee policy. A partnership with a coupon policy is beneficial only for online retailers with a higher budget.

3.5.2 Multiple Pickup Locations

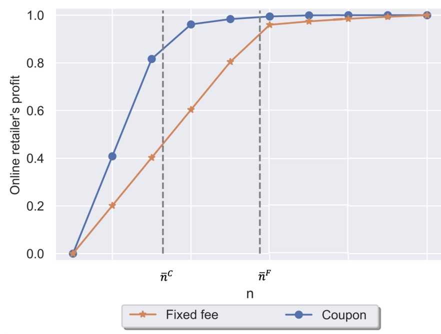
So far, we have assumed that the offline partner had only one pickup location. While the single-location assumption allows us to characterize the fundamental features of pickup partnership policies, in practice, offline partners can designate multiple stores as pickup locations. In such cases, it is managerially important to understand (i) whether the optimal pickup partnership policy depends on the number of in-store pickup locations, and (ii) how retailers should determine the optimal number of in-store pickup locations for a given pickup partnership policy. Next, we extend our model to examine the scenario with multiple pickup locations.

We assume that the offline partner has $n \geq 1$ stores that are spread uniformly along the circumference of a circular city. Thus, the distance between two nearby stores is $1/n$. Since customers are uniformly distributed on the circumference of the circular city (Salop 1979), $x \sim U[0, 1/2n]$. This implies that customers' hassle cost for visiting the nearest store to pick up an order (i.e., $h_p x$) decreases as the number of pickup locations increases (regardless of partnership policy). Consequently, more customers will find the in-store pickup option more convenient than the direct-delivery option, increasing the demand for in-store pickups.

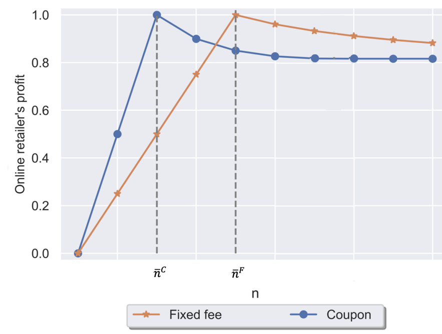
Under this modeling framework, we find that the number of pickup locations can change the optimal policy structure derived in Proposition 6 only when the coupon policy is optimal (i.e., when $c_s \leq \underline{c}_s$). In Figure 3.5.1, we plot the relationship between the number of pickup locations and the optimal policy structure for that case. Since the demand for in-store pickup orders increases with n, once the number of pickup locations in the pickup partnership reaches a certain threshold, the online retailer will cover the entire market (i.e., each customer in the market will make a purchase via either the direct-delivery or in-store pickup option). The notations $\bar{n}^F$ and $\bar{n}^C$ in Figure 3.5.1 represent this threshold for the fixed fee and coupon policies, respectively, when $c_s \leq \underline{c}_s$. Since the total demand (i.e., demand for both direct-delivery and in-store pickup) is higher under the coupon policy relative to the fixed fee policy (Proposition 4), the online retailer can cover the entire market with fewer pickup locations under the coupon policy (i.e., $\bar{n}^C \leq \bar{n}^F$).

We next summarize our findings. First, the online retailer's profit increases with n under both partnership policies so long as $n < \bar{n}^C$ and a partnership is beneficial. Indeed, as n increases, the market expansion and demand shift effects under both policies become more substantial. Consequently, the coupon policy remains more profitable than the fixed fee policy. Second, when $\bar{n}^C \leq n < \bar{n}^F$, since the market is still not fully covered under the fixed fee policy, the market expansion and demand shift effects are still at play, and thus, profit continues to rise with n under the fixed fee policy. However, since the entire market is fully covered under the coupon policy, any increase in n will only amplify the demand shift effect of the coupon policy, whereas the market expansion effect remains the same. If the margin from an in-store pickup order is greater than the margin from a direct-delivery order (as illustrated in Figure 3.5.1(a)), then the online retailer's profit will continue to increase with n, albeit at a decreasing rate, under the coupon policy. In this case, the coupon policy still remains more profitable than the fixed fee policy. By contrast, if the in-store pickup fulfillment is more costly to the online retailer (as illustrated in Figure 3.5.1(b)), the profit under the coupon policy will decrease with n due to customers changing their preferences from direct delivery to in-store pickup. In this case, the two policies become equally profitable with a certain number of pickup locations (refer to the point in Figure 3.5.1(b) at which both lines intersect), and any increase in n beyond that point will make the fixed fee policy optimal. Third, when $n > \bar{n}^F$, the entire market is covered under both policies. Therefore, any increase in n will only amplify the demand shift effect under both policies while keeping the relative profitability the same. Hence, the coupon policy is optimal when the direct-delivery fulfillment is more costly (Figure 3.5.1(a)), whereas the fixed fee policy is optimal when the in-store pickup fulfillment is more costly (Figure 3.5.1(b)). Fourth, when n becomes large enough, the entire market makes a purchase only via the in-store pickup option under both policies, so that the online retailer becomes indifferent between the two policies (as illustrated in Figure 3.5.1(a)).

The above discussion suggests that the optimal number of pickup locations for a partnership depends on the comparison between the profit margin from an in-store pickup order and the profit margin from a direct-delivery order. When the margin is higher for the in-store pickup order, as illustrated in Figure 3.5.1(a), the minimum number of pickup locations, which enables the online retail to fully cover the market using only the in-store pickup option, is optimal under both partnership policies. When the margin is higher for the direct-delivery order, as illustrated in Figure 3.5.1(b), then the optimal number of pickup locations is equal to the smallest number that ensures



(a) Direct-delivery Fulfillment is more Costly



(b) In-store Pickup Fulfillment is more Costly

Figure 3.5.1: The Effect of **n** on the Online Retailer's Profit under Fixed Fee and Coupon Policies

full market coverage (i.e., $\bar{n}^F$ for the fixed fee policy and $\bar{n}^C$ for the coupon policy).

Overall, our analysis suggests that a larger number of pickup locations does not necessarily yield a higher profit for the partnership. Before negotiating the number of pickup locations, online retailers should carefully compare their profit margins between the direct-delivery and in-store pickup options. When the in-store pickup fulfillment is more costly than the direct-delivery fulfillment, online retailers can establish a partnership under the coupon policy using a smaller number of pickup locations. Otherwise, a larger number of locations will increase the partnership profitability.

3.5.3 Total Welfare

In our main model, the optimal partnership policy was chosen to maximize the profit of the online retailer while ensuring that the offline partner is not worse off relative to the setting without a partnership. With the increasing awareness of social responsibility, retailers may alternatively seek partnership solutions that also consider the customers' interests. Therefore, in this subsection, we evaluate how the optimal policy changes when the online retailer seeks to maximize the total welfare earned by all stakeholders (i.e., herself, the offline partner, and customers).

Our analysis reveals that when maximizing total welfare (i.e., the sum of the online retailer profit, offline partner profit, and consumer surplus), the threshold $\bar{c}_s$ increases. In other words, the range under which a beneficial pickup partnership exists will expand. This result is expected because the pickup partnership (under any policy) will always increase the consumer surplus relative to the baseline policy, since it offers an additional delivery option for customers. Therefore, the increase in consumer surplus is yet another benefit of establishing a pickup partnership. It is worth noting that even in this case, the pickup partnership may still not be beneficial for high values of c_s.

More importantly, our analysis shows that when setting the objective as the total welfare, the fixed fee is no longer the optimal policy for a pickup partnership. Recall that as shown in Proposition 6, when the online retailer maximizes her own profit, the fixed fee policy is optimal when the profit from the additional new customers induced by the coupon policy cannot offset the loss from the additional existing customers (switching from the direct-delivery to in-store pickup option), again induced by the coupon policy. When the online retailer maximizes total welfare, the consumer surplus is higher under the coupon policy than under the fixed fee policy, since the probability of redeeming the coupon provides an additional utility to customers. Consequently, compared to the fixed fee policy, the profit from additional new customers, coupled with the higher consumer surplus, will always offset the loss from additional existing customers under the coupon policy. Therefore, the online retailer is always better off by establishing the pickup partnership under the coupon policy relative to the fixed fee policy when considering the total welfare.

3.6 Pickup Partnership with Hybrid Policy

An optimal partnership policy, as we identified in Proposition 6, aims to maximize the online retailer's profit subject to the offline partner's rationality constraint. Thus, to establish a pickup partnership, the online retailer has to offer either a fixed fee policy or a coupon policy while ensuring that the offline partner is not worse off relative to the baseline policy. In some situations, this may force the online retailer to select a partnership parameter that is not necessarily a profit maximizer for herself, implying that an optimal partnership policy can entail inefficiencies for the online retailer. Equivalently, we investigate when the offline partner's rationality constraint is tight. In this section, we first examine under which cases such inefficiencies exist for the two policies (fixed fee and coupon) used in current practices. We then prescribe a novel pickup partnership policy that alleviates such inefficiencies.

3.6.1 Inefficiency from Fixed Fee and Coupon Policies

We start by examining the optimal fixed fee policy. Recall from Proposition 6 that, when $\underline{c}_s < c_s \leq \bar{c}_s$, the online retailer prefers the fixed fee policy over the coupon policy with a moderately high β^*. Figure 3.6.1 illustrates the corresponding market segmentation under the optimal fixed fee policy. We observe that the optimal fixed fee policy with α generates a partial market coverage so that some customers (as depicted by the dotted region at the top-right corner in Figure 3.6.1) leave the market without making a purchase. Note that the online retailer can achieve the same market segmentation under the coupon policy by setting the coupon value to zero (i.e., $\beta = 0$). Consistent with Proposition 4, this implies that with any positive coupon value under a hypothetical coupon policy, the online retailer would generate additional sales from some of the customers in the dotted region, although a positive coupon value would also motivate some existing customers to change their delivery mode from direct-delivery to in-store pickup. As shown in Proposition 6, when the coupon value is low enough, the net profit change with any positive coupon value under that hypothetical coupon policy relative to the optimal fixed fee policy would be positive, representing an opportunity cost for the online retailer under the optimal fixed fee policy. When $\underline{c}_s < c_s \leq \bar{c}_s$, since the online retailer can encourage the offline partner to establish a partnership under the coupon policy using only a moderately high β^* (which makes the net profit change under the coupon policy compared to the optimal fixed fee policy negative), she prefers to establish the pickup partnership under a fixed fee policy despite its opportunity cost. Therefore, the aforementioned opportunity cost represents the inefficiency of the pickup partnership under the optimal fixed fee policy (relative to the hypothetical coupon policy with a low β).

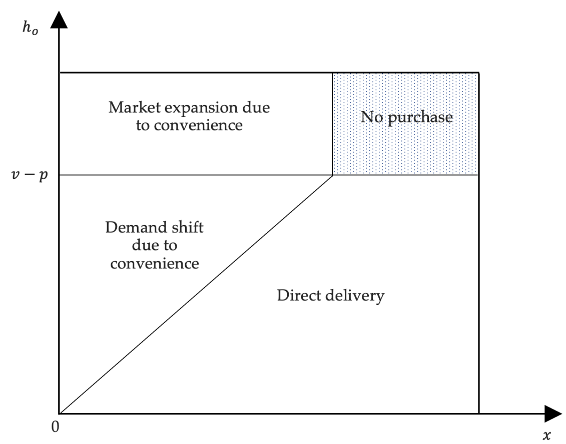


Figure 3.6.1: Market Segmentation under the Optimal Fixed Fee Policy

We next examine the optimal coupon policy. As discussed in Proposition 6, when $c_s \leq \underline{c}_s$, the online retailer can induce the offline partner to establish a pickup partnership under the coupon policy with a relatively low β^*, making the coupon policy optimal. Figure 3.6.2 illustrates the corresponding market segmentation under the optimal coupon policy. We observe that the optimal coupon policy with β^* allows the online retailer to cover the entire market (i.e., all customers will make a purchase via direct-delivery or in-store pickup). However, as shown in Figure 3.6.2, β^* under the optimal coupon policy is greater than the minimum coupon value β_m under a hypothetical coupon policy that can allow the online retailer to just cover the entire market. This implies that the market expansion effect of the coupon policy is maximized when $\beta = \beta_m < \beta^*$. Thus, when the coupon value increases from β_m to β^*, the online retailer no longer generates new customers. Rather, as illustrated by the dotted region in Figure 3.6.2, an increase in the coupon value beyond β_m only induces more existing customers to change their delivery option from direct-delivery to in-store pickup. When the profit margin from an in-store pickup order is lower than the profit margin from a direct-delivery order, the customers in the dotted region will decrease the online retailer's profit relative to the profit under the hypothetical coupon policy with β_m. Despite this profit loss, when $c_s \leq \underline{c}_s$, the online retailer constructs the optimal coupon policy with $\beta^* > \beta_m$, because any β lower than β^* will make the offline partner worse off under the partnership. Thus, in order to fairly compensate the offline partner under the optimal coupon policy, the online retailer will absorb the loss from the customers in the dotted region. Hence, the absorbed loss from these customers represents the inefficiency of the pickup partnership under the optimal coupon policy (relative to the hypothetical coupon policy with β_m).

Overall, we find that, despite being optimal, both the fixed fee and coupon policies may entail inefficiencies for the online retailer. In the next subsection, we prescribe an alternative partnership policy to alleviate such inefficiencies.

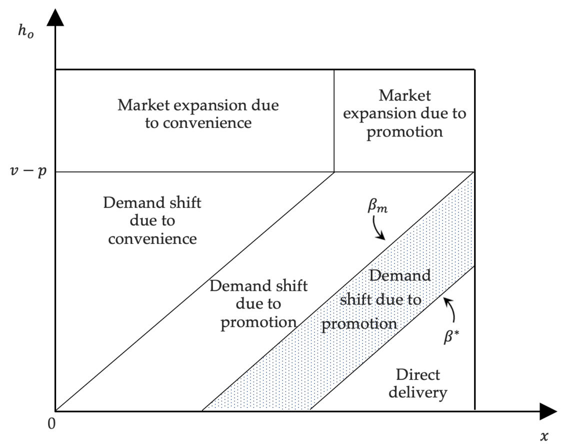


Figure 3.6.2: Market Segmentation under the Optimal Coupon Policy

3.6.2 Hybrid Policy

We show that an alternative partnership can be established such that for each in-store pickup, the online retailer can compensate the offline partner with a total compensation γ, of which α_h is paid to the offline partner as a fixed fee, and β_h (where $\beta_h = \frac{\gamma-\alpha_h}{\theta}$) is offered to the customers as a coupon. We term this policy the *hybrid policy*. We note that in this setting, the average compensation per in-store pickup order becomes equivalent to those under the fixed fee and coupon policies (i.e., $\gamma = \alpha = \theta\beta$).

Customer utilities for direct-delivery and in-store pickup options under the hybrid policy remain the same as under the coupon policy. We let d_o^H and d_p^H denote the online retailer's demand for direct-delivery and in-store pickup options, respectively, where the superscript H denotes the hybrid policy. Then, the online retailer's profit is equal to

$$\pi_o^H(\gamma) = (p - c_o)d_o^H + (p - c_p - \gamma)d_s^H, \tag{3.6.1}$$

whereas the offline partner's profit is equal to

$$\pi_s^H(\gamma) = (r + \gamma - c_s)d_s^H. \tag{3.6.2}$$

Proposition 7 characterizes how the optimal policy structure presented in Proposition 6 changes in the presence of the hybrid policy.

proposition 7. *There exist thresholds* $\bar{c}_p$, $\underline{\underline{c}}_s$, *and* $\bar{\bar{c}}_s$ *such that*

(a) if $c_p < \bar{c}_p$, *it is optimal for the online retailer and the offline partner*

- *not to establish a pickup partnership when* $c_s > \bar{c}_s$,
- *to establish a pickup partnership under the fixed fee policy with parameter* $\alpha \in [c_s - r, \bar{c}_s - r]$ *when* $\bar{\bar{c}}_s < c_s \leq \bar{c}_s$,
- *to establish a pickup partnership under the hybrid policy with parameters* $\beta_h = \beta_m$ *and* $\alpha_h \in [\max\{c_s - r - \theta\beta_m, 0\}, \bar{\bar{c}}_s - r - \theta\beta_m]$ *when* $\underline{\underline{c}}_s < c_s \leq \bar{\bar{c}}_s$, *and*
- *to establish a pickup partnership under the coupon policy with parameter* $\beta \in [\max\{0, \frac{c_s - r}{\theta}\}, \frac{\underline{\underline{c}}_s - r}{\theta}]$ *when* $c_s \leq \underline{\underline{c}}_s$.

(b) otherwise (i.e., $c_p \geq \bar{c}_p$*), the optimal policy structure from Proposition 6 remains the same.*

The closed-form expressions for $\bar{c}_p$, $\underline{\underline{c}}_s$, and $\bar{\bar{c}}_s$ appear in Appendix A2.4. Proposition 7 reveals that the hybrid policy can be a valuable lever for the online retailer to improve partnership efficacy only when the profit margin of the in-store pickup order is smaller than the profit margin of the direct-delivery order (i.e., $c_p < \bar{c}_p$ and $\underline{\underline{c}}_s < c_s \leq \bar{\bar{c}}_s$). In this case, the optimal policy structure characterized in Proposition 6 (labeled as "Without hybrid policy" in Figure 3.6.3) changes as shown by the "With hybrid policy" case in Figure 3.6.3. The figure shows that the retailer is better off under the hybrid policy relative to the coupon policy in Region I and relative to the fixed fee policy in Region II.

In Region I of Figure 3.6.3 (i.e., when $\underline{\underline{c}}_s \leq c_s < \underline{c}_s$), in the absence of the hybrid policy, the online retailer establishes the partnership using the coupon policy with a relatively low β^* while absorbing the loss from the customers in

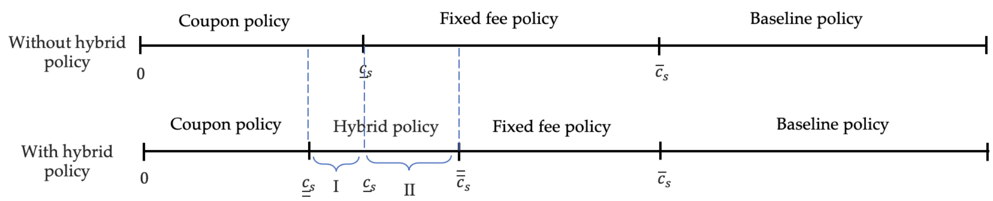


Figure 3.6.3: Optimal Policy with and without the Hybrid Policy

the dotted region of Figure 3.6.2, as discussed in Section 3.6.1. The hybrid policy enables the online retailer to mitigate this loss. In particular, as illustrated in Figure 3.6.4, the online retailer will set the optimal hybrid policy coupon value β_h^* to β_m to cover the entire market, making the customers in Region A better off with the direct-delivery option and hence eliminating the loss from these customers. Since β_h^* is not high enough for the offline partner to accept the partnership, the online retailer will pay the remaining α_h of the hybrid policy compensation γ as a fixed fee to ensure that the offline partner is not worse off under the partnership.

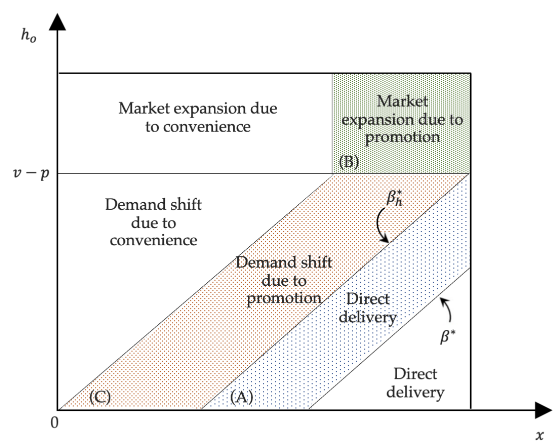


Figure 3.6.4: Market Segmentation under the Optimal Hybrid Policy

In Region II of Figure 3.6.3 (i.e., when $\underline{c}_s \leq c_s < \bar{\bar{c}}_s$), in the absence of the hybrid policy, the online retailer establishes the partnership using the fixed fee policy, while missing the additional profit opportunity from customers who leave the market without making any purchase (i.e., customers in the blue dotted region in Figure 3.6.1), as discussed in Section 3.6.1. As illustrated in Figure 3.6.4, the hybrid policy enables the online retailer to attract those customers. In particular, by setting the optimal hybrid policy coupon value β_h^* to β_m, the online retailer will cover the entire market, including the customers in Region B, thus generating an additional profit. However, this will also induce some of the existing customers (illustrated by Region C in Figure 3.6.4) to change their preference from direct-delivery under the fixed fee policy to in-store pickup under the hybrid policy, generating an additional loss. Since β_h^* represents a low β, as implied by Proposition 6, the former's additional profit becomes greater than the latter's additional loss, ultimately making the online retailer better off under the hybrid policy.

We next conduct a numerical study to quantify the extent to which the hybrid policy can benefit the online retailer. Using a wide range of model parameters, our study results in 5,540 instances for which the partnership is beneficial to both parties under either the fixed fee or coupon policies.[7] We find that the hybrid policy can improve the online retailer's profit for 24.06% (1,333 of 5,540) of those instances. The average improvement in the online retailer's profit amounts to 5.14% (with a standard deviation of 3.13%), with minimum and maximum values of 0.78% and 14.45%.

To summarize, our proposed hybrid policy allows the online retailer to minimize the potential inefficiencies of the pickup partnership established under either the fixed fee policy or the coupon policy while ensuring that the partnership remains beneficial for the offline partner. To our knowledge, such a hybrid policy has not yet been deployed in practice, potentially due to

[7]The parameter values are as follows: $p \in [0.7, 0.95]$ with 0.05 increments, $c_o \in [0.05p, 0.25p, 0.5p, 0.75p, 0.95p]$, $c_p \in [c_o, 0.75c_o, 0.5c_o, 0.25c_o, 0]$, $h_p \in [0.45, 0.75]$ with 0.1 increments, and $c_s \in [0, 1]$ with 0.05 increments. Other parameters are set to $v = 1$, $r = 0$, and $\theta = 1$. After dropping the instances that do not satisfy model assumptions, the numerical study results in 5,946 instances, of which, 5,540 (i.e., 93.17%) lead to a beneficial pickup partnership for both parties.

the fact that pickup partnerships in retail are still in a fairly nascent stage. Nevertheless, our results suggest that online retailers should consider implementing a hybrid policy when establishing a pickup partnership since such a policy is likely to yield a more efficient partnership, especially when in-store pickup delivery fulfillment is costly.

3.7 Conclusion

To survive in the omnichannel era, many pure online retailers (e.g., Amazon, Cookit, Maturin) have recently started to form pickup partnerships with offline stores to provide their customers with convenient in-store pickup services. Despite this evolving business model, the literature on how to design and assess the impact of a pickup partnership is non-existent. To our knowledge, this paper is the first to develop an analytical model to theoretically examine the benefits of implementing a pickup partnership. In particular, we first analyze the two policies, the fixed fee and the coupon policies, that practitioners use to form pickup partnerships. We then characterize whether and how online retailers should choose between these two policies. We also convey that despite being optimal, both policies entail inefficiencies leading to an opportunity cost for the online retailer. We then prescribe a new type of policy that mitigates such inefficiencies.

Our results indicate that the cost structures of the online retailer and offline partner determine whether the two parties should form a pickup partnership and if so, which policy they should use. Specifically, we find that the coupon policy is particularly suitable for offline partners that can manage the in-store pickup process efficiently (i.e., those with low in-store pickup handling cost) and online retailers with a high direct-delivery fulfillment cost, a low in-store pickup fulfillment cost, or high-priced products. In contrast, the fixed fee policy is suitable for offline partners with a moderate in-store pickup handling cost and online retailers with a moderate direct-delivery fulfillment cost, a moderate in-store fulfillment cost, or low-priced products. We also find that the partnership will not be beneficial for offline partners with a high in-store pickup handling cost or for online retailers with a low direct-delivery fulfillment cost or a high in-store fulfillment cost.

We later extend our model to examine the pickup partnership under three different scenarios. First, we find that the presence of a budget constraint may reduce the online retailer's willingness to adopt a coupon policy, making the fixed fee policy a more suitable option for online retailers with limited budgets. Second, we consider an offline partner with multiple pickup locations and find that a larger number of pickup locations does not necessarily yield a higher profit for the partnership, especially when in-store pickup fulfillment is more costly than direct-delivery fulfillment. Third, we consider an online retailer maximizing total welfare for all parties (including customers) and find that the coupon policy always outperforms the fixed fee policy due to the additional utility earned by customers from the discounted coupons.

Finally, we find that both policies entail inefficiencies due to the misaligned incentives of the online retailer. More precisely, to ensure that the offline partner is not worse off with the partnership (relative to no partnership), in some cases the online retailer has to propose a partnership that does not necessarily maximize her own profit. In such cases, the online retailer cannot fully leverage the potential of the partnership. To address this issue, we propose a hybrid policy that allows the online retailer to split the compensation amount per in-store pickup order between the offline partner (in the form of a fixed fee) and customers (in the form of a coupon). We then show that our proposed hybrid policy allows the online retailer to minimize these inefficiencies, while ensuring that the pickup partnership remains attractive for the offline partner. In a numerical study, we find that such inefficient partnerships can be common and that the profit improvement generated by the hybrid policy is substantial.

Our results provide several managerial implications for retailers seeking to establish a pickup partnership. First, a newly established partnership may require a fixed investment cost for the offline partner (e.g., setting up a pickup point in the store, assigning staff to process pickups) and the volume of in-store pickup orders is likely to increase over time as customers become more familiar with this service. Thus, the offline partner's handling cost is likely to be high in a newly established partnership. However, it will likely decline over time as the volume of in-store pickup orders increases and the offline partner improves the process through learning-by-doing. Therefore, a direct implication of our study is that a pickup partnership should initially be established using the fixed fee policy. As the offline partner becomes more efficient in processing in-store pickup orders, both parties can be better off by switching from the fixed fee policy to the coupon policy. In fact, our glance at the current pickup partnerships in the industry reveals an observation consistent with our implication. We observe that at this nascent stage of pickup partnerships, as implied by our study, firms mostly prefer the fixed fee policy relative to the coupon policy. Second, our results suggest that the parties are better off by customizing the partnership not only based on the stage of their relationship but also on the specific characteristics of the business setting. For example, for low-priced staple items, the fixed fee policy is more beneficial, whereas for high-priced niche items, the coupon policy is better. Since the hybrid policy encompasses both the fixed fee and coupon policies, it allows the online retailer to be more flexible in the partnership implementation. Third, although our results show that the optimal parameters for both the fixed fee and coupon policies depend on the characteristics of the business setting, we believe that the coupon policy may be an easier-to-implement option. For example, under the fixed fee policy, if the online retailer wants to set a fee that depends on the product price, both the online and offline partners need to keep track of actual transactions to determine the total transfer amount owing to the pickup partnership. Under the coupon policy, however, the online retailer can easily set a different coupon value based on the product price and make a transfer to the offline partner based on the redeemed amount.

Admittedly, more studies are needed to examine the growing trend of pickup partnerships. We can think of at least two potential avenues for future research. First, it would be interesting to investigate how the online retailer's and offline partner's operational decisions (e.g., assortment, price, inventory decisions) are affected by partnership policies. Second, it might be beneficial to empirically investigate the long- and short-term effects of each partnership policy for both the online retailer and the offline partner.

Bibliography

Akturk, M. S. and Ketzenberg, M. (2021). Exploring the competitive dimension of omnichannel retailing, *Management Science.* . ePub ahead of print July 9, `https://doi.org/10.1287/mnsc.2021.4008`.

Baardman, L., Cohen, M. C., Panchamgam, K., Perakis, G. and Segev, D. (2019). Scheduling promotion vehicles to boost profits, *Management Science* **65**(1): 50–70.

Balasubramanian, S. (1998). Mail versus mall: A strategic analysis of competition between direct marketers and conventional retailers, *Marketing Science* **17**(3): 181–195.

Bell, D., Gallino, S. and Moreno, A. (2018). Offline showrooms in omnichannel retail: Demand and operational benefits, *Management Science* **64**(4): 1629–1651.

Boston Retail Partners (2021). POS/Customer engagement survey. Technical report. `https://www.retailconsultingpartners.com/2021-pos-survey-pr`.

Cachon, G. and Lariviere, M. (2005). Supply chain coordination with revenue-sharing contracts: Strengths and limitations, *Management Science* **51**(1): 30–44.

Cao, J., So, K. C. and Yin, S. (2016). Impact of an "online-to-store" channel on demand allocation, pricing and profitability, *European Journal of Operational Research* **248**(1): 234–245.

Chen, J., Liang, Y., Shen, H., Shen, Z.-J. M. and Xue, M. (2021). Offline-channel planning in smart omnichannel retailing, *Manufacturing & Service Operations Management.* . ePub ahead of print December 20, `https://doi.org/10.1287/msom.2021.1036`.

Chevalier, S. (2021). Change in click and collect retail sales in the United States from 2019 to 2024. `https://www.statista.com/statistics/1132011/click-and-collect-retail-sales-growth-us/`.

Cohen, M. C. and Zhang, R. (2022). Competition and coopetition for two-sided platforms, *Production and Operations Management.* . ePub ahead of print January 11, `https://doi.org/10.1111/poms.13661`.

Dan, B., Zhang, H., Zhang, X., Guan, Z. and Zhang, S. (2021). Should an online manufacturer partner with a competing or noncompeting retailer for physical showrooms?, *International Transactions in Operational Research* **28**(5): 2691–2714.

Danaher, P., Smith, M., Ranasinghe, K. and Danaher, T. (2015). Where, when, and how long: Factors that influence the redemption of mobile phone coupons, *Journal of Marketing Research* **52**(2): 710–725.

Ertekin, N., Gümüş, M. and Nikoofal, M. E. (2021). Online-exclusive or hybrid? Channel merchandising strategies for ship-to-store implementation, *Management Science.* . ePub ahead of print November 17, `https://doi.org/10.1287/mnsc.2021.4180`.

Fang, J., Giuliano, G. and Wu, A. (2019). The spatial dynamics of Amazon Lockers in Los Angeles County, *Report MF-5.4c*, University of Southern California, Los Angeles, CA.

Gallino, S. and Moreno, A. (2014). Integration of online and offline channels in retail: The impact of sharing reliable inventory availability information, *Management Science* **60**(6): 1434–1451.

Gallino, S., Moreno, A. and Stamatopoulos, I. (2017). Channel integration, sales dispersion, and inventory management, *Management Science* **63**(9): 2813–2831.

Gao, F., Agrawal, V. V. and Cui, S. (2022). The effect of multichannel and omnichannel retailing on physical stores, *Management Science* **68**(2): 809–826.

Gao, F. and Su, X. (2017). Omnichannel retail operations with buy-online-and-pick-up-in-store, *Management Science* **63**(8): 2478–2492.

Glaeser, C. K., Fisher, M. and Su, X. (2019). Optimal retail location: Empirical methodology and application to practice, *Manufacturing & Service Operations Management* **21**(1): 86–102.

Gupta, S. (1988). Impact of sales promotions on when, what, and how much to buy, *Journal of Marketing Research* **25**(4): 342–355.

Heilman, C., Nakamoto, K. and Rao, A. (2002). Pleasant surprises: Consumer response to unexpected in-store coupons, *Journal of Marketing Research* **39**(2): 242–252.

Hu, M., Xu, X., Xue, W. and Yang, Y. (2022). Demand pooling in omnichannel operations, *Management science* **68**(2): 883–894.

Hwang, E. H., Nageswaran, L. and Cho, S.-H. (2021). Value of online–off-line return partnership to off-line retailers, *Manufacturing & Service Operations Management.* . ePub ahead of print November 16, `https://doi.org/10.1287/msom.2021.1026`.

Krishna, A. and Shoemaker, R. W. (1992). Estimating the effects of higher coupon face values on the timing of redemptions, mix of coupon redeemers, and purchase quantity, *Psychology & Marketing* **9**(6): 453–467.

Li, Z., Yang, W., Liu, X. and Si, Y. (2020). Coupon promotion and its two-stage price intervention on dual-channel supply chain, *Computers & Industrial Engineering* **145**: 106543.

Loch, C. and Wu, Y. (2008). Social preferences and supply chain performance: An experimental study, *Management Science* **54**(11): 1835–1849.

Martin-Herran, G. and Sigué, S. (2015). Trade deals and/or on-package coupons, *European Journal of Operational Research* **241**(2): 541–554.

Morganti, E., Dablanc, L. and Fortin, F. (2014). Final deliveries for online shopping: The deployment of pickup point networks in urban and suburban areas, *Research in Transportation Business & Management* **11**: 23–31.

Nagarajan, M. and Sošić, G. (2007). Stable farsighted coalitions in competitive markets, *Management Science* **53**(1): 29–45.

Nageswaran, L., Hwang, E. H. and Cho, S.-H. (2021). Offline returns for online retailers via partnership. Working paper, `https://dx.doi.org/10.2139/ssrn.3635026`.

Narasimhan, C. (1984). A price discrimination theory of coupons, *Marketing Science* **3**(2): 128–147.

Neslin, S., Henderson, C. and Quelch, J. (1985). Consumer promotions and the acceleration of product purchases, *Marketing Science* **4**(2): 147–165.

Pauwels, K., Leeflang, P., Teerling, M. and Huizingh, K. (2011). Does online information drive offline revenues? Only for specific products and consumer segments, *Journal of Retailing* **87**(1): 1–17.

Pavlov, V., Katok, E. and Zhang, W. (2022). Optimal contract under asymmetric information about fairness, *Manufacturing & Service Operations Management* **24**(1): 305–314.

Reibstein, D. and Traver, P. (1982). Factors affecting coupon redemption rates, *Journal of Marketing* **46**(4): 102–113.

Reimers, I. and Xie, C. (2019). Do coupons expand or cannibalize revenue? Evidence from an e-market, *Management Science* **65**(1): 286–300.

Salop, S. C. (1979). Monopolistic competition with outside goods, *The Bell Journal of Economics* pp. 141–156.

Sethuraman, R. and Mittelstaedt, J. (1992). Coupons and private labels: A cross-category analysis of grocery products, *Psychology & Marketing* **9**(6): 487–500.

Shulman, J., Coughlan, A. and Savaskan, R. (2009). Optimal restocking fees and information provision in an integrated demand-supply model of product returns, *Manufacturing & Service Operations Management* **11**(4): 577–594.

Song, P., Wang, Q., Liu, H. and Li, Q. (2020). The value of buy-online-and-pickup-in-store in omni-channel: Evidence from customer usage data, *Production and Operations Management* **29**(4): 995–1010.

Su, M., Zheng, X. and Sun, L. (2014). Coupon trading and its impacts on consumer purchase and firm profits, *Journal of Retailing* **90**(1): 40–61.

Tsay, A., Nahmias, S. and Agrawal, N. (1999). Modeling supply chain contracts: A review, *Quantitative Models for Supply Chain Management*, Springer, pp. 299–336.

UPS (2015). UPS pulse of the online shopper: Empowered shoppers propel retail change. Technical report. United Parcel Service of America, Atlanta, GA. `https://www.ups.com/assets/resources/media/knowledge-center/UPS_Pulse_of_the_Online_Shopper.pdf`.

Yuan, X., Dai, T., Chen, L. G. and Gavirneni, S. (2021). Co-opetition in service clusters with waiting-area entertainment, *Manufacturing & Service Operations Management* **23**(1): 106–122.

A2 Appendix for Article 2

A2.1 Tables and Figures

Table A2.1: Summary of Notation

Symbol	Definition
	Notation related to the customer
v	Customer's valuation for the product
h_o	Customer's hassle cost of using the direct-delivery option (e.g., shipping cost and delivery time)
x	Distance between customer's location and pickup location
h_p	Customer's hassle cost per unit of distance to visit the pickup location
θ	Probability that the customer redeems the coupon at the offline partner's store when picking up the order
r	Offline retailer's cross-selling profit per customer who picks up the order in-store
p	Price of the product
	Notation related to the online retailer
c_o	Online retailer's handling cost for each direct-delivery order (e.g., direct shipping cost)
c_p	Online retailer's handling cost for each in-store pickup order (e.g., cost of shipping to the pickup location)
α	Compensation value paid by the online retailer to the offline partner for each in-store pickup order
β	Monetary value of the coupon offered by the online retailer to be redeemed at the pickup location
d_o^i	Online retailer's expected demand for direct-delivery orders under policy $i \in \{F, C\}$
d_s^i	Online retailer's expected demand for in-store pickup orders under policy $i \in \{F, C\}$
π_o^i	Online retailer's expected profit under policy $i \in \{F, C\}$
	Notation related to the offline partner
c_s	Offline partner's handling cost per in-store pickup order (e.g., staff and storage)
π_s^i	Offline partner's expected profit from the partnership under policy $i \in \{F, C\}$

A2.2 Demand Functions

To avoid trivial cases, in all appendices, we assume that under the baseline and fixed fee policies, there are some customers who leave the market, that is, $v - p < 1$ and $(v - p)/h_p < 1/2$.

A2.3 Baseline Policy

Under the baseline policy, based on their utility, customers can either purchase the product via direct delivery or leave the market. Since the customer's utility from leaving the market is zero, a customer will purchase the product if her utility from buying is positive. Thus, since $h_o \sim U[0, 1]$ and $x \sim U[0, 1/2]$, the online retailer's demand under the baseline policy is simply $d_o^B = v - p$.

Fixed Fee Policy

Under the fixed fee policy, customer decisions are as follows:

a. When $h_o \leq \min\{v-p, xh_p\}$, the customer purchases via direct delivery.

b. When $x \leq \min\{\frac{h_o}{h_p}, \frac{v-p}{h_p}\}$, the customer purchases via in-store pickup.

c. When $v-p < \min\{h_o, xh_p\}$, the customer leaves the market.

Therefore, the demand for the direct-delivery and in-store pickup options are given by:

$$d_s^F = \frac{(v-p)(2-v+p)}{h_p} \quad \text{and} \quad d_o^F = \left[1 - \frac{(v-p)}{h_p}\right](v-p).$$

Coupon Policy

Under the coupon policy, customer decisions are as follows:

a. When $h_o \leq \min\{v-p, xh_p - \theta\beta\}$, the customer purchases via direct delivery.

b. When $x \leq \min\{\frac{h_o+\theta\beta}{h_p}, \frac{v-p+\theta\beta}{h_p}\}$, the customer purchases via in-store pickup.

c. When $v-p < \min\{h_o, xh_p - \theta\beta\}$, the customer leaves the market.

Therefore, the demand for the direct-delivery and in-store pickup options are given by:

$$d_s^C = \begin{cases} \frac{2\theta\beta+(2-v+p)(v-p)}{h_p}, & 0 \leq \beta \leq \frac{h_p-2(v-p)}{2\theta} \\ (1 - \frac{(h_p-2\theta\beta)^2}{4h_p}) & \frac{h_p-2(v-p)}{2\theta} \leq \beta \leq \frac{h_p}{2\theta} \\ 1 & \beta \geq \frac{h_p}{2\theta} \end{cases} \quad \text{and} \quad d_o^C = \begin{cases} (1 - \frac{(v-p+2\theta\beta)}{h_p})(v-p), & 0 \leq \beta \leq \frac{h_p-2(v-p)}{2\theta} \\ \frac{(h_p-2\theta\beta)^2}{4h_p}, & \frac{h_p-2(v-p)}{2\theta} \leq \beta \leq \frac{h_p}{2\theta} \\ 0, & \beta \geq \frac{h_p}{2\theta} \end{cases}$$

A2.4 Proofs of Statements

Proof of Proposition 2. The proof of Proposition 2 follows directly from the demand function derived in Appendix A2.2.

Proof of Proposition 3. (a) For the online retailer, we have

$$\Delta\pi_o^F = (p-c_o)\Delta d_o^B + (p-c_p-\alpha)\Delta d_s^F \implies \Delta\pi_o^F = \frac{2(v-p)(1-v+p)}{h_p}(p-c_p-\alpha) + \frac{(v-p)^2}{h_p}(c_o-c_p-\alpha),$$

and for the offline partner, we have

$$\Delta\pi_s^F = (r+\alpha-c_s)\Delta d_s^F = \frac{(v-p)(2-v+p)}{h_p}(r+\alpha-c_s).$$

(b) The offline partner will accept the pickup partnership offer if and only if $\Delta\pi_s^F \geq 0$. Since $\Delta d_s^F > 0$ and $\alpha \geq 0$, $\Delta\pi_s^F$ is positive if

$$r+\alpha-c_s \geq 0 \implies \alpha \geq \max\{0, c_s - r\} = \underline{\alpha}.$$

Similarly, the online retailer will initiate the partnership under the fixed fee policy if and only if $\Delta\pi_o^F \geq 0$, and thus

$$\Delta\pi_o^F = \frac{2(v-p)(1-v+p)}{h_p}(p-c_p-\alpha)+\frac{(v-p)^2}{h_p}(c_o-c_p-\alpha) \geq 0 \implies \alpha \leq p-c_p-\frac{v-p}{2-v+p}(p-c_o) = \bar{\alpha}.$$

Therefore, the partnership under the fixed fee policy is beneficial only when $\alpha \in [\underline{\alpha}, \bar{\alpha}]$.

Proof of Proposition 4. The proof of Proposition 4 directly follows from the demand functions derived in Appendix A2.2. We note that $\hat{\beta}$ and $\hat{\hat{\beta}}$ are given by:

$$\hat{\beta} = \begin{cases} 2\theta\beta & 0 \leq \beta \leq \frac{h_p-2(v-p)}{2\theta} \\ \frac{4\theta\beta(hp+\theta\beta)-h_p^2}{v-p} + h_p - v + p & \frac{h_p-2(v-p)}{2\theta} \leq \beta \leq \frac{h_p}{2\theta} \\ (h_p - v + p) & \beta \geq \frac{h_p}{2\theta} \end{cases} \quad \text{and} \quad \hat{\hat{\beta}} = \begin{cases} 2\theta\beta & 0 \leq \beta \leq \frac{h_p-2(v-p)}{2\theta} \\ \frac{(h_p-2v+2p)}{2} & \beta \geq \frac{h_p-2(v-p)}{2\theta} \end{cases}$$

Proof of Proposition 5. (a) This result can be shown by substituting the demand function into the profit function reported in Section 3.3.

(b) The existence proof of $\underline{\beta}$ follows a similar argument to that of $\underline{\alpha}$. Thus, we only show the existence of $\bar{\beta}$, and we can then find β^*. To show the existence of $\bar{\beta}$, it is enough to show that $\pi_o^C(\beta)$ is a continuous and unimodal function of β (i.e., there exists a $\check{\beta}$ such that $\pi_o^C(\beta)$ is increasing for $\beta < \check{\beta}$ and decreasing for $\beta \geq \check{\beta}$) and that $\pi_o^B \leq \pi_o^C(\beta = 0)$.

One can easily show that $\pi_o^B < \pi_o^C(0)$ and that $\pi_o^C(\beta)$ is a continuous function.

Thus, we only need to show that $\pi_o^C(\beta)$ is a unimodal function of β. The online retailer's profit under the coupon policy can be written as the following piece-wise function of β:

$$\pi_o^C(\beta) = \begin{cases} (1-\frac{2\theta\beta+v-p}{h_p})(v-p)(p-c_o)+\frac{2\theta\beta+(2-v+p)(v-p)}{h_p}(p-c_p-\theta\beta), & 0\leq\beta\leq\frac{h_p-2(v-p)}{2\theta} \\ \frac{(h_p-2\theta\beta)^2}{4h_p}(p-c_o)+(1-\frac{(h_p-2\beta)^2}{4h_p})(p-c_p-\theta\beta), & \frac{h_p-2(v-p)}{2\theta}\leq\beta\leq\frac{h_p}{2\theta} \\ (p-c_p-\theta\beta) & \beta\geq\frac{h_p}{2\theta} \end{cases} \tag{A2.1}$$

The first term in $\pi_o^C(\beta)$ is a quadratic function of β, which is denoted by $A(\beta)$. We thus have

$$\frac{\partial A(\beta)}{\partial\beta}=\frac{2\theta}{h_p}\Big[p-c_p-(v-p)(p-c_o)-\frac{(2-v+p)(v-p)}{2}-2\theta\beta\Big].$$

The root of $\frac{\partial A(\beta)}{\partial\beta}$ is given by:

$$\beta_A^*=\frac{p-c_p}{2\theta}-\frac{(v-p)(p-c_o)}{2\theta}-\frac{(2-v+p)(v-p)}{4\theta}.$$

This root maximizes $A(\beta)$ if $0\leq\beta_A^*\leq\frac{h_p-2(v-p)}{2\theta}$. In other words, β_A^* maximizes $A(\beta)$ if

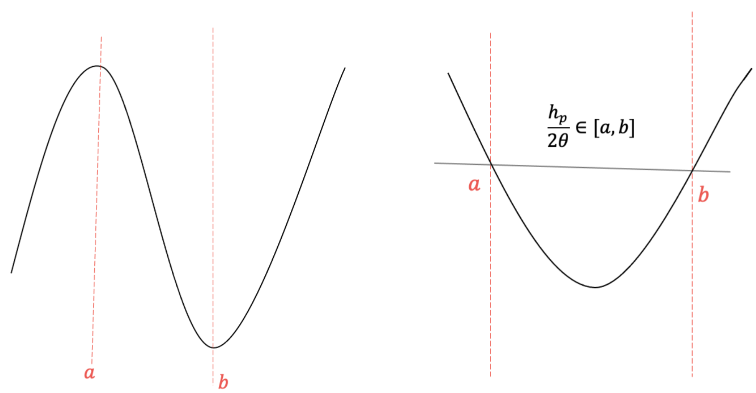
$$(v-p)c_o+(1-v+p)p+\frac{(2+v-p)(v-p))}{2}-h_p\leq c_p\leq(v-p)c_o+(1-v+p)p-\frac{(v-p)(2-v+p)}{2}.$$

If $c_p>(v-p)c_o+(1-v+p)p-\frac{(v-p)(2-v+p)}{2}$, then $\beta_A^*<0$, meaning that $A(\beta)$ is decreasing in β for $0\leq\beta\leq\frac{h_p-2(v-p)}{2}$. If $c_p<(v-p)c_o+(1-v+p)p+\frac{(2+v-p)(v-p))}{2}-h_p$, then $\beta_A^*>\frac{h_p-2(v-p)}{2\theta}$, meaning that $A(\beta)$ is increasing in β for $0\leq\beta\leq\frac{h_p-2(v-p)}{2\theta}$.

The second term in $\pi_o^C(\beta)$ is a cubic function of β, which is denoted by $B(\beta)$ and can be written as follows:

$$B(\beta) = (p - c_p - \theta\beta) + \frac{h_p^2 + 4(\theta\beta)^2 - 4h_p\theta\beta}{4h_p}(c_p + \theta\beta - c_o)$$
$$= \frac{1}{h_p}(\theta\beta)^3 + \frac{c_p - c_o - h_p}{h_p}(\theta\beta)^2 + (c_o - c_p - 1 + \frac{h_p}{4})\theta\beta + (\frac{h_p}{4}(c_p - c_o) + p - c_p).$$

Since $\frac{\theta^3}{h_p} > 0$, Figure A2.1(a) depicts the only possible pattern for $B(\beta)$.



(a) Possible Pattern for $B(\beta)$ (b) Plot of $\frac{\partial B(\beta)}{\partial \beta}$

Figure A2.1

The first derivative of $B(\beta)$ is given by:

$$\frac{\partial B(\beta)}{\partial \beta} = -\theta + \frac{h_p - 2\theta\beta}{h_p}\theta(c_o - c_p) - \frac{h_p - 2\theta\beta}{h_p}\theta^2\beta + \frac{(h_p - 2\theta\beta)^2}{4h_p}\theta, \quad \frac{h_p - 2(v - p)}{2\theta} < \beta \leq \frac{h_p}{2\theta}.$$

When $\beta = \frac{h_p}{2\theta}$, we have $\frac{\partial B(\beta)}{\partial \beta} = -\theta < 0$. Therefore, $\frac{h_p}{2\theta} \in [a, b]$, where a and b are the roots of $\frac{\partial B(\beta)}{\partial \beta}$ (see Figure A2.1).

When $\beta = \frac{h_p - 2(v-p)}{2\theta}$, we have

$$\begin{aligned}\frac{\partial B(\beta)}{\partial \beta} &= \frac{2(v-p)}{h_p}\theta(c_o - c_p) - \frac{2(v-p)}{h_p}(\frac{h_p}{2} - v + p)\theta + \frac{(v-p)^2}{h_p}\theta - \theta \\ &= \Big[\frac{2(v-p)}{h_p}(c_o - c_p) - (v-p) + \frac{3(v-p)^2}{h_p} - 1\Big]\theta.\end{aligned}$$

If $c_p \leq c_o + \frac{3(v-p)}{2} - \frac{h_p}{2} - \frac{h_p}{2(v-p)}$, then $\frac{\partial B(\beta)}{\partial \beta} \geq 0$ for $\beta = \frac{h_p-2(v-p)}{2\theta}$. It means that $\frac{h_p-2(v-p)}{2\theta} < a$, where a is the first root of $\frac{\partial B(\beta)}{\partial \beta}$ (see Figure A2.1). Therefore, when $c_p \leq c_o + \frac{3(v-p)}{2} - \frac{h_p}{2} - \frac{h_p}{2(v-p)}$, a maximizes $B(\beta)$. Since $\frac{\partial B(\beta)}{\partial \beta}$ is a quadratic function of β, a is given by:

$$a = \frac{1}{6\theta}\Big[2(h_p + c_o - c_p) - \sqrt{(2c_p + h_p)^2 + 4c_o(c_o - h_p - 2c_p) + 12h_p}\Big].$$

When $c_p > c_o + \frac{3(v-p)}{2} - \frac{h_p}{2} - \frac{h_p}{2(v-p)}$, $\frac{\partial B(\beta)}{\partial \beta} < 0$ for $\beta = \frac{h_p-2(v-p)}{2\theta}$. Therefore, $\frac{h_p-2(v-p)}{2\theta} \in [a, b]$, where a and b are the roots of $\frac{\partial B(\beta)}{\partial \beta}$. Thus, $B(\beta)$ is decreasing in β when $\beta \in [\frac{h_p-2(v-p)}{2\theta}, \frac{h_p}{2\theta}]$ and $\beta = \frac{h_p-2(v-p)}{2\theta}$ maximizes $B(\beta)$.

The third term of $\pi_o^C(\beta)$ is a linear function of β, which is denoted by $C(\beta)$. Since $\frac{\partial C(\beta)}{\partial \beta} = -\theta < 0$, $C(\beta)$ is decreasing in β.

We define $c_1 = c_o - \frac{h_p}{2} - \frac{h_p}{2(v-p)} + \frac{3(v-p)}{2}$, $c_2 = (v-p)c_o + (1-v+p)p + \frac{(2+v-p)(v-p)}{2} - h_p$, and $c_3 = (v-p)c_o + (1-v+p)p - \frac{(v-p)(2-v+p)}{2}$. Since we assume that $v - p < 1$ and $\frac{v-p}{h_p} \leq \frac{1}{2}$ (i.e., there are some customers who leave the market under the baseline and fixed fee policies), we can show that $c_3 \geq c_2 \geq c_1$. Since $\pi_o^C(\beta)$ is continuous, $\pi_o^C(\beta)$ can have only one of four possible patterns as shown in Figure A2.2. We thus conclude that $\pi_o^C(\beta)$ is a continuous and unimodal function.

Since $\pi_o^C(\beta)$ is a continuous and unimodal function, there exists a unique $\check{\beta}$ (see closed-form expression below) that maximizes $\pi_o^C(\beta)$.

$$\check{\beta} = \begin{cases} \frac{1}{6\theta}\Big[2(h_p + c_o - c_p) - \sqrt{(2c_p + h_p)^2 + 4c_o(c_o - h_p - 2c_p) + 12h_p}\Big], & c_p \leq c_1 \\ \frac{1}{2\theta}\Big[h_p - 2(v-p)\Big], & c_1 < c_p \leq c_2 \\ \frac{1}{4\theta}\Big[2(p - c_p) - 2(v-p)(p-c_o) - 2(v-p) + (v-p)^2\Big], & c_2 < c_p \leq c_3 \\ 0, & c_p > c_3 \end{cases}$$

where $c_1 = c_o - \frac{h_p}{2} - \frac{h_p}{2(v-p)} + \frac{3(v-p)}{2}$, $c_2 = (v-p)c_o + (1-v+p)p + \frac{(2+v-p)(v-p)}{2} - h_p$, and $c_3 = (v-p)c_o + (1-v+p)p - \frac{(v-p)(2-v+p)}{2}$.

The variable $\check{\beta}$ maximizes the online retailer's profit under the coupon policy without considering the offline partner's rationality constraint. If $\beta = \check{\beta}$ does not satisfy the rationality constraint, then the online retailer should increase the value of β. Since $\pi_o^C(\beta)$ is decreasing for $\beta > \check{\beta}$, in that case, the optimal value is the smallest value that satisfies the offline partner's rationality constraint, that is, $\frac{c_s - r}{\theta}$. As a result, the optimal value of β is $\min\{\frac{c_s-r}{\theta}, \check{\beta}\}$.

Proof of Proposition 6. First, we show that there exists a unique $\bar{\bar{\beta}}$ such that under condition $\theta\beta = \alpha$, the online retailer will prefer the coupon policy over the fixed fee policy when $\beta \leq \bar{\bar{\beta}}$ (i.e., for $\alpha = \theta\beta \leq \theta\bar{\bar{\beta}}$, $\pi_o^F(\alpha) \leq \pi_o^C(\beta)$). We also show that $\theta\bar{\bar{\beta}} \leq \bar{\alpha}$ and $\bar{\bar{\beta}} < \bar{\beta}$. By combining these findings with Propositions 3 and 5, it is straightforward to show that the optimal policy is (i) the coupon policy when $c_s \leq r + \theta\bar{\bar{\beta}} = \underline{c}_s$, (ii) the fixed fee policy when $r + \theta\bar{\bar{\beta}} = \underline{c}_s < c_s \leq r + \bar{\alpha} = \bar{c}_s$, and (iii) the baseline policy when $c_s > \bar{c}_s$.

The existence of $\bar{\bar{\beta}}$ can be shown using the characteristics of $\pi_o^C(\beta)$ as discussed in the proof of Proposition 5. For $\frac{\alpha}{\theta} = \beta > \check{\beta}$, the online retailer's profit is decreasing in the partnership parameter under both the fixed fee and the coupon policies, and $|\frac{\partial \pi_o^C(\beta)}{\partial \beta}| > \theta |\frac{\partial \pi_o^F(\alpha)}{\partial \alpha}|$ for $\beta > \frac{h_p - 2(v-p)}{2\theta}$. Therefore, $\pi_o^C(\beta) = \pi_o^F(\alpha)$ has a unique solution.

Since $\pi_o^F(\alpha = 0) = \pi_o^C(\beta = 0)$, $\pi_o^C(\beta)$ is a unimodal function, and $\pi_o^F(\alpha)$ is a decreasing function of α, if we show that $\theta\bar{\beta} \leq \bar{\alpha}$, then we can conclude that $\theta\bar{\bar{\beta}} < \bar{\alpha}$

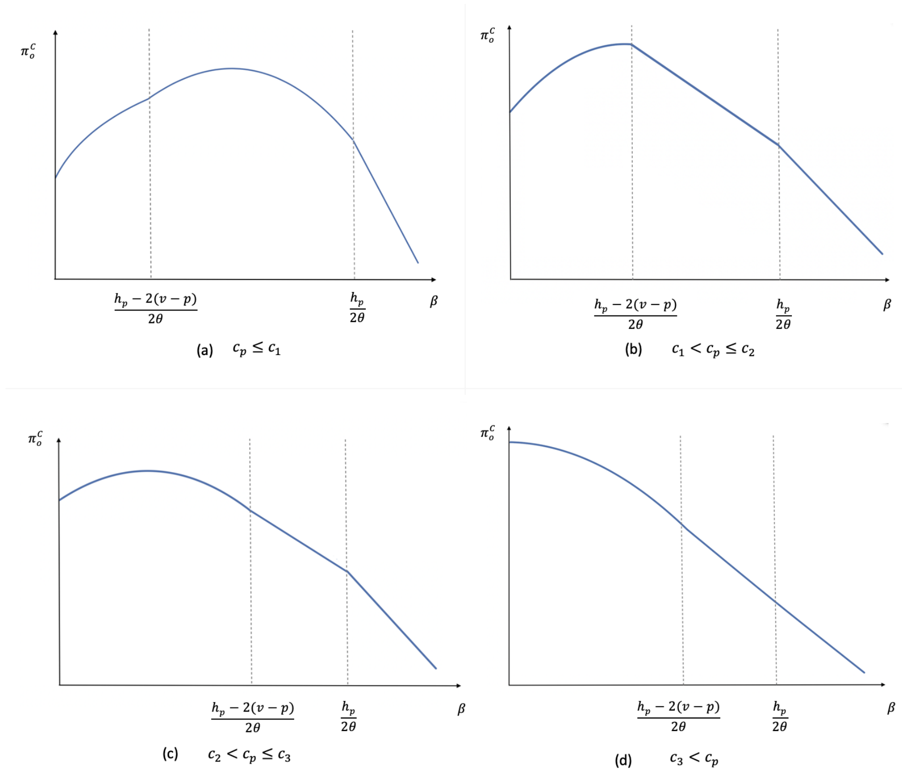


Figure A2.2: The Online Retailer's Profit under the Coupon Policy

(see Figure A2.3). To show this, it is enough to show that $\pi_o^C(\frac{\bar{\alpha}}{\theta}) < \pi_o^F(\bar{\alpha}) = \pi_o^B$ where $\bar{\alpha} = p - c_p - \frac{v-p}{2-v+p}(p - c_o)$. The following tree situations are then possible:

1. $\bar{\alpha} > \frac{h_p}{2}$. In this case, we have

$$\pi_o^C(\frac{\bar{\alpha}}{\theta}) = (p - c_p - \bar{\alpha}) = \frac{v-p}{2-v+p}(p - c_o) < (v-p)(p - c_o) = \pi_o^B,$$

where the last inequality follows from $v - p < 1$.

2. $\frac{h_p - 2(v-p)}{2} < \bar{\alpha} \le \frac{h_p}{2}$. In this case, we have

$$\pi_o^C(\frac{\bar{\alpha}}{\theta}) = \Big[p - c_p - \bar{\alpha} + (c_p + \bar{\alpha} - c_o)\frac{(h_p - 2\bar{\alpha})^2}{4h_p}\Big] = \Big[1 - \frac{(h_p - 2\bar{\alpha})^2}{4h_p}\Big]\frac{v-p}{2-v+p}(p - c_o) + \frac{(h_p - 2\bar{\alpha})^2}{4h_p}(p - c_o).$$

We next show that if $\pi_o^C(\frac{\bar{\alpha}}{\theta}) \le \pi_o^B$, we then reach to a strict inequality. We

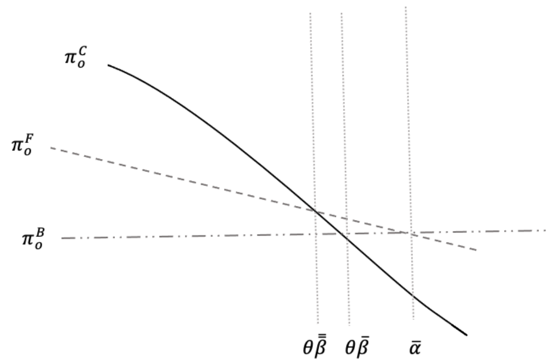


Figure A2.3: Possible Relationship Between $\bar{\bar{\beta}}$, $\bar{\beta}$, and $\bar{\alpha}$

have

$$\Big[1-\frac{(h_p-2\bar{\alpha})^2}{4h_p}\Big](\frac{v-p}{2-v+p})(p-c_o)+\frac{(h_p-2\bar{\alpha})^2}{4h_p}(p-c_o) \leq (v-p)(p-c_o) \implies (h_p-2\bar{\alpha})^2 \leq 2h_p(v-p)$$

$$\implies (h_p-2(p-c_p)+\frac{2(v-p)}{2-v+p}(p-c_o))^2 \leq 2h_p(v-p).$$

Since $4(v-p)^2 \leq 2h_p(v-p)$, then

$$\Big[h_p-2(p-c_p)+\frac{2(v-p)}{2-v+p}(p-c_o)\Big] < 2(v-p) \implies \frac{h_p-2(v-p)}{2} \leq (p-c_p)-\frac{v-p}{2-v+p}(p-c_o) = \bar{\alpha}.$$

3. $\bar{\alpha} < \frac{h_p-2(v-p)}{2}$. As in the second case, if $\pi_o^C(\frac{\bar{\alpha}}{\theta}) \leq \pi_o^B$, then we reach a strict inequality.

$$\pi_o^C(\frac{\bar{\alpha}}{\theta}) = \frac{2\bar{\alpha}+(2-v+p)(v-p)}{h_p}(p-c_p-\bar{\alpha})+(1-\frac{v-p+2\bar{\alpha}}{h_p})(v-p)(p-c_o) \leq (p-c_o)(v-p)$$

$$\implies 2\bar{\alpha}(p-c_p-\bar{\alpha})+(2-v+p)(v-p)(p-c_p-\bar{\alpha}) < (v-p+2\bar{\alpha})(v-p)(p-c_o).$$

Since $\bar{\alpha} = (p-c_p)-\frac{v-p}{2-v+p}(p-c_o)$, we have

$$2\bar{\alpha}\frac{v-p}{2-v+p}(p-c_o)+(2-v+p)(v-p)\frac{v-p}{2-v+p}(p-c_o) \leq (v-p+2\bar{\alpha})(v-p)(p-c_o)$$

$$\implies \frac{2\bar{\alpha}}{2-v+p} + v - p \leq v - p + 2\bar{\alpha} \implies 1 \leq 2 - v + p,$$

where last inequity follows from $v - p < 1$. We then have $\theta\bar{\bar{\beta}} \leq \bar{\alpha}$.

Therefore, we conclude that when $\alpha = \theta\beta < \theta\bar{\bar{\beta}}$, we have $\pi_o^C(\beta) > \pi_o^F(\alpha) > \pi_o^B$, when $\theta\bar{\bar{\beta}} < \alpha = \theta\beta < \bar{\alpha}$, we have $\pi_o^F(\alpha) > \max\{\pi_o^C(\beta), \pi_o^B\}$, and when $\alpha = \theta\beta > \bar{\alpha}$, we have $\pi_o^B > \pi_o^F(\alpha) > \pi_o^C(\alpha)$. Thus, based on Propositions 3 and 5, when $c_s \leq r + \theta\bar{\bar{\beta}} = \underline{c}_s$, the online retailer's profit under a beneficial coupon policy is higher relative to the fixed fee policy, and the coupon policy is beneficial for any $\beta \in [\max\{0, \frac{c_s - r}{\theta}\}, \bar{\bar{\beta}}]$. When $\underline{c}_s < c_s \leq r + \bar{\alpha} = \bar{c}_s$, the online retailer's profit under a beneficial fixed fee policy is greater than a beneficial coupon policy, and the fixed policy is beneficial for any $\alpha \in [c_s - r, \bar{c}_s - r]$. Lastly, when $c_s > \bar{c}_s$, neither a fixed fee policy nor a coupon policy can improve the online retailer's profit when compared to the baseline policy, and hence the baseline policy is optimal.

Proof of Proposition 7. We first show that when $c_p > v - \frac{h_p}{2} - \frac{(v-p)(p-c_o)}{h_p} = \bar{c}_p$, the hybrid policy cannot be optimal, and so the optimal policy remains the same as in Proposition 6.

When $c_p > \bar{c}_p$, we have $\bar{\bar{\beta}} \leq \frac{h_p - 2(v-p)}{2\theta}$, so that $\pi_o^C(\beta = \frac{h_p - 2(v-p)}{2\theta}) < \pi_o^F(\alpha = \frac{h_p - 2(v-p)}{2})$. We next show that when $c_p > \bar{c}_p$, the optimal policy is either the coupon policy or the fixed fee policy.

Although the online retailer's profit is a piece-wise function of β, in this case, it is enough to focus on the first part of the function (i.e., when $\beta \leq \frac{h_p - 2(v-p)}{2\theta}$) because the coupon policy can be optimal only when $\beta \leq \frac{h_p - 2(v-p)}{2\theta}$. We assume that there exists an optimal hybrid policy with parameters β_h^* and α_h^* (where $\beta_h^*, \alpha_h^* > 0$). Namely, β_h^* and α_h^* maximize the online retailer's profit, while satisfying the offline partner's rationality constraint $(r + \alpha_h^* + \theta\beta_h^* - c_s)d_s^H \geq 0$. Thus, the online retailer's profit under the hybrid policy must be higher than the profit under the coupon policy with the coupon value $\beta^* = \frac{\theta\beta_h^* + \alpha_h^*}{\theta}$ and the fixed fee policy with $\alpha^* = \theta\beta_h^* + \alpha_h^*$.

We let $\pi_o^H(\gamma = \theta\beta_h + \alpha_h)$ denote the online retailer's profit under the hybrid policy. For the online retailer's profit under the hybrid policy to be higher relative to the coupon policy, we must have the following (without loss of generality, we assume $\theta = 1$):

$$\begin{aligned}\pi_o^H(\gamma^* = \beta_h^* + \alpha_h^*) > \pi_o^C(\beta^*) \implies & \frac{2\beta_h^* + (2-v+p)(v-p)}{h_p}(p - c_p - \beta_h^* - \alpha_h^*) + (1 - \frac{v-p+2\beta_h^*}{h_p})(v-p)(p-c_o) > \\ & \frac{2\beta^* + (2-v+p)(v-p)}{h_p}(p - c_p - \beta^*) + (1 - \frac{v-p+2\beta^*}{h_p})(v-p)(p-c_o) \\ \implies & \beta_h^* + \alpha_h^* > p - c_p - (v-p)(p-c_o). \quad \text{(A2.2)}\end{aligned}$$

In addition, the online retailer's profit under the hybrid policy must be higher relative to fixed fee policy, that is,

$$\begin{aligned}\pi_o^H(\gamma^* = \beta_h^* + \alpha_h^*) > \pi_o^F(\alpha^*) \implies & \frac{2\beta_h^* + (2-v+p)(v-p)}{h_p}(p - c_p - \beta_h^* - \alpha_h^*) + (1 - \frac{v-p+2\beta_h^*}{h_p})D(v-p)(p-c_o) > \\ & \frac{(2-v+p)(v-p)}{h_p}(p - c_p - \alpha^*) + (1 - \frac{v-p}{h_p})(v-p)(p-c_o) \\ \implies & \beta_h * + \alpha_h^* < p - c_p - (v-p)(p-c_o). \quad \text{(A2.3)}\end{aligned}$$

Inequalities (A2.2) and (A2.3) cannot hold simultaneously, and so when $c_p > \bar{c}_p$, the hybrid policy cannot be optimal. The online retailer will opt either for the fixed fee policy or for coupon policy based on Proposition 6.
We next show that when $c_p < \bar{c}_p$ (i.e., $\bar{\bar{\beta}} > \frac{h_p - 2(v-p)}{2\theta}$), the hybrid policy can be optimal only when $\underline{\underline{c}}_s = r + \frac{h_p-2(v-p)}{2} \leq c_s < r + p - c_p - (p-c_o)(v-p) = \bar{\bar{c}}_s$. When $c_p < \bar{c}_p$, there are three possible cases:

1. $c_s \leq \theta\check{\beta} + r$. Based on Propositions 5 and 6, the online retailer's profit is maximized when $\beta = \check{\beta}$, and since $c_s \leq \theta\check{\beta} + r$, $\beta = \check{\beta}$ satisfies the offline partner's rationality constraint, so that the coupon policy is optimal.
2. $\theta\check{\beta} + r < c_s \leq r + \bar{\alpha}$. In this case, the online retailer can maximize her profit under the fixed fee and coupon policies with parameters $c_s - r$ and $\frac{c_s - r}{\theta}$, respectively (i.e., to maximize her profit, the online retailer will pay the minimum compensation value under either policy, and so the offline partner's rationality constraint is binding). Therefore, the online retailer determines

the optimal decision by using the following optimization formulation:

$$\begin{aligned}\max_{\beta_h,\alpha_h} \quad & \pi_o(\gamma) = (p-c_o)d_o^H + (p-c_p-\alpha_h-\theta\beta_h)d_s^H \\ & (r+\alpha_h+\theta\beta_h-c_s)d_s^H \geq 0 \\ & \alpha_h+\theta\beta_h = c_s - r \\ & \alpha_h, \beta_h \geq 0.\end{aligned} \tag{A2.4}$$

As a result, it is enough to solve

$$\max_{\beta_h} \pi_o(\beta_h) = (p-c_o)d_o^H + (p-c_p-c_s+r)d_s^H. \tag{A2.5}$$

If $\beta_h^\dagger$ is the solution of Equation (A2.5), then the solution of Equation (A2.4) is $(\beta_h^*, \alpha_h^*) = (\beta_h^\dagger, c_s - r - \beta_h^\dagger)$. We will find $\beta_h^\dagger$ based on the first-order condition. The profit function $\pi_o(\beta_h)$ in Equation (A2.5) can be written as

$$\pi_o(\beta_h) = \begin{cases} (p-c_o)(1-\frac{v-p+2\theta\beta_h}{h_p})(v-p) + (p-c_p-c_s+r)\frac{2\theta\beta_h+(2-v+p)(v-p)}{h_p}, & 0 \leq \beta_h \leq \frac{h_p-2(v-p)}{2\theta} \\ (p-c_p-c_s+r) - \frac{(h_p-2\theta\beta_h)^2}{4h_p}(r+c_o-c_p-c_s), & \frac{h_p-2(v-p)}{2\theta} < \beta_h \leq \frac{h_p}{2\theta} \\ (p-c_p-c_s+r) & \beta_h \geq \frac{h_p}{2\theta} \end{cases}$$

Thus, the first derivative of $\pi_o(\beta_h)$ is given by:

$$\frac{\partial \pi_o}{\partial \beta_h} = \begin{cases} \frac{2\theta}{h_p}\Big[p-c_p-c_s+r-(p-c_o)(v-p)\Big], & 0 \leq \beta_h < \frac{h_p-2(v-p)}{2\theta} \\ \frac{\theta}{h_p}(h_p-2\theta\beta_h)(r+c_o-c_p-c_s), & \frac{h_p-2(v-p)}{2\theta} < \beta_h \leq \frac{h_p}{2\theta} \\ 0 & \beta_h \geq \frac{h_p}{2\theta} \end{cases}$$

Therefore, when $c_s > r+p-c_p-(p-c_o)(v-p)$, we have $\beta_h^\dagger = 0$. In this case, the optimal policy is the fixed fee policy. When $c_s < r+p-c_p-(p-c_o)(v-p)$, we have $\beta_h^\dagger = \frac{h_p-2(v-p)}{2\theta}$. In this case, the optimal policy can be either the coupon policy or the hybrid policy. More specifically, if $c_s \geq r + \frac{h_p-2(v-p)}{2}$, then the hybrid policy with $\beta_h^* = \frac{h_p-2(v-p)}{2\theta}$ and $\alpha_h^* = c_s - r - \frac{h_p-2(v-p)}{2\theta}$ is optimal. Otherwise, the coupon policy with $\beta^* = \frac{c_s - r}{\theta}$ is optimal.

3. $c_s > r + \bar{\alpha}$. In this case, based on Proposition 6, none of the policies are beneficial.

A2.5 Details for Comparative Statics

We let $\Delta\pi_o^{C-F}$ denote the difference in the online retailer's profit between the coupon policy and the fixed fee policy with the same average compensation value per pickup order (i.e., $\alpha = \theta\beta$). We first chracterize how $\Delta\pi_o^F$, $\Delta\pi_o^C$, and $\Delta\pi_o^{C-F}$ are changing with respect to c_p and c_o. We have

$$\frac{\partial \Delta\pi_o^F}{\partial c_p} = -\frac{(v-p)(2-v+p)}{h_p} < 0 \quad \text{and} \quad \frac{\partial \Delta\pi_o^C}{\partial c_p} = -\Big[\frac{(v-p)(2-v+p)}{h_p} + \frac{(1-v+p)}{h_p}\hat{\hat{\beta}} + \frac{(v-p)}{h_p}\hat{\beta}\Big] < 0.$$

Thus, we conclude that the profitability of the fixed fee and coupon policies decreases with c_p. We also have

$$\frac{\partial \Delta\pi_o^{C-F}}{\partial c_p} = -\Big[\frac{(1-v+p)}{h_p} + \frac{(v-p)}{h_p}\hat{\beta}\Big] < 0.$$

Since $\frac{\partial \Delta\pi_o^{C-F}}{\partial c_p} < 0$, we conclude that the profitability of the coupon policy decreases with c_p faster than that of the fixed fee policy.

Similarly, for c_o, we have

$$\frac{\partial \Delta\pi_o^F}{\partial c_o} = \frac{(v-p)^2}{h_p} > 0 \quad \text{and} \quad \frac{\partial \Delta\pi_o^C}{\partial c_o} = \frac{(v-p)^2}{h_p} + \frac{(v-p)}{h_p}\hat{\beta} > 0.$$

Thus, we conclude that the profitability of the fixed fee and coupon policies increases with c_o. We also have

$$\frac{\partial \Delta\pi_o^{C-F}}{\partial c_o} = \frac{(v-p)}{h_p}\hat{\beta} > 0.$$

Since $\frac{\partial \Delta\pi_o^{C-F}}{\partial c_o} > 0$, we conclude that the profitability of the coupon policy increases with c_o faster than the fixed fee policy. Next, we show how $\Delta\pi_o^{C-F}$ changes with respect to p.

$$\Delta\pi_o^{C-F}=\begin{cases}\frac{2\theta\beta(1-v+p)}{h_p}(p-c_p-\theta\beta+\frac{2\theta\beta(v-p)}{h_p}(c_o-c_p-\theta\beta)) & p\geq\theta\beta+v-\frac{h_p}{2}\\ \frac{(h_p-2v+2p)(1-v+p)}{2h_p}(p-c_p-\theta\beta)+(\theta\beta-\frac{(\theta\beta)^2}{h_p}-\frac{(h_p-2(v-p))^2}{4h_p})(c_o-c_p-\theta\beta) & p\leq\theta\beta+v-\frac{h_p}{2},\beta\leq\frac{h_p}{2\theta}\\ \frac{(h_p-2(v-p))(1-v+p)}{2h_p}(p-c_p-\theta\beta)+(1-\frac{v-p}{h_p})(v-p)(c_o-c_p-\theta\beta) & p\leq\theta\beta+v-\frac{h_p}{2},\beta>\frac{h_p}{2\theta}\end{cases}$$

More specifically, $\frac{\partial\Delta\pi_o^{C-F}}{\partial p}$ can be characterized as follows:

- When $p\geq\theta\beta+v-\frac{h_p}{2}$, we have

$$\frac{\partial\Delta\pi_o^{C-F}}{\partial p}=\frac{2\theta\beta}{h_p}(1-v+2p-c_o)\geq 0.$$

- When $p\leq\theta\beta+v-\frac{h_p}{2}$, $\theta\beta\leq\frac{h_p}{2}$, we have

$$\frac{\partial\Delta\pi_o^{C-F}}{\partial p}=\frac{1}{2h_p}\Big[(p-c_p-\theta\beta)(2(1-v+p)+(h_p-2(v-p))+(h_p-2(v-p))(1-v+p-2(c_o-c_p-\theta\beta))\Big].$$

  When $c_o\leq c_p+\theta\beta$, we can show that $\frac{\partial\Delta\pi_o^{C-F}}{\partial p}>0$. When $c_o>c_p+\theta\beta$, it is enough to show that

$$2(p-c_p-\theta\beta)(1-v+p)+(p-c_p-\theta\beta)(h_p-2(v-p))+(h_p-2(v-p))(1-v+p)\geq 2(c_o-c_p-\theta\beta)(h_p-2(v-p))$$

  so that

$$\frac{2(p-c_p-\theta\beta)(1-v+p)}{(h_p-2(v-p))}+(p-c_p-\theta\beta)+(1-v+p)\geq 2(c_o-c_p-\theta\beta),$$

  where the above inequality follows from $p+c_p+\theta\beta+1-v+p\geq 2c_o$.

- When $p\leq\theta\beta+v-\frac{h_p}{2}$, $\theta\beta>\frac{h_p}{2}$, we have

$$\frac{\partial\Delta\pi_o^{C-F}}{\partial p}=\frac{1}{h_p}((1-v+p)(p-c_p-\theta\beta)+\frac{(h_p-2(v-p))(p-c_p-\theta\beta)}{2}+\frac{(h_p-2(v-p))(1-v+p)}{2}+(c_o-c_p-\theta\beta)(2(v-p)-h_p)).$$

Since $p + c_p + \theta\beta + 1 - v + p \geq c_o$, we can show that $\frac{\partial \Delta\pi_o^{C-F}}{\partial p} > 0$ when $p \leq \theta\beta + v - \frac{h_p}{2}$ and $\theta\beta > \frac{h_p}{2}$.

As a result, by increasing p, the coupon policy becomes more profitable for the online retailer.

A2.6 Proofs of Statements from Section 3.5

Budget Constraint

In this subsection, we provide the details of our findings related to the model in the presence of the budget constraint. We let K denote the total budget. There exists $\underline{K} = \max\{0, c_s - r\}d_s^F$ so that when $K < \underline{K}$, neither the fixed fee policy nor the coupon policy are feasible. This follows from the fact that the minimum compensation value under which the offline partner is not worse off under the fixed fee policy is $\max\{0, c_s - r\}$, and the demand for in-store pickup orders is d_s^F. As a result, the minimum budget under which the fixed fee policy is beneficial is $\max\{0, c_s - r\}d_s^F$. Note that the minimum budget under the coupon policy is higher than $\max\{0, c_s - r\}d_s^F$. Indeed, based on Proposition 4, the demand for in-store pickup orders is always higher under the coupon policy than under the fixed fee policy (i.e., $d_s^F < d_s^C$). Thus, when $K > \underline{K}$, the fixed fee policy is beneficial. The maximum fixed fee value that the online retailer can pay to the offline partner under the budget constraint is $\frac{K}{d_s^F}$. However, based on Proposition 6, the fixed fee policy will be beneficial only when $\alpha \in [\underline{\alpha}, \bar{c}_s - r]$. Therefore, when $K > \underline{K}$ the fixed fee policy with parameter $\alpha \in [\underline{\alpha}, \max\{\bar{c}_s - r, \frac{K}{d_s^F}\}]$ is beneficial.

We next identify conditions under which the coupon policy is beneficial (for the model with a budget constraint). Recall that the offline partner is not worse off under the coupon policy only when $\beta > \underline{\beta} = \max\{0, \frac{c_s - r}{\theta}\}$. Therefore, the minimum budget under which the coupon policy is beneficial is $\bar{K} = \max\{0, \frac{c_s - r}{\theta}\}d_s^C$, which is characterized as

$$\bar{K} = \begin{cases} 0, & c_s - r \leq 0 \\ \frac{2(c_s-r)+(2-v+p)(v-p)}{h_p}(c_s - r), & 0 < c_s - r \leq \frac{h_p-2(v-p)}{2} \\ (1 - \frac{(h_p-2(c_s-r))^2}{4h_p})(c_s - r) & \frac{h_p-2(v-p)}{2} \leq c_s - r \leq \frac{h_p}{2} \\ c_s - r & c_s - r \geq \frac{h_p}{2} \end{cases}$$

As a result, the coupon policy is beneficial only when $K \geq \bar{K}$. In this case, the maximum coupon value that the online retailer can pay is $\frac{K}{\theta d_s^C}$. Based on Proposition 6, the coupon policy is beneficial when $\beta \in [\underline{\beta}, \frac{c_s-r}{\theta}]$. Therefore, when $K > \bar{K}$, the coupon policy is beneficial for any $\beta \in [\underline{\beta}, \min\{\frac{c_s-r}{\theta}, \frac{K}{\theta d_s^C}\}]$. Since $\bar{K} > \underline{K}$, when $K > \bar{K}$, both the fixed fee and the coupon policies are beneficial and they become optimal strategies based on the results of Proposition 6.

Multiple Pickup Locations

First, we examine how the optimal policy changes when there is more than one pickup location. To do so, we evaluate how the demand function varies when there are n pickup locations under the fixed fee and coupon policies. We will then infer the optimal policy.

The demand for the direct-delivery (d_o^F) and in-store pickup (d_s^F) options, when there are n pickup locations under the fixed fee policy are given by (recall that when n pickup locations are available, the maximum distance between a customer and a pickup location is $\frac{1}{2n}$ based on the circle model from Salop (1979)):

$$d_o^F = \begin{cases} (1 - \frac{n(v-p)}{h_p})(v - p), & n < \lceil \frac{h_p}{2(v-p)} \rceil \\ \frac{h_p}{4n}, & n \geq \lceil \frac{h_p}{2(v-p)} \rceil \end{cases} \quad \text{and} \quad d_s^F = \begin{cases} \frac{n(2-v+p)(v-p)}{h_p}, & n < \lceil \frac{h_p}{2(v-p)} \rceil \\ (1 - \frac{h_p}{4n}), & n \geq \lceil \frac{h_p}{2(v-p)} \rceil \end{cases}$$

The demand under the baseline policy is independent of n and remains to be $(v - p)$. We can now substitute the demand function into the online retailer's and offline partner's profit functions under the fixed fee and baseline policies, and then, by comparing them, we can find the conditions under which the fixed fee policy is beneficial.

As in the main model, the offline partner is better-off under the fixed fee policy if and only if $\alpha \geq \underline{\alpha} = \max\{c_s - r, 0\}$ (using the same argument as in the main model). The online retailer is better-off under the fixed fee policy if and only if $\alpha \leq \bar{\alpha}$, where $\bar{\alpha}$ is given by:

$$\bar{\alpha} = \begin{cases} (p - c_p) - \frac{v-p}{2-v+p}(p - c_o), & 1 \leq n < \lceil \frac{h_p}{2(v-p)} \rceil \\ (p - c_p) - \frac{4n(v-p)-h_p}{4n-h_p}(p - c_o), & n \geq \lceil \frac{h_p}{2(v-p)} \rceil \end{cases}$$

Thus, similar to Proposition 3, the fixed fee policy is beneficial $\forall \alpha \in [\underline{\alpha}, \bar{\alpha}]$.

Under the coupon policy, the demand functions for the direct-delivery and in-store pickup options are given by:

$$d_o^C = \begin{cases} \left[1 - \frac{n(v-p+2\theta\beta)}{h_p}\right](v - p), & 0 \leq \beta < \max\{0, \frac{h_p-2n(v-p)}{2n\theta}\} \\ \frac{(h_p-2n\theta\beta)^2}{4nh_p}, & \max\{0, \frac{h_p-2n(v-p)}{2n\theta}\} \leq \beta < \frac{h_p}{2n\theta} \\ 0 & \beta \geq \frac{h_p}{2n\theta} \end{cases}$$

$$d_s^C = \begin{cases} \frac{n}{h_p}(2\theta\beta + (2 - v + p)(v - p)), & 0 \leq \beta < \max\{0, \frac{h_p-2n(v-p)}{2n\theta}\} \\ \left[1 - \frac{(h_p-2n\theta\beta)^2}{4nh_p}\right], & \max\{0, \frac{h_p-2n(v-p)}{2n\theta}\} \leq \beta < \frac{h_p}{2n\theta} \\ 1 & \beta \geq \frac{h_p}{2n\theta} \end{cases}$$

Thus, the online retailer's profit under the coupon policy with n pickup locations is given by:

$$\pi_o^C(\beta) = \begin{cases} \Big[1 - \frac{n(v-p+2\theta\beta)}{h_p}\Big](v-p)(p-c_o) + & 0 \le \beta < \max\{0, \frac{h_p - 2n(v-p)}{2n\theta}\} \\ \frac{n}{h_p}\Big[2\theta\beta + (2-v+p)(v-p)\Big](p - c_p - \theta\beta) & \\ \frac{(h_p - 2n\theta\beta)^2}{4nh_p}(p-c_o) + (1 - \frac{(h_p-2n\theta\beta)^2}{4nh_p})(p - c_p - \theta\beta), & \max\{\frac{h_p-2n(v-p)}{2n\theta}, 0\} \le \beta < \frac{h_p}{2n\theta} \\ (p - c_p - \theta\beta) & \beta \ge \frac{h_p}{2n\theta} \end{cases}$$

As in Proposition 5, we can show that $\pi_o^C(\beta)$ is unimodal so that there exist unique $\bar{\beta}$ and $\underline{\beta}$ such that the coupon policy is beneficial when $\beta \le [\underline{\beta}, \bar{\beta}]$. Consequently, the optimal coupon value from the online retailer's perspective is $\beta^* = \max\{c_s - r, \check{\beta}\}$. The expression of $\check{\beta}$ depends on the vale of n as follows:
(i) If $n < \lceil \frac{h_p}{2(v-p)} \rceil$, then we have

$$\check{\beta} = \begin{cases} \frac{1}{6n\theta}\Big[2(n(c_o - c_p) + h_p) - \sqrt{4n(c_o-c_p)(n(c_o-c_p) - h_p) + h_p(h_p + 12n)}\Big], & c_p \le c_1 \\ \frac{h_p - 2n(v-p)}{2n\theta}, & c_1 < c_p \le c_2 \\ \frac{(p-c_p)}{2\theta} - \frac{(v-p)(p-c_o)}{2\theta} - \frac{(2-v+p)(v-p)}{4\theta}, & c_2 < c_p \le c_3 \\ 0, & c_p > c_3 \end{cases}$$

Here $c_1 = c_o - \frac{h_p}{2n} + \frac{3(v-p)}{2} - \frac{h_p}{2n(v-p)}$, $c_2 = p - (v-p)(p - c_o - 2) + \frac{(v-p)^2}{2} - \frac{h_p}{n}$, and $c_3 = p - (v-p)(p - c_o + 1) + \frac{(v-p)^2}{2}$.
(ii) If $n \ge \lceil \frac{h_p}{2(v-p)} \rceil$, then we have

$$\check{\beta} = \begin{cases} \frac{1}{6n\theta}\Big[2(n(c_o - c_p) + h_p) - \sqrt{4n(c_o-c_p)(n(c_o-c_p) - h_p) + h_p(h_p + 12n)}\Big], & c_p \le c_o + \frac{h_p}{4n} - 1 \\ 0 & c_p > c_o + \frac{h_p}{4n} - 1 \end{cases}$$

The derivation of the above expressions is similar to the derivation of the expressions in the proof of Proposition 5.

Therefore, when $n < \lceil \frac{h_p}{2(v-p)} \rceil$, all of our previous findings still hold. When $n > \lceil \frac{h_p}{2(v-p)} \rceil$ (i.e., the case of market saturation under the pickup partnership), the coupon policy can be optimal only when the profit margin of in-store pickup orders is higher than the profit margin of direct-delivery orders (i.e., $c_p + \theta\beta < c_o$).

Based on the above findings, we can now obtain the optimal number of pickup locations under each policy. Suppose that $\alpha \in [\underline{\alpha}, \bar{\alpha}]$, the online retailer's profit as a function of n is given by:

$$\pi_o^F = \begin{cases} (p-c_o)(1-\frac{n(v-p)}{h_p})(v-p) + (p-c_p-\alpha)(2-v+p)\frac{(v-p)n}{h_p}, & 1 \le n \le \lceil \frac{h_p}{2(v-p)} \rceil \\ \frac{h_p}{4n}(p-c_o) + (p-c_p-\alpha)(1-\frac{h_p}{4n}), & n \ge \lceil \frac{h_p}{2(v-p)} \rceil \end{cases}$$

We assume that n is a continuous variable, so that the derivative of π_o^F can be written as

$$\frac{\partial \pi_o^F}{\partial n} = \begin{cases} -\frac{(p-c_o)(v-p)^2}{h_p} + (p-c_p-\alpha)\frac{(2-v+p)(v-p)}{h_p}, & 1 \le n < \frac{h_p}{2(v-p)} \\ \frac{h_p}{4n^2}(c_o-c_p-\alpha), & n \ge \frac{h_p}{2(v-p)} \end{cases}$$

Since we assume that $\alpha \in [\underline{\alpha}, \bar{\alpha}]$, we have $\frac{\partial \pi_o^F}{\partial n} > 0$ when $n < \frac{h_p}{2(v-p)}$. Thus, the optimal n under the fixed fee policy is given by:

$$n^* = \begin{cases} +\infty & \alpha \le c_o - c_p \\ \lfloor \frac{h_p}{2(v-p)} \rfloor & \alpha > c_o - c_p \end{cases}$$

Under the coupon policy, we have:

$$\frac{\partial \pi_o^C}{\partial n} = \begin{cases} \frac{(v-p)}{h_p}\Big[(2-v+p)(p-c_p-\theta\beta)-(p-c_o)(v-p)\Big]+ & 1 \le n < \lceil \frac{h_p}{2(\theta\beta+v-p)} \rceil \\ \frac{2\theta\beta}{h_p}\Big[p-c_p-\theta\beta-(v-p)(p-c_o)\Big], & \\ \frac{(h_p^2-4n^2\theta\beta^2)(c_o-c_p-\theta\beta)}{4h_p n^2}, & \lceil \frac{h_p}{2(\theta\beta+v-p)} \rceil \le n < \lceil \frac{h_p}{2\theta\beta} \rceil \\ 0 & n \ge \lceil \frac{h_p}{2\theta\beta} \rceil \end{cases}$$

Therefore, the optimal n under the coupon policy is given by:

$$n^* = \begin{cases} \lceil \frac{h_p}{2\theta\beta} \rceil & \theta\beta \le c_o - c_p \\ \lfloor \frac{h_p}{2(\theta\beta+v-p)} \rfloor & \theta\beta > c_o - c_p \end{cases}$$

By comparing n^* under the fixed fee and coupon policies, we conclude that the online retailer prefers a larger number of pickup locations under the fixed fee policy. We next evaluate how the profitability of the coupon policy depends on n. We let $\Delta\pi_o^{C-F}$ denote the difference between the online retailer's profit under the fixed fee and the coupon policies when $\alpha = \theta\beta$. We have

$$\Delta\pi_o^{C-F} = \begin{cases} \frac{2n\alpha}{h_p}\Big[p-c_p-\alpha-(v-p)(p-c_o)\Big], & 1 \le n < \lceil \frac{h_p}{2(\alpha+v-p)} \rceil \\ \frac{(h_p-2n\alpha)^2}{4nh_p}(c_p+\alpha-c_o) - \frac{n(v-p)(1-v+p)}{h_p}(p-c_p-\alpha), & \lceil \frac{h_p}{2(\alpha+v-p)} \rceil \le n < \lceil \frac{h_p}{2(v-p)} \rceil \\ +(1-\frac{n(v-p)}{h_p})\Big[p-c_p-\alpha-(v-p)(p-c_o)\Big] & \\ (\frac{h_p}{4n} - \frac{(h_p-2n\alpha)^2}{4nh_p})(c_o-c_p-\alpha), & \lceil \frac{h_p}{2(v-p)} \rceil \le n < \lceil \frac{h_p}{2\alpha} \rceil \\ \frac{h_p}{4n}(c_o-c_p-\alpha) & n \ge \lceil \frac{h_p}{2\alpha} \rceil \end{cases}$$

When $\alpha = \theta\beta \le c_o - c_p$, we have $\Delta_o\pi^{C-F} \ge 0$. We next compute the first-order derivative:

$$\frac{\partial \Delta\pi_o^{C-F}}{\partial n} = \begin{cases} \frac{2\alpha}{h_p}\Big[p - c_p - \alpha - (v-p)(p-c_o)\Big] & 1 \le n < \lceil \frac{h_p}{2(\alpha+v-p)} \rceil \\ \frac{h_p - 2n\alpha}{nh_p}\Big[\alpha + \frac{(h_p-2n\alpha)}{4n}\Big](c_o - \alpha - c_p) + & \lceil \frac{h_p}{2(\alpha+v-p)} \rceil \le n < \lceil \frac{h_p}{2(v-p)} \rceil \\ \frac{(v-p)}{h_p}\Big[(v-p)(p-c_o) - (2-v+p)(p-c_p-\alpha)\Big] & \\ \Big[\frac{(h_p-2n\alpha)^2}{4n^2h_p^2} + \frac{\alpha(h_p-2n\alpha)}{nh_p} - \frac{h_p}{4n^2}\Big](c_o - c_p - \alpha) & \lceil \frac{h_p}{2(v-p)} \rceil \le n < \lceil \frac{h_p}{2\alpha} \rceil \\ -\frac{h}{4n^2}(c_o - c_p - \alpha) & n \ge \lceil \frac{h_p}{2\alpha} \rceil \end{cases}$$

Therefore, when $\alpha = \theta\beta \le c_o - c_p$, one can show that $\Delta\pi_o^{C-F}$ is increasing in n when the market is not saturated (i.e., $1 \le n < \lceil \frac{h_p}{2(\alpha+v-p)} \rceil$) and is decreasing for $n > \lceil \frac{h_p}{2(\alpha+v-p)} \rceil$. In addition, when $n \to \infty$, we have $\Delta\pi_o^{C-F} \to 0$.
When $\alpha = \theta\beta > c_o - c_p$, since $\alpha \le \frac{\bar{\beta}}{\theta} \le \bar{\alpha} = p - c_p - \frac{v-p}{2-v+p}(p - c_o)$, we can conclude that $\Delta\pi_o^{C-F}$ is increasing in n when the market is not saturated (i.e., $1 \le n < \lceil \frac{h_p}{2(\alpha+v-p)} \rceil$). When the market is saturated, the profitability of the coupon policy over the fixed fee policy decreases with n, and there exists a unique $n \in [\lceil \frac{h_p}{2(\alpha+v-p)} \rceil, \lceil \frac{h_p}{2(v-p)} \rceil]$, so that $\Delta\pi_o^{C-F}$ becomes negative. Then, for $n > \lceil \frac{h_p}{2(v-p)} \rceil$, $\Delta\pi_o^{C-F}$ starts to increase with n again, and when $n \to \infty$, we have $\Delta\pi_o^{C-F} \to 0$.

Total Welfare

We find that when the objective is set to maximize the total welfare, the optimal policy is either the baseline policy when c_s is high or the coupon policy when c_s is low. To show this, we first prove that when $c_s > p + r - c_p - \frac{v-p}{2-v+p}[p - c_o + \frac{2(v-p)}{3} - 1]$, the total welfare under the baseline policy is higher than the total welfare under the fixed fee policy. Since the values of total welfare under the fixed fee and baseline policies are constant, this can be easily shown by comparing the total welfare under both policies. We let TW^F and TW^B denote the total welfare under the fixed fee and baseline policies, respectively. Then, TW^F and TW^B can be characterized as follows:

$$TW^B = (p - c_o)(v - p) + \frac{(v-p)^2}{2},$$

$$TW^F = (p-c_o)\Big[1-\frac{(v-p)}{h_p}\Big](v-p)+(p+r-c_p-c_s)\frac{(v-p)(2-v+p)}{h_p}+\frac{(v-p)^2}{2}\Big[\frac{2}{h_p}+1-\frac{4(v-p)}{3h_p}\Big].$$

By comparing TW^B to TW^F, we can show that when $c_s > p + r - c_p - \frac{v-p}{2-v+p}(p - c_o + \frac{2(v-p)}{3} - 1)$, we have $TW^B \geq TW^F$.

We next show that when $c_s < p + r - c_p - \frac{v-p}{2-v+p}[p - c_o + \frac{2(v-p)}{3} - 1]$, the coupon policy always improves the total welfare relative to the fixed fee policy. To do so, we show that when $\beta = 0$, the total welfare is equal between the coupon and the fixed fee policies, and then the total welfare under the coupon policy increases with β. Since the total welfare under the fixed fee policy is constant, this will conclude the argument. We let $TW^C(\beta)$ denote the total welfare under the coupon policy. It is then enough to show that the first derivative of $TW^C(\beta)$ is positive. The total welfare under the coupon policy is given by:

$$TW^C(\beta) = \begin{cases} (p-c_o)\Big[1-\frac{(v-p+2\theta\beta)}{h_p}\Big](v-p) + (p+r-c_p-c_s)\Big[\frac{2\theta\beta+(2-v+p)(v-p)}{h_p}\Big] + \\ \frac{(\theta\beta)^2+\theta\beta(v-p)(2-v+p)}{h_p} + \frac{(v-p)^2}{2}\Big[\frac{2}{h_p}+1-\frac{4(v-p)}{3h_p}\Big], & 0 \leq \beta \leq \frac{h_p-2(v-p)}{2\theta} \\ (v+r-c_p-c_s) - \frac{(h_p-2\theta\beta)^2}{4h_p}(r+c_o-c_p-c_s) + \\ \Big[\frac{(\theta\beta)^2}{2} - \frac{(\theta\beta)^3}{3h_p} + \theta\beta(1-\frac{h_p}{4}) + h_p(\frac{h_p}{24}-\frac{1}{4})\Big], & \frac{h_p-2(v-p)}{2\theta} \leq \beta \leq \frac{h_p}{2\theta} \\ v+\theta\beta+r-c_s-c_p-\frac{h_p}{4}, & \beta \geq \frac{h_p}{2\theta} \end{cases}$$

It is clear that when $\beta = 0$, $TW^C(\beta) = TW^F$. The first derivative of $TW^C(\beta)$ is given by:

$$\frac{\partial TW^C(\beta)}{\partial \beta} = \begin{cases} \frac{\theta}{h_p}\Big[2(p+r-c_p-c_s) - 2(v-p)(p-c_o) + \theta\beta + (v-p)(2-v+p)\Big], & 0 \leq \beta \leq \frac{h_p-2(v-p)}{2\theta} \\ \theta\Big[(\frac{h_p-2\theta\beta}{h_p})(r+c_o-c_p-c_s) + \theta\beta - \frac{(\theta\beta)^2}{h_p} + 1 - \frac{h_p}{4}\Big], & \frac{h_p-2(v-p)}{2\theta} \leq \beta \leq \frac{h_p}{2\theta} \\ \theta & \beta \geq \frac{h_p}{2\theta} \end{cases}$$

For the first part, we need to show that for $\beta \in [0, \frac{h_p-2(v-p)}{2\theta}]$, we have

$$\frac{\theta}{h_p}\Big[2(p+r-c_p-c_s) - 2(v-p)(p-c_o) + \theta\beta + (v-p)(2-v+p)\Big] \geq 0.$$

Therefore, it is enough to show that

$$p+r-c_p-c_s-(v-p)(p-c_o)+\frac{(v-p)(2-v+p)}{2} \geq 0 \implies c_s \leq p+r-c_p+\frac{(v-p)(2-v+p)}{2}-(v-p)(p-c_o).$$

Since we assume that $c_s \leq p + r - c_p - \frac{v-p}{2-v+p}[p - c_o + \frac{2(v-p)}{3} - 1]$, it is enough to show that

$$p+r-c_p+\frac{(v-p)(2-v+p)}{2}-(v-p)(p-c_o) \geq p+r-c_p-\frac{v-p}{2-v+p}\Big[p-c_o+\frac{2(v-p)}{3}-1\Big] \tag{A2.6}$$

or equivalently

$$\frac{v-p}{2-v+p}\Big[1-\frac{2(v-p)}{3}-p+c_o\Big] \leq \frac{(v-p)(2-v+p)}{2}-(v-p)(p-c_o).$$

Note that $[1 - \frac{2(v-p)}{3} - p + c_o] > 0$, $v - p > 0$, and $2 - v + p > 1$, so we have

$$\frac{v-p}{2-v+p}\Big[1-\frac{2(v-p)}{3}-p+c_o\Big] \leq (v-p)\Big[1-\frac{2(v-p)}{3}-p+c_o\Big]$$
$$\leq (v-p)\Big[1-\frac{v-p}{2}-p+c_o\Big] = \frac{(v-p)(2-v+p)}{2}-(v-p)(p-c_o).$$

For the second part, we need to show that for $\beta \in [\frac{h_p-2(v-p)}{2\theta}, \frac{h_p}{2\theta}]$, we have

$$\theta\Big[(\frac{h_p-2\theta\beta}{h_p})(r+c_o-c_p-c_s)+\theta\beta-\frac{(\theta\beta)^2}{h_p}+1-\frac{h_p}{4}\Big] \geq 0.$$

The above equation is a continuous concave quadratic function of β and, thus, it is enough to show that this function is positive at both threshold values (i.e., $\frac{h_p-2(v-p)}{2\theta}$ and $\frac{h_p}{2\theta}$). For $\beta = \frac{h_p}{2\theta}$, we have

$$\frac{h_p}{2}-\frac{1}{h_p}(\frac{h_p}{2})^2+1-\frac{h_p}{4} = 1 > 0,$$

and for $\beta = \frac{h_p - 2(v-p)}{2\theta}$, we have

$$\theta\Big[\frac{2(v-p)}{h_p}(r+c_o-c_p-c_s)+\frac{h_p}{2}-(v-p)-\frac{1}{4h_p}(h_p^2+4(v-p)^2-4h_p(v-p))\Big] \geq 0$$

$$\frac{2(v-p)}{h_p}(r+c_o-c_p-c_s)-\frac{(v-p)^2}{h_p}+1 \geq 0 \implies c_s \leq r+c_o-c_p-\frac{v-p}{2}+\frac{h_p}{2(v-p)}.$$

Since we assume $c_s \leq p+r-c_p-\frac{v-p}{2-v+p}(p-c_o+\frac{2(v-p)}{3}-1)$, it is enough to show that

$$r+p-c_p+\frac{v-p}{2-v+p}\Big[1-\frac{2(v-p)}{3}-p+c_o\Big] \leq r+c_o-c_p-\frac{v-p}{2}+\frac{h_p}{2(v-p)}.$$

By using Inequality (A2.6), it is enough to show that

$$p+r-c_p+\frac{(v-p)(2-v+p)}{2}-(v-p)(p-c_o) \leq r+c_o-c_p-\frac{v-p}{2}+\frac{h_p}{2(v-p)}$$

$$\implies \frac{h_p}{2(v-p)}-\frac{v-p}{2}-p+c_o \geq (v-p)\Big[1-\frac{v-p}{2}-p+c_o\Big],$$

where the last inequality follows from $v-p \leq 1$ and $\frac{v-p}{h_p} \leq \frac{1}{2}$, and hence concludes the proof.

4

Vertical Product Location Effect on Sales: A Field Experiment in Convenience Stores

In the previous chapter, we introduced a stylized model and demonstrated its efficacy in optimizing retail operations. Building upon this foundation, subsequent chapters will delve into utilizing data to further enhance retail efficiency. A significant challenge faced by most brick-and-mortar stores revolves around strategically placing assortments within limited space to attract customers in today's fiercely competitive marketplace. Addressing this challenge necessitates an understanding of how product placement influences sales. However, deciphering this relationship is hindered by confounding variables within current store layouts, making it impossible to directly discern the impact from observational data alone.

To overcome this hurdle, we propose innovative experiments that leverage data to provide insights into one of the most pressing questions for brick-and-mortar retailers. By designing experiments tailored to isolate the effects of product placement, we aim to offer actionable solutions that empower retailers to optimize their store layouts and drive sales in an increasingly competitive landscape.

4.1 Introduction

It is a well-known fact that customers rely on shelf layout to infer product quality (Valenzuela and Raghubir 2015, Atan et al. 2023), product popularity (Valenzuela et al. 2013), and brand value (Parker and Koschmann 2018), which collectively influence their product selection. As a result, retailers seek effective shelf layout strategies to convert shoppers into buyers. At the same time, manufacturers routinely negotiate with retailers for favorable placements on retailers' limited shelf space to enhance their products' visibility and increase their market share. In the present study, among the various dimensions defining shelf layout (e.g., total space, vertical location, horizontal location, number of facings), we focus on the vertical shelf location of products.

Vertical location indicates the height of the position occupied by a specific product (Chen et al. 2021) and has been identified as a critical dimension of shelf layout because of its direct impact on product visibility. Vertical location is divided into four zones: stretch-level, eye-level, touch-level, and stoop-level. A shelf at the eye-level is located at (or near) an average adult's eye-level when standing upright in front of the shelf. A shelf at the touch-level is located approximately at an adult's waist height and is considered the "eye-level" for kids. A shelf at the stretch-level is located above the eye-level and requires customers to stretch upward to reach products. Finally, a shelf at the stoop-level is located below the touch-level and requires customers to bend down to reach products (Ebster 2011). As implied by the famous motto "eye-level is buy level," retailers strategically place products with higher profitability at the eye-level to boost sales (Ausick 2017), and most retailers require an additional fee (called slotting fee) from manufacturers that want to secure the prime eye-level location in retail stores (Alexander 2003, Rivlin 2016, Meyersohn 2022, Logie 2022). In the early 2000s, 85% of retailers were charging slotting fees to manufacturers, and today, slotting fees have become a standard in the retail industry (Chelstad 2018). As such, the Center for Science and the Public Interest (CSPI) estimates that the annual slotting fees paid by manufacturers to retailers amount to $50 billion (Rivlin 2016), representing more than 16% of manufacturers' new product introduction costs (Touche et al. 1990, Sudhir and Rao 2006).

Although the practice of offering the eye-level location as a sales-enhancing lever to manufacturers is prevalent and well established in practice, the way that retailers reorganize their shelf layout to take advantage of the eye-level effect is not as straightforward. As a motivation, consider the situation where a retailer sells three products on a shelving unit using the following planogram: product A is located on the stretch-level shelf, product B is located on the eye-level shelf, and product C is located on the stoop-level shelf. Suppose that the manufacturer of product A and the manufacturer of product C both would like to secure the eye-level shelf. Which product should the retailer move there? Once the retailer has decided on the eye-level, which product should be displayed at the stretch-and stoop-levels, respectively? Consider the two planogram changes displayed in Figure 4.1.1, where products A and C are moved to the eye-level in the left and right planograms, respectively. In essence, these changes lead to two different planograms compared with the initial planogram in the middle. Although moving product A or C to the eye-level will likely increase the sales of these products, it is not

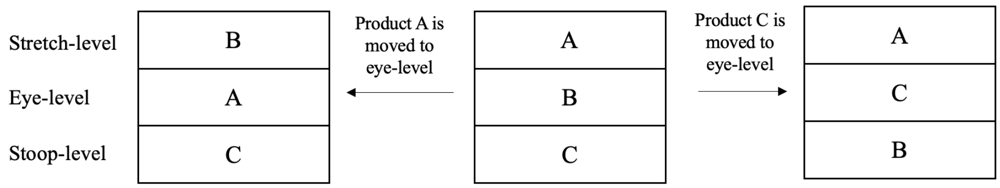


Figure 4.1.1: Illustration of two potential new planograms derived from the base planogram (shown in the middle).

clear what will happen to the sales of the other products that are moved to the stretch- and stoop-levels. Will the sales increase for the eye-level shelf be offset by the decrease in sales for products on other shelves? Does this interplay differ in the left and right planograms? The answers to these questions are of practical importance to retailers when deciding which new planogram to operationalize, and uncovering this requires a complete understanding of the effect of simultaneously changing the vertical locations for multiple products. Surprisingly, we are not aware of an empirical study that examines (i) the interplay among a set of products for which the vertical locations change simultaneously and (ii) how this interplay influences the overall sales. This leads to the following research question: *How does simultaneously changing vertical product locations of a set of products influence product-level sales and overall sales?*

The aforementioned interplay may not be the same for all products. Consider products with different market shares, price points, or market concentration levels. Does the eye-level effect increase or decrease if we place at the eye-level shelf a product with a high market share compared to a product with a low market share? Similarly, what happens to the eye-level effect of a product on promotion compared to a product with no promotion? These questions are practically important because, despite being placed in favorable shelf locations, 80–90% of products fail to generate the desired sales, according to a report published by the U.S. Federal Trade Commission (Federal Trade Commission 2001). Consistent with this estimate, many manufacturers re-

port that some products fail to earn enough revenue to cover the slotting fees (Alexander 2003). Overall, these anecdotal observations imply that the eye-level effect—and more broadly the vertical location effect—is likely to be heterogeneous across products. Particularly, if the increase in revenue from displaying a product at the eye-level does not outweigh the overall revenue loss (if any) across all products available in a planogram, the retailer would be better off by not bringing that product to the eye-level or by charging its manufacturer a higher slotting fee. The empirical literature has remained silent on the heterogeneity of the vertical product location effect across all products in a planogram. This brings us to our second research question: *What type of products benefit the most from being displayed at the eye-level?*
To answer our research questions, we conducted a field experiment in collaboration with one of the largest retail chains operating convenience stores (C-stores). Targeting 818 products displayed on the stretch-level, eye-level, and stoop-level shelves, our field experiment was carried out across six convenience stores located in a North American metropolitan area for 20 weeks in 2022. Designing a field experiment within the constraints of the real-world retail realm while controlling for all sources of variation across observations is naturally challenging (Gallino and Moreno 2018, Bianchi-Aguiar et al. 2021, Bandi et al. 2022). To address this challenge, we designed a novel two-stage field experiment using a treatment-switch approach where we systematically switched the planograms over time at the treatment stores while keeping them constant at the control stores. This design enabled us to use an identification strategy to accurately estimate the causal effect of the vertical location change on sales.
Our research provides several fundamental contributions, which can be summarized as follows:

- We find that the effect of changing the vertical location of a product depends on how the vertical locations of other products are reorganized as a result of that change. For instance, we estimate that bringing a product from the stretch-level to the eye-level with a particular planogram change increases sales by 9.2%. However, moving a product from the eye-level back to the stretch-level with a different planogram change decreases sales by only 6.9%, as opposed to an expected decrease of 9.2%. The difference in these two estimated effects (despite representing the same—yet reverse—vertical location changes) arises because the vertical locations of the other products differ between the two planograms. This highlights the interplay among products when the vertical location changes simultaneously for multiple products.

- When we average these asymmetric effects, we find strong evidence for the vertical location effect and quantify it in our setting. Relative to displaying a product at the stoop-level, displaying the same product at the eye-level and stretch-level increases its sales on average by 13.8% and 5.3%, respectively. Despite these vertical location effects on individual products, when we consider all products across the three shelves, we find that the overall sales remain the same among all planograms. This implies that the vertical location effect on sales is likely to arise because of a pure substitution effect among products. As such, the incremental sales for products displayed at the eye-level just compensate for the sales loss for products displayed at the stretch- and stoop-levels.

- Our results demonstrate that the vertical location effect is amplified when we consider dollar sales. Relative to the stoop-level, the sales are, on average, 31.2% and 14.6% higher when a product is placed at the eye- and stretch-levels, respectively. When we consider the sales for all products across the three shelves, unlike the volume sales, we find that the overall dollar sales are significantly higher for certain planograms. This implies that, for certain planograms, the incremental dollar sales for products displayed at the eye-level can be higher than the dollar sales loss from the products displayed at the other levels, hence resulting in increased revenue for retailers. Therefore, despite the potential pure substitution among products displayed on different shelves, retailers can still increase the overall dollar sales by selectively placing certain products at the eye-level. To identify such products, we examine the heterogeneity of the vertical location effect with respect to several product characteristics. We find that the vertical location effect varies by product profile. In particular, products with high prices, low discounts, that are displayed in a complete assortment, that are in a category with a low market concentration, and that have a high number of substitute products will benefit more from being displayed at the eye-level. This finding suggests that optimizing planograms while accounting for the heterogeneity in the vertical location effect across products can boost profits.

- We conduct a counterfactual analysis and find that, compared with the business-as-usual planograms, planograms that are optimized by considering the average effect of the interplay among a set of products (arising from simultaneously changing vertical locations) can increase the profit by 2.2%. Given the fact that profit margins in the retail sector hover around single low digits, an increase of 2.2% in profits shows the importance of considering planogram optimization as a holistic process rather than just focusing on the eye-level. In addition to considering the effect of the interplay, when we also account for the heterogeneity in that effect across products, we find that, compared with the business-as-usual planograms, the planogram optimization can increase the profit by 3%, hence testifying to the economic significance of our empirical findings.

- As a secondary contribution, our study develops a novel two-stage experimental design that is less invasive to retailers' daily store operations, reduces the time and budget required for a full factorial design by half, and still accurately captures the interplay among a set of products displayed at different vertical locations.

4.2 Related Literature

The retail practice of offering eye-level location as a sales-enhancing lever has received significant attention in academic research. Most of empirical research on the vertical location effect has been conducted using controlled laboratory experiments, with a few exceptions that use secondary data. These experimental studies either use eye-tracking technology in a physical store to examine the visual attention toward products placed on different shelves (e.g., Chen et al. 2021) or are based on a virtual setting that aims to mimic a physical retail store on a computer screen to understand customer intentions toward vertical location placement (e.g., Chandon et al. 2009, Valenzuela et al. 2013, Valenzuela and Raghubir 2015). Consistent with retail practice, these studies demonstrate that products located on eye-level or touch-level shelves gain more visual attention relative to those placed on stretch-level or stoop-level shelves. The additional attention attributed to the eye-level and touch-level positions has been shown (i) to influence brand evaluation (Chandon et al. 2009), price perception, and quality perception (Valenzuela and Raghubir 2015), (ii) to attenuate as customers navigate through the store (Chen et al. 2021) or when the store layouts are not informative regarding product popularity (Valenzuela et al. 2013), and (iii) not to arise from a tendency to look more at the center (Atalay et al. 2012). Consistent with the experimental studies, empirical studies that leverage secondary data from grocery retailers also report that sales increase as the location of a product gets closer to the eye-level (Dreze et al. 1994, Van Nierop et al. 2008). One exception is Frank and Massy (1970), who find an insignificant vertical location effect on sales. However, the authors acknowledge that their results can be because of the bias arising from the uncontrolled variables in their model. In summary, the literature provides ample empirical evidence for the eye-level effect. However, this stream of the literature examines the vertical location effect on sales from the perspective of an individual product without considering the lateral effect on the products located on other shelves.

Thus, it ignores how (i) moving a product to an eye-level shelf influences sales for products that are simultaneously moved to other shelves and (ii) the rearrangement of vertical locations of several products on a shelving unit influences the overall sales on that shelving unit. Furthermore, even though the empirical literature demonstrates that the vertical location effect can vary across product categories displayed at different shelving units (e.g., the eye-level effect for canned soup shelves vs. bath tissues shelves) (Dreze et al. 1994), we are not aware of any paper that empirically examines the heterogeneity in the vertical location effect across products within the same shelving unit (e.g., the eye-level effect for Progresso chicken noodle soup vs. Pacific chicken and wild rice soup, which are both displayed on the same shelving unit).

Even when we turn to the literature relying on analytical models, we do not find any clear suggestions for these aforementioned gaps. For instance, with respect to the overall sales, some papers (e.g., Hansen et al. 2010, Gencosman and Begen 2022) assume that the vertical location change simply results in demand substitution among products, without affecting the overall sales, whereas others (e.g., Smirnov and Huchzermeier 2019, Hübner et al. 2021) assume that such a change can enhance the overall sales. Similarly, although most analytical papers assume a homogeneous vertical location effect across products within a shelving unit (e.g., Bianchi-Aguiar et al. 2016, Smirnov and Huchzermeier 2019, Hübner et al. 2021, Gencosman and Begen 2022), a few papers consider that the vertical location effect is heterogeneous (e.g., Van Nierop et al. 2008, Russell and Urban 2010). Collectively, these studies motivate competing hypotheses regarding (i) the effect of the vertical location change on the overall sales, and (ii) the heterogeneity in the vertical location effect across products within a shelving unit, which are amenable to empirical validation.

Overall, the present paper contributes to the literature by conducting the first field experiment (i) to identify the effect of the vertical location change on individual product and overall sales, and (ii) to assess the heterogeneity in the vertical location effect across products within a shelving unit.

4.3 Experiment Design

In this section, we present the design of our field experiment. We partnered with a global convenience store chain that has an extensive presence in many countries, encompassing thousands of stores worldwide. Our field experiment was conducted in a metropolitan area in North America, where our partner company has over 100 stores. These stores operate 24 hours a day, seven days a week. Strategic store-level operational decisions (i.e., product assortments, pricing, planogram design, and promotions) are centrally made by the regional marketing and operations teams, whereas day-to-day operational decisions (i.e., inventory and ordering) are made in a decentralized manner by the store managers. The planograms are designed by using a simple approach that relies on expert opinions (rather than using data-driven optimization techniques). For instance, each shelving unit in a store is assigned to a specific product category, and products are typically arranged vertically based on their size, with smaller items occupying the top shelves and larger items occupying the bottom shelves. Based on the contractual agreements with the suppliers, certain products need to be placed in specific store locations. However, because of the requirements of our field experiments, we received permission to temporarily deviate from these arrangements.

In collaboration with our retail partner, we designed a novel, two-stage experiment using a treatment-switch approach where the planograms would be systematically switched over time at the treatment stores, whereas they would remain constant at the control stores. This design provides an identification strategy that can accurately estimate the causal effect of the vertical location on the focal products' sales, as well as on total sales. Similar designs have been used in the context of airline pricing (Cohen et al. 2023) and hotel revenue management (Lopez Mateos et al. 2022). We next discuss the various parameters of our design.

Shelving unit: The retailer carries over 1,500 stock-keeping units (SKUs) across 35 product categories. The products are displayed on different shelving units, including fridges, gondolas, checkout desks, and promotion stands, which differ in height and number of vertical shelves. Our study focuses on fridges for the following reasons: First, similar fridges are available in other retail settings, such as supermarkets and grocers, hence enhancing the generalizability of our study. Second, unlike other shelving units, such as gondolas and checkout desks, whose heights range from 40″ to 60″, the height of fridges is 85″, allowing us to examine all vertical positions, including the stretch-level. Third, products displayed on fridges at our retail partner represent 38% of all product offerings and account for 45% of the overall sales, hence allowing us to examine the vertical location effect across a broad range of products. The product categories stored in fridges include juices, water, energy drinks, sports drinks, carbonated soft drinks, beer, and wine.

Treated shelves: In C-stores, fridges typically consist of six shelves. As illustrated in Figure 4.3.1, our intervention in the experiment focuses on the first (from the top), second, and fifth shelves. The height from the floor is 72″ for the first shelf, 60″ for the second shelf, and 24″ for the fifth shelf. These heights correspond to the commonly known stretch-level, eye-level, and stoop-level in the literature (Chen et al. 2021). Thus, we name the first shelf *the stretch-level*, the second shelf *the eye-level*, and the fifth shelf *the stoop-level* throughout the paper.[1] Focusing on these three shelves makes a full-factorial field experiment (i.e., testing all possible combinations of product locations) feasible,[2] while capturing the essence of the vertical location dimension of shelf layout (i.e., stretch-, eye-, and stoop-levels).

[1] We exclude the sixth shelf (i.e., the closest to the floor) from our analysis for examining the stoop-level because the products placed on this shelf are large and heavy and, thus, cannot be moved to other shelves.

[2] An experiment involving three shelves results in $3! = 6$ different planograms, whereas an experiment with six shelves would result in $6! = 720$ different planograms, which is clearly cost-prohibitive and infeasible.

Figure 4.3.1: Illustration of a fridge and treated shelves.

Experimental unit: In the experiment, we switch the planograms by moving all the products on a shelf to a different shelf. This allows us to keep the number of facings and horizontal location of each product constant throughout the experiment and attribute any effect on sales to the change in the vertical location (i.e., our treatment). Consistent with this approach, we define the experiment unit as the set of products placed on a given shelf (i.e., hereafter referred to as a *product set*), rather than an individual product on a shelf. This selection allows us to estimate the average vertical location effect across all potential numbers of facings and horizontal locations (whereas the vertical location effect estimated with an individual product being the treatment unit would be conditional on the number of facings and the horizontal location). This selection is also consistent with the retailer's store execution practices, given that managers often prefer to display products within the same brand and/or the same category in adjacent locations.

Design and treatment: Our goal is to design an experiment within the real-world retail realm while controlling for any source of variation across observations. Changing the vertical location of product sets at the three treated shelves would result in six different planograms, as illustrated in Figure 4.3.2. The effect of a vertical location change can be identified if a product set is observed across all planogram combinations. However, a store can implement only one planogram for a fridge at a time. To overcome this shortcoming, we consider several stores and observe a product set across all planograms at the same time by launching different planograms across several stores. This, however, may introduce between-store variation across observations. To remove this variation, we also implement all combinations of planograms within stores over time, by launching one planogram at a time while ensuring that a planogram can be observed in at least one store at a given time. Although this approach is robust to between-store variation, it results in temporal variation within the treated stores. To disentangle the temporal variation, we use a group of control stores where the planograms

	P1	P2	P3	P4	P5	P6
Stretch-level	A	C	B	C	B	A
Eye-level	B	A	C	B	A	C
Stoop-level	C	B	A	A	C	B

Figure 4.3.2: All potential planograms with three vertical locations.

remain unchanged throughout the experiment so the change in sales at those stores is only subject to temporal factors.

Overall, a robust full-factorial design requires six treated stores (to implement each planogram in one store at a time) and six rounds of planogram changes over time (to implement all planograms in each store over time). Considering one control store for each treated store, the full-factorial design implies that the research team would need to intervene with the daily store operations across 12 stores for a long period of time. The retailer perceive this design to be invasive to its daily store operations and expressed reluctance to accommodate the full-factorial design.

To address the retailer's request for a less invasive experiment while still ensuring a robust design, we propose the following *reduced design*: Consider that across all six planograms (in Figure 4.3.2) required for a full-factorial design, a specific vertical location change for a product in a planogram can be operationalized by replacing that planogram with one of the two potential other planograms. For instance, product set B at the stretch-level in P3 can be moved to the eye-level by switching the planogram either from P3 to P1 or from P3 to P4. In this example, if one assumes that the effect of the vertical location change remains the same for product set B between switching to P1 and switching to P4, the effect of the vertical location change from the stretch-level to eye-level for product set B in P3 can be estimated using only one of the two planograms (i.e., P1 or P4). Note that having the same vertical location change effect for product set B between switching from P3 to P1 and from P3 to P4 would be tantamount to having no change in sales for product set B between P1 and P4. In other words, the assumption implies that when the neighboring product sets are swapped (as is the case for product set B between P1 and P4), the sales for the focal product set remain the same. A similar argument can be developed for P1 and P5 for moving product set C to the stoop-level, as well as P1 and P6 for moving product set A to the stretch-level. We term this assumption the *insensitivity assumption*. This captures the essence of our reduced experiment. In particular, if we can demonstrate that the insensitivity assumption holds, then the effect of the vertical location change can be estimated efficiently using only three planograms. Thus, the full-factorial design with 12 stores over six rounds can be reduced to an experiment with six stores (i.e., three treatment stores and three control stores) over three rounds, translating into a 50% reduction in time/budget and, hence, making the experiment less invasive and costly. We formally test this assumption through another field experiment (which we call the pilot experiment).

A natural question that arises is which three planograms (as shown in Figure 4.3.2) should be used in the reduced design. Among all planograms in Figure 4.3.2, we let P1 represent the business-as-usual status quo at the retailer. Hence, we label any product set that, before the experiment, is located at the stretch-level as *product set A*, at the eye-level as *product set B*, and at the stoop-level as *product set C*. For the reduced experiment, we choose a combination of three planograms, which enable us to observe any product set at all three vertical locations across the combination. There are only two such combinations: P1-P2-P3 and P4-P5-P6. Between these two, we choose the combination P1-P2-P3 for the following reason: the retailer allows us to test a different planogram in a store for a duration of three weeks. Thus, we switch planograms across the treatment stores every three weeks. With these frequent changes, some customers may be confused with new planograms and may not immediately find their desired products. As a result, the changes in sales we observe over time may commingle both the vertical location effect and that from the potential confusion experienced by customers. The combination P1-P2-P3 allows us to observe the business-as-usual planogram (i.e., P1) both before and during the experiment (yet at different times). Therefore, if customer confusion is likely, we can identify it by comparing P1 before the experiment to P1 during the experiment and account for this factor in our identification of the vertical location effect. We use the combination P4-P5-P6 to test the insensitivity assumption.

In light of the above background, we design a two-stage experiment, as illustrated in Figure 4.3.3. In particular, we first observe the business-as-usual planogram (i.e., P1) across all treatment and control stores before the experiment between March 21, 2022, and April 16, 2022 (i.e., Round 0).[3] We then use the data from this pre-intervention round to conduct an A/A test between the treatment and control group stores (as detailed in Section 4.4). We next conduct a pilot experiment between April 18, 2022, and May 28, 2022 (i.e., Round 1). In the pilot experiment, the business-as-usual planograms are replaced with P4, P5, and P6 in each of the three treated stores. The pilot experiment aims to validate the insensitivity assumption (as detailed in Section 4.4). Finally, having established the validity of the reduced design, we conduct the main experiment between May 30, 2022, and August 6, 2022 (i.e., Rounds 2, 3, and 4). In the main experiment, we implement the combination P1-P2-P3 within a store, as illustrated in Figure 4.3.3. We ultimately draw our research insights using the main experiment (as detailed in Section 4.5).

To summarize, our reduced design from Figure 4.3.3 has three features that make our main experiment robust, conditional on the insensitivity assumption. First, it enables us to observe a product set across all three vertical

[3]On March 14, 2022, we visited all six stores and ensured that all the fridges were equally organized based on the P1 planogram. We used the first week in this phase as a warm-up period to allow the store staff and the field team to familiarize themselves with the process and with their responsibilities.

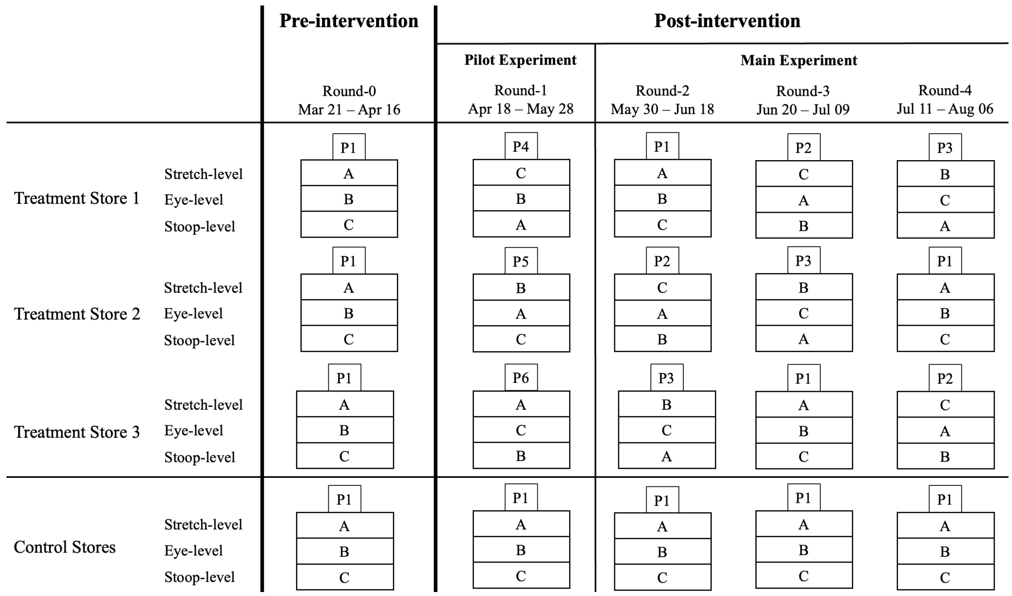


Figure 4.3.3: Field experiment design.

locations at the treated stores. Second, it allows us to observe a product set across all three vertical locations over time within the treated stores and, thus, to control for the between-store variation across observations from the treated stores. Third, by maintaining the business-as-usual planogram (i.e., P1) at the control stores throughout the experiment, we can control for temporal variations across observations from the treated stores.

Treated and control stores: When selecting stores for the treatment and control groups, we considered two factors. First, because our objective is to compare different planograms between the two groups, we identified stores that carry the same assortment of products in refrigerated shelving units. Second, the retailer has both traditional stand-alone C-stores and C-stores located in gas stations. Because the customers of these two types of stores are likely to differ in terms of shopping behavior, we limit our study to the traditional stand-alone C-stores. These two criteria resulted in 20 stores available for our study. Among these, we randomly selected six stores and randomly assigned three of them to the treatment group and three of them to the control group.

Implementation and compliance: To deploy our field experiment, we collaborate with the marketing managers and store staff (i.e., store managers and cashiers). In addition, we form a field team of six trained assistants, each assigned to one of the six stores throughout the experiment. At the beginning of the study, before Round 0, the field team visited all six stores to ensure that the business-as-usual planograms (i.e., P1) are properly respected. During the pilot and main experiments, all planogram implementations at the treated stores were scheduled on Sundays and handled by the field team alongside the store staff. We ensured compliance in two ways. First, the field team visited each store on a daily basis to monitor each planogram, take planogram pictures, ensure that the planograms are organized as planned, and record the inventory, as well as the number of facings per product on each treated shelf. Second, the marketing managers used the same assortment, pricing, and marketing decisions (e.g., promotions, advertising campaigns, etc.) for products in the fridges between the treated and control stores throughout our field experiment.

Data collection, unit of analysis, and variable definitions: We obtained transactional data that included the number of items sold in each transaction, the product price, and the promotional discounts. If we do not observe any sales for a particular product on a specific day, we use the planogram pictures to identify whether this is because of a stockout or absence of demand. Thus, our data also include stockout occurrences.

We define the unit of analysis as a triplet of product set-store-day (denoted by i, j, and t, respectively) and aggregate the transactions over these triplets to obtain sales figures. We operationalize two outcome variables to measure sales: sales quantity and sales revenue. Because the number of facings on a shelf can vary from one product set to another due to the physical size of the products, we normalize our outcome variables by dividing them by the number of facings in a product set. To this end, we operationalize sales quantity, $Quantity_{ijt}$, as the natural logarithm of the average number of items sold per facing in product set i at store j on day t and sales revenue, $Revenue_{ijt}$, as the natural logarithm of the average dollar sales per facing in product set i at store j on day t.[4]

We use several control variables in our analyses. At the product level, we control for the price of all the products in set i on day t ($Price_{it}$) by using the average price of all unique SKUs included in that set. We also control for promotional activities in a fridge using two variables: $SetPromotion_{it}$ represents the average monetary value of discounts across all unique SKUs present in product set i on day t, whereas $FridgePromotion_{it}$ represents the average monetary discount value across all unique SKUs displayed on all other shelves in the same fridge on day t (excluding products in product set i). We also control for stockouts in a fridge using two variables: $SetStockout_{it}$ represents the number of SKUs that are out of stock in product set i on day t, while $FrdigeStockout_{it}$ represents the number of SKUs that are out of stock on all other shelves in the same fridge on day t (excluding products in product set i). The variable $NumFaces_it$ represents the number of faces in product set i on day t. We also control for differences in product categories using indicator variables representing the category where the products in product set i belong to ($ProdCat_i$). At the store level, we control for potential unobserved differences across stores with fixed effects using an indicator variable representing store j that product set i belongs to ($Store_j$). Finally, we control for seasonality using three fixed effects: $Holiday_t$ represents whether day t is a national holiday, $Weekend_t$ represents whether day t is a weekend day, and DoW_t represents the day of the week.

[4]To avoid non-normality because of having a distribution with a long-tail, we take the natural logarithm of a variable when needed, as is commonly done.

4.4 A/A Test and Pilot Experiment

In this section, we conduct the A/A test to assess whether the control group would provide an accurate baseline for the treatment group and examine the insensitivity assumption in our setting.

4.4.1 A/A Test

In our study, the treatment and control groups include stores with the same assortment, subject to the same set of operational decisions (i.e., pricing, inventory, and promotions) under a centralized management style, located in the same metropolitan area, serving similar customers, and, thus, being exposed to comparable geographical and macroeconomic trends. Nonetheless, it is likely that some unobserved factors (e.g., store employee competence, competing businesses in the same neighborhood, etc.) may cause the randomly selected treatment and control stores to be inherently different. If such factors result in treatment and control groups responding to the same intervention differently, then the estimated effects could also be attributed to the unobserved inherent differences between the two groups. We investigate this matter by conducting an A/A test using pre-intervention data (i.e., Round 0) from both the control and treatment group stores.

The data from the A/A test include all the observations in Round 0. Despite all the collaboration efforts with the marketing team to prevent stockouts in the targeted fridges across the six stores, stockouts are alas inevitable (e.g., unpredicted demand spikes, unexpected delays in replenishment lead times caused by manufacturers or third-party logistics providers). As a rule of thumb, we exclude all daily observations from a fridge if more than 30% of the products in that fridge (including those placed on the other shelves not considered in our study) are out of stock. We retain 6,245 observations for the A/A test after removing around 10% of all observations because of stockouts. By conducting a t-test, we can demonstrate that the removed observations are statistically balanced between the control and treated stores

($t = 12.56, p - val = 0.13$). Although we follow the same data exclusion rule across all our analyses throughout the present paper, our results are robust to the inclusion of observations that are removed because of stockouts. Table 4.4.1 provides the descriptive statistics for all continuous variables for the treatment and control group stores. The correlation matrices for all experiments are reported in Appendix A3.1.

Table 4.4.1: Descriptive statistics for the data sample from our A/A test.

	Treated stores		Control stores	
	Mean	SD	Mean	SD
1. Quantity	0.28	0.33	0.53	0.38
2. Revenue	0.64	0.69	1.21	0.77
3. Price	5.53	4.09	5.28	3.98
4. SetPromotion	0.69	0.93	0.70	0.87
5. FridgePromotion	0.51	0.60	0.62	0.73
6. SetStockout	0.46	0.74	0.80	1.07
7. FridgeStockout	0.53	0.43	0.89	0.67
8. NumFaces	5.83	1.51	5.83	1.52
Number of product sets	146		109	
Sample Size	6,245			

As detailed in Section 4.5, we identify the vertical location change effect as the change in sales trend when a product set is moved from one vertical location to another. Here, the control group can serve as a baseline for the treatment group if the sales trend for a product set is the same between both groups when using the same planogram. This is the main essence of our A/A test.

Our dataset represents a panel containing daily observations from 255 product sets over 27 days. Thus, we use a longitudinal panel data regression to compare the trend in outcome variables between the treatment and control groups when operating under the same planogram setting. In this and all subsequent analyses, we use a fixed-effect model because the Hausman test suggests that the random effect formulation is not consistent. We specify our model as follows:

$$Y_{ijt} = \beta_0 + \beta_1 Time_t \times Treatment_j + \beta_2 Time_t + TVControls_{ijt} + Seasonality_t + u_i + \epsilon_{ijt}, \tag{4.4.1}$$

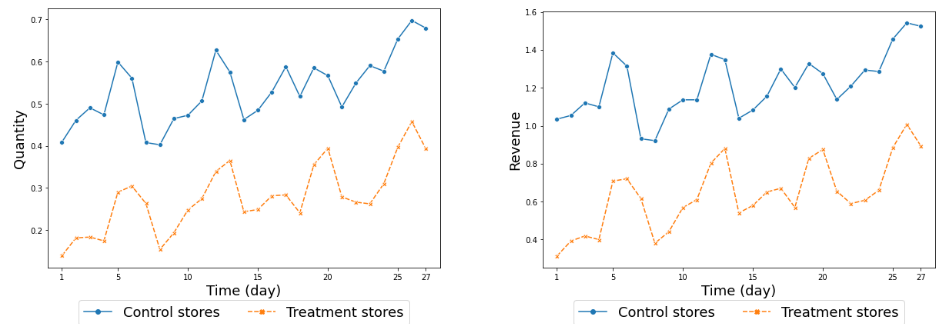
where Y_{ijt} can be either $Quantity_{ijt}$ or $Revenue_{ijt}$, $Time_t$ counts the number of days between the first day of Round 0 and day t, $Treatment_j$ represents whether $store_j$ is a treated store (i.e., $Treatment_j = 1$) or a control store (i.e., $Treatment_j = 0$), $TVControls_{ijt}$ represents all time-variant control variables, including $Price$, $SetPromotion$, $FridgePromotion$, $SetStockout$, $FridgeStockout$, and $NumFaces$, $Seasonality_t$ represents the three seasonality fixed effects, u_i represents the fixed effect for product set i, and ϵ_{ijt} is the random error term. We account for heteroscedasticity (in Equation (4.4.1) and in all other estimation processes throughout our analyses) by using robust standard errors clustered at the product set level.

Table 4.4.2: Estimation results for our A/A test.

	(1) Quantity	(2) Quantity	(3) Revenue	(4) Revenue
$Time \times Treatment$	0.0000	-0.0005	0.0020	0.0014
	(0.0014)	(0.0014)	(0.0026)	(0.0026)
$Time$	0.0056***	0.0065***	0.0098***	0.0115***
	(0.0012)	(0.0012)	(0.0021)	(0.0022)
$TVControls$	No	Yes	No	Yes
Sample size	6,245	6,245	6,245	6,245
Adj. R^2	0.081	0.105	0.078	0.086

Notes: Standard errors are clustered at the product set level and reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

In this specification, a statistically significant β_1 would indicate that the trend in sales is different between the treatment and control groups when using the same planogram, hence failing the A/A test. Table 4.4.2 demonstrates the results from estimating Equation (4.4.1) for both outcome variables with and without control variables. In all models, β_1 is statistically nonsignificant



(a) Quantity (b) Revenue

Figure 4.4.1: Average pre-intervention sales across treatment and control groups.

($p > 0.428$). Figure 4.4.1 plots the average *Quantity* and *Revenue* for both groups during Round 0. Collectively, these results support that the trend in sales between the two groups remains the same when operating under the same planogram, establishing that the control group is comparable to the treatment group.

4.4.2 Pilot Experiment

We next conduct our pilot experiment to justify the choice of a reduced design over a full-factorial design for our main field experiment by empirically testing the validity of the insensitivity assumption. As discussed, this assumption implies that the effect estimated for a vertical location is not sensitive to swapping the product sets in the two other vertical locations.

We examine the insensitivity assumption using both pre-intervention data (i.e., Round 0) and data from the pilot experiment (i.e., Round 1). As illustrated in Figure 4.3.3, relative to Round 0, in the pilot experiment, we keep the vertical location of one product set at each treated store the same and swap the product sets on the other two shelves. For example, in Treatment Store 1, compared with Round 0, we keep the product sets on the eye-level shelves (labeled as B) fixed and swap the product sets on the stretch-level and stoop-level shelves (labeled as A and C). The vertical locations of all product sets in the control stores remain the same in both rounds.

We use only the data from the product sets whose vertical locations are kept constant between Rounds 0 and 1 (i.e., product set B in Treatment Store 1, product set C in Treatment Store 2, product set A in Treatment Store 3, and all product sets in the control stores). After removing 6% of observations based on our stockout filtering rule, we retain 9,822 observations for our analysis. Table 4.4.3 reports the descriptive statistics for the data used in our analysis.

Table 4.4.3: Descriptive statistics for the data sample from our pilot experiment.

	Round 0				Round 1			
	Treated stores		Control stores		Treated stores		Control stores	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1. Quantity	0.29	0.34	0.53	0.38	0.40	0.37	0.55	0.39
2. Revenue	0.66	0.70	1.21	0.77	0.92	0.81	1.26	0.78
3. Price	5.52	4.27	5.28	3.98	5.88	4.22	5.71	3.99
4. SetPromotion	0.69	1.02	0.70	0.87	1.36	1.33	1.46	1.27
5. FridgePromotion	0.50	0.56	0.62	0.73	1.07	1.00	1.21	1.03
6. SetStockout	0.49	0.74	0.80	1.07	0.68	0.89	1.03	0.99
7. FridgeStockout	0.53	0.43	0.89	0.67	0.81	0.55	1.10	0.61
8. NumFaces	5.93	1.42	5.83	1.53	5.87	1.50	5.54	1.55
Number of product sets	49		109		48		123	
Sample Size	9,822							

To test the insensitivity assumption, we measure the change in sales in Round 1 (relative to Round 0) for product sets whose vertical locations are kept fixed in both rounds at the treated stores relative to the control stores. If the insensitivity assumption holds, then we would expect that a change in sales for those targeted shelves at the treated stores (where the product sets in the other two shelves are swapped) would not be statistically different from that at the control stores (where the product sets in the other two shelves are not swapped).

More formally, we use a difference-in-differences (DiD) model. The DiD model estimates double differences between treatment and control groups in two steps: (i) the difference in the outcome variables across time within each group, and (ii) the difference between the two group-specific differences estimated in the first step. This method makes our identification robust to both the omitted variable bias that may arise from any time-invariant unobserved heterogeneity and trend-specific differences between treatment and control groups (Card and Krueger 1994, Cui et al. 2020, Arslan et al. 2022).

The product sets whose vertical locations are kept constant in both rounds are at the eye-level in Treatment Store 1, at the stoop-level in Treatment Store 2, and at the stretch-level in Treatment Store 3. We capture this variation by estimating vertical location-specific DiD estimators. More specifically, to test the insensitivity assumption in each vertical location, we estimate a triple difference (DDD) model. We specify the fixed-effect DDD model as follows:

$$\begin{aligned} Y_{ijt} =& \beta_0 + \beta_1 Post_d \times Treatment_j + \beta_2 Post_t + \beta_3 VerLoc_i \times Post_t \times Treatment_j \\ &+ VerLoc_i \times Post_t + TVControls_{ijt} + Seasonality_t + u_i + \epsilon_{ijt}, \end{aligned} \tag{4.4.2}$$

where $Post_t$ is a binary variable that equals one for time periods in Round 1 and 0 for time periods in Round 0, and $VerLoc_i$ is an indicator variable representing whether the vertical location of product set i is kept fixed in both rounds at the stretch-level, eye-level, or stoop-level. In our estimation, we set the stoop-level to be the reference level for the variable $VerLoc_i$. The coefficients of interest to test the insensitivity assumption are β_1 and β_3.

The fundamental assumption of the DiD model is the parallel trends assumption, which implies that, in the absence of the treatment, the trend in the outcome variable should be the same between the treatment and control groups. In our study, Round 0 corresponds to the period before the intervention. As we have estimated in our A/A test (see Table 4.4.2), the trend in sales is not statistically different between the treatment and control group stores during the preintervention period. Thus, this provides support for the parallel trends assumption for our data.

Table 4.4.4 reports the results from the estimation of Equation (4.4.2) for *Quantity* and *Revenue* with and without control variables. Across all estimation results, (i) the DiD estimator for the stoop-level ($\beta = 0.149$, $p > 0.05$) is not statistically significant and (ii) the DiD estimators for the stretch-level ($\beta = -0.0873$, $p > 0.05$) and the eye-level ($\beta = -0.144$, $p > 0.05$) are not significantly different from the DiD estimator for the stoop-level. Furthermore, we compare the DiD estimators for the stretch-level and the eye-level using a Wald test, finding that they are not statistically different from each other ($F = 1.46$, $p = 0.24$). These results establish that the sales of a product set at a vertical location are not sensitive to swapping the product sets in the other two vertical locations. This provides evidence that the insensitivity assumption is satisfied in our setting. In Appendix A3.2, we conduct an additional robustness test, finding that the insensitivity assumption for a vertical location is also robust to the type of product set that is kept at that location.

In summary, the A/A test and pilot experiment establish that we can conduct our main field experiment with the selected treatment and control group stores using our proposed reduced design to efficiently estimate the vertical location effect. We next present our main experiment.

4.5 Main Experiment: Measuring the Vertical Location Effect

In this section, we first discuss the data and implementation of our main experiment, examining whether customer confusion because of frequent planogram changes exists. We then estimate the vertical location effect for a product set

Table 4.4.4: Estimation results for our pilot experiment.

	(1) Quantity	(2) Quantity	(3) Revenue	(4) Revenue
$Post \times Treatment$	0.149	0.146	0.520	0.512
	(0.0835)	(0.0844)	(0.270)	(0.270)
$VerLoc(StretchLevel) \times Post \times Treatment$	-0.0873	-0.113	-0.413	-0.463
	(0.0907)	(0.0916)	(0.278)	(0.279)
$VerLoc(EyeLevel) \times Post \times Treatment$	-0.144	-0.123	-0.509	-0.487
	(0.100)	(0.100)	(0.294)	(0.295)
$TVControls$	No	Yes	No	Yes
Sample size	9,822	9,822	9,822	9,822
Adj. R^2	0.069	0.092	0.072	0.087

Standard errors clustered at the product set-level are presented in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

and for the overall sales. Next, we explore the heterogeneity of the vertical location effect with respect to several product characteristics.

4.5.1 Data and Implementation

We conducted our main experiment using the reduced design. In particular, after the pilot experiment (i.e., Round 1), we rotated the planograms in the treated stores in three rounds, as illustrated in Figure 4.3.3, while keeping the vertical locations of all product sets in the control stores unchanged.
The data for our main experiment consist of the observations in Round 0 (i.e., pre-intervention period), Round 2, Round 3, and Round 4 (i.e., post-intervention period) from all six stores. After removing 9.02% of observations based on the stockout filtering rule, the final dataset contains 20,291 observations. The descriptive statistics for this dataset are provided in Table 4.5.1.

4.5.2 Customer Confusion

As discussed in Section 4.3, because of the frequent planogram changes, some customers may not immediately adapt to the new planogram and, subsequently, decide not to purchase anything or to purchase a substitute prod-

Table 4.5.1: Descriptive statistics for the data sample from our main experiment.

	Round 2				Round 3				Round 4			
	Treatment		Control		Treatment		Control		Treatment		Control	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1. Quantity	0.42	0.39	0.59	0.40	0.53	0.40	0.67	0.39	0.50	0.41	0.63	0.41
2. Revenue	0.97	0.78	1.38	0.82	1.24	0.83	1.52	0.84	1.18	0.81	1.48	0.84
3. Price	6.04	4.16	5.86	4.00	5.97	4.13	5.43	3.68	6.25	4.30	5.94	4.06
4. SetPromotion	1.36	1.28	1.46	1.21	1.37	1.27	1.48	1.25	1.28	1.23	1.37	1.21
5. FridgePromotion	1.15	1.04	1.23	1.00	1.16	1.06	1.35	1.08	1.09	1.06	1.24	1.06
6. SetStockout	0.74	0.92	0.76	0.99	0.75	0.98	0.64	1.05	0.87	1.01	0.63	0.95
7. FridgeStockout	0.82	0.63	0.82	0.75	0.95	0.71	0.79	0.70	1.08	0.74	0.80	0.75
8. NumFaces	5.64	1.68	5.77	1.76	5.64	1.66	6.12	1.82	5.53	1.73	5.82	1.64
Product sets	142		119		137		108		139		121	
Sample Size	4,649				3,786				5,611			

Notes. Descriptive statistics for Round 0 are provided in Table 1.

uct. After they become familiar with the new planogram and locate their preferred product, they may reinstate their original purchasing behavior. In this case, the initial change in purchasing behavior (i.e., not purchasing anything or purchasing a substitute product) arises from not adapting to the new planogram (i.e., customer confusion) and cannot be attributed to the vertical location change. Hence, any change in sales in our main experiment may commingle both the vertical location effect and potential customer confusion effect.

To identify whether the aforementioned customer confusion is present in our main experiment, we first compare the sales under the business-as-usual planogram (i.e., P1) during the preintervention period (i.e., Round 0) to the sales when the same planogram is reinstated during the postintervention period (i.e., Round 2 for Treatment Store 1, Round 4 for Treatment Store 2, and Round 3 for Treatment Store 3). If the sales under P1 are significantly different between these two periods, the difference can be attributed to customer confusion because the vertical locations of all products remain the same. Our identification relies on comparing the change in sales for the product sets in the postintervention period (relative to the preintervention period) at the treated stores relative to the control stores. To do so, we modify the traditional DiD model, as follows:

- Unlike the traditional DiD setting where there is a single intervention, our main experiment consists of three interventions (i.e., Rounds 2-3-4). We capture these interventions by using three binary variables: (i) $Intervention_t^1$ is equal to 1 for time periods in Round 2 and 0 otherwise, (ii) $Intervention_t^2$ is equal to 1 for time periods in Round 3 and 0 otherwise, and (iii) $Intervention_t^3$ is equal to 1 for time periods in Round 4 and 0 otherwise.

- Another difference is that the treated stores in our experiment are exposed to different treatments (i.e., planograms) across all three interventions (as opposed to the same treatment). For instance, during the first intervention in Round 2, the planograms applied to Treatment Stores 1, 2, and 3 are P1, P2 and P3, respectively. We capture the difference in treatment during the intervention using the categorical variable $Treated_j^l$, where $l \in \{1, 2, 3\}$, which indicates whether store j is Treatment Store 1 ($l = 1$), Treatment Store 2 ($l = 2$), or Treatment Store 3 ($l = 3$). We set the control group as the reference group.

Following these modifications, we specify our fixed-effect DiD model as follows:

$$Y_{ijt} = \beta_0 + \sum_{k \in \{1,2,3\}} \beta_{1k} Intervention_t^k + \sum_{l \in \{1,2,3\}} \sum_{k \in \{1,2,3\}} \beta_{2lk} Treated_t^l \times Intervention_t^k + TVControls_{ijt} + Seasonality_t + u_i + \epsilon_{ijt}. \quad (4.5.1)$$

We present the results from the estimation of Equation (4.5.1) in Table A3.6, which is relegated to the appendix. In Table 4.5.2, we report the summary of our estimation results. First, the change in product set sales between preintervention P1 and postintervention P1 is not statistically significant ($\beta = 0.035$, $p = 0.171$). This implies that there is no statistical evidence for customer confusion in our main experiment. Thus, any significant change in product set sales in our experiment when the planogram is switched from P1 to P2 or from P1 to P3 can be attributed solely to the vertical location effect. Second, product set sales significantly change when the planogram is switched from P1 to P2 ($\beta = 0.043$ $p = 0.09$) and from P1 to P3 ($\beta = 0.057$ $p = 0.02$). This provides evidence for the presence of the vertical location effect. Despite this evidence, it is still not clear how much the relative sales gain/loss is when a product is placed at a specific vertical location compared with other vertical locations. We next rigorously estimate these gains and losses.

Table 4.5.2: Estimation results to examine customer confusion.

Planogram change From	To	Coefficient	Std. err.	p-value
P1	P1	0.0350	0.0255	0.172
P1	P2	0.0430	0.0257	0.094
P1	P3	0.0511	0.0256	0.047
Sample size			20,291	

4.5.3 Vertical Location Effect

To identify the vertical location effect, we extend the DiD model in Equation (4.5.1) to a triple difference model. We do so because, within a store, launching a new planogram implies different interventions for our experimental units (i.e., product sets). For instance, in Treatment Store 1, launching P2 in Round 3 implies that product set A is moved from the stretch-level to eye-level, product set B is moved from the eye-level to stoop-level, and product set C is moved from the stoop-level to stretch-level. To capture this product set–specific variation in interventions, we define the categorical variable $ProductSet_i^m$, which indicates whether product set i is labeled as A ($m = A$) or B ($m = B$) in our study. We set product set C as the reference group. We interact $ProductSet_i^m$ with the DiD estimators in Equation (4.5.1) (i.e., $Treated \times Intervention$) to identify product set–specific treatment effects, making our model a triple difference model, as follows:

$$
\begin{aligned}
Y_{ijt} =& \beta_0 + \sum_{k \in \{1,2,3\}} \beta_{1k} Intervention_t^k + \sum_{l \in \{1,2,3\}} \sum_{k \in \{1,2,3\}} \beta_{2lk} Treated_t^l \times Intervention_t^k \\
&+ \sum_{m \in \{A,B\}} \sum_{k \in \{1,2,3\}} \beta_{3mk} ProductSet_i^m \times Intervention_t^k \\
&+ \sum_{m \in \{A,B\}} \sum_{l \in \{1,2,3\}} \sum_{k \in \{1,2,3\}} \beta_{4mlk} ProductSet_i^m \times Treated_t^l \times Intervention_t^k \\
&+ TVControls_{ijt} + Seasonality_t + u_i + \epsilon_{ijt}.
\end{aligned} \tag{4.5.2}
$$

To provide intuition, let α_1, α_2, and α_3 denote the coefficients of $Treated^2 \times Intervention^1$, $ProductSet^A \times Treated^2 \times Intervention^1$, and $ProductSet^B \times Treated^2 \times Intervention^1$, respectively. These coefficients capture the change in sales at Treatment Store 2 between Round 2 and Round 0. In particular, (i) α_1 represents the effect of bringing product set C from the stoop-level to stretch-level on product set C sales, (ii) $\alpha_1 + \alpha_2$ represents the effect of bringing the product set A from the stretch-level to eye-level on product set A sales, and (iii) $\alpha_1 + \alpha_3$ represents the effect of bringing the product set B from the eye-level to stoop-level on product set B sales.

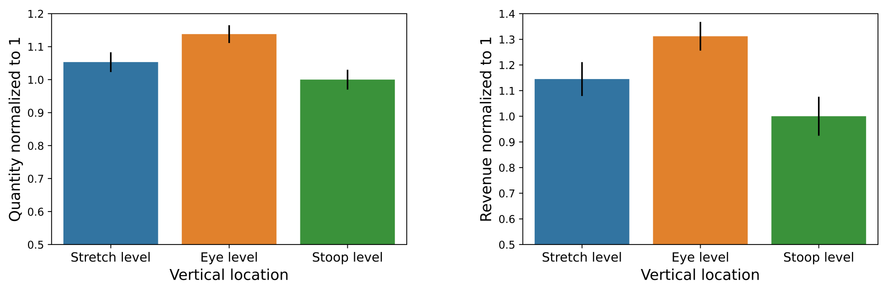
We present the results from estimating the model in Equation (4.5.2) in Table A3.7 in the appendix. We summarize the key results in Table 4.5.3 and draw several observations from these results. First, we find that the vertical location change significantly affects the quantity sold for a product set, yet this effect is not necessarily symmetric. For instance, moving a product set from the stretch-level to the eye-level increases the sales by 9.2% ($\beta = 0.092$, $p = 0.012$), whereas the opposite move (i.e., from the eye-level to stretch-level) decreases the sales by only 6.9% ($\beta = -0.069$, $p = 0.089$). Similarly, moving a product set from the eye-level to the stoop-level decreases the sales by 10% ($\beta = -0.100$, $p = 0.009$), whereas the opposite move (i.e., from the stoop-level to eye-level) increases the sales by 19.6% ($\beta = 0.196$, $p = 0.000$). Finally, moving a product set from the stoop-level to stretch-level increases

Table 4.5.3: Estimation results for the vertical location effect.

Planogram change From	Planogram change To	Vertical location change From	Vertical location change To	Quantity	Revenue
P1	P2	Stretch-level	Eye-level	0.092^{*} (0.036)	0.212^{**} (0.060)
		Eye-level	Stoop-level	-0.100^{**} (0.038)	-0.165^{+} (0.088)
		Stoop-level	Stretch-level	0.128^{**} (0.049)	0.357^{**} (0.116)
P1	P3	Stretch-level	Stoop-level	0.025 (0.036)	0.085 (0.058)
		Eye-level	Stretch-level	-0.069^{+} (0.041)	-0.111 (0.094)
		Stoop-level	Eye-level	0.196^{***} (0.047)	0.583^{***} (0.126)
Sample size				20,291	

Standard errors are clustered at the product set-level and presented in parentheses. $^{+}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

the sales by 12.8% ($\beta = 0.128$, $p = 0.009$), whereas the opposite move (i.e., from the stretch-level to stoop-level) does not change the sales ($\beta = 0.025$, $p > 0.1$). These asymmetric effects arise because the reorganization of the vertical locations of the other product sets when moving a product set from one shelf to another is different than that when moving a product set in the opposite direction. For instance, when moving a product set from the stretch-level to eye-level (while changing the planogram from P1 to P2), the product set in the stoop-level is moved to the stretch-level. However, when moving a product set from the eye-level to stretch-level (while changing the planogram from P1 to P3), the product set in the stoop-level is moved to the eye-level. This indicates that the effect of the vertical location change depends on how other products are reorganized as a result of that change, signifying the importance of the interplay among products whose vertical locations change simultaneously. Using the estimated coefficients in Table 4.5.3, we obtain the average vertical location effects for *Quantity* and illustrate the normalized results in Figure 4.5.1.a. On average, we find that displaying the products at the stoop-level results in the lowest sales. Compared with the stoop-level, the sales are 5.3% higher when the products are placed at the stretch-level and 13.8% higher when they are placed at the eye-level.



(a) Quantity (b) Revenue

Figure 4.5.1: Normalized vertical location effect on sales.
Note: All the errors bars are at 95% confidence intervals.

Second, as can be inferred from Table 4.5.3, the vertical location change effect is amplified when considering revenue sales. Figure 4.5.1.b illustrates the corresponding normalized vertical location effects for *Revenue*. Here, compared with the stoop-level, on average, the dollar sales are 14.6% higher when the products are placed at the stretch-level and 31.2% higher when products are placed at the eye-level. This suggests that the vertical location effect in dollar sales likely depends on both the change in demand and product prices.

Overall, these results provide the first rigorous empirical evidence from a field experiment for the causal effect of vertical location on sales, supporting that "eye-level is buy level." Our results also demonstrate that the eye-level effect potentially comes at the expense of a sales loss for products placed at the stretch-level and/or stoop-level. These results lead to the following question: How are the overall sales in a fridge affected when the vertical locations of three shelves change simultaneously? We proceed to answer this question.

4.5.4 Vertical Location Effect on Overall Sales

In this section, we investigate whether the gain in sales at the eye-level compensates for the loss in sales at the stretch- or stoop-levels. More formally, we examine whether retailers can increase their sales across all treated products with a planogram change or if such a change simply results in pure substitution among the treated products. For this analysis, we change the experimental unit from a product set to a fridge, subsequently aggregating the data from the product set-level to the fridge-level. We define Y_{ijt} as the outcome variable (i.e., *Quantity* or *Revenue*) representing the sales of all product sets on all three shelves (i.e., stretch-, eye-, and stoop-levels) in fridge i in store j at time t. Using this aggregate outcome variable, we estimate the DiD model specified in Equation (4.5.1), presenting the results in Table A3.8 in the appendix.

Table 4.5.4 summarizes the key results from the estimated model. Here, changing planograms from P1 to P2 or from P1 to P3 does not have a significant effect on the overall quantity sold in a fridge. Hence, the sales increase at the eye-level comes at the expense of a sales loss at the stretch- and stoop-levels, suggesting that the planogram change results in a pure substitution pattern among products. Combined with the *Quantity* results in Table 4.5.3, these results demonstrate that although the market share of individual products can change depending on the vertical location, for retailers, the overall sales across all products remain the same, regardless of the planogram used. To our knowledge, this finding provides the first empirical evidence for the literature assuming pure substitution in the theoretical modeling of planogram optimization problems (e.g., Hansen et al. 2010, Gencosman and Begen 2022).

The results regarding the overall revenue are noteworthy. Here as with the quantity, changing planograms from P1 to P2 does not have a significant effect on the overall revenue in a fridge. However, changing planograms from P1 to P3 results in a significant increase in revenue from all treated products

Table 4.5.4: Estimation results for the overall sales.

Planogram Change From	To	Quantity	Revenue
P1	P1	0.036 (0.029)	0.116 (0.079)
P1	P2	0.034 (0.028)	0.094 (0.076)
P1	P3	0.043 (0.020)	0.158† (0.080)
Sample size (N)		7,753	

Standard errors are clustered at the fridge-level and presented in parentheses.
† $p < 0.10$, $^{*}p < 0.05$.

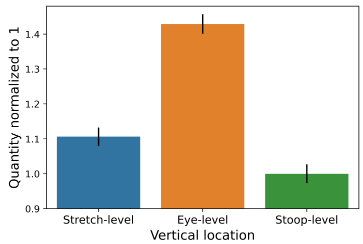
in a fridge by 15.8% ($\beta = 0.158$, $p = 0.05$). In summary, the analysis on the overall sales demonstrates that, although the overall sales for the retailer remain the same across different planograms, the overall revenue can vary depending on the planogram used. This arises because of two factors. First, as shown in Table 4.5.3, the effect of vertical location change on sales depends on both the specific level at which the product is placed and on how the other products are reorganized as a result of that change. Second, the products placed at different levels have different price points. Consequently, the overall effect of a vertical location change on the entire shelving unit's revenue depends not only on which products are moved to the eye-level, but also on which products are simultaneously moved to the stretch- and stoop-levels. In the next subsection, we proceed to characterize the type of products better suited for the eye-level shelves versus stretch- and stoop-level shelves.

4.5.5 Heterogeneous Vertical Location Effect

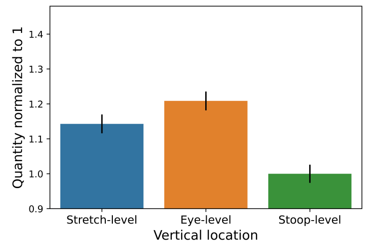
In the previous section, we estimated the vertical location effect by averaging all the products in a set. In this section, using a product-level analysis, we examine the heterogeneity of the vertical location effect with respect to several product characteristics, including price, promotions, broken assortment, market concentration, number of substitute products, package size, and flashiness. We provide the technical details behind this analysis in Appendix A3.4. Here, we summarize the key insights of this analysis and interpret the results.[5]

First, as illustrated in Figure 4.5.2, product price positively moderates the vertical location effect. This suggests that vertical location is more important for high-priced than low-priced items. We conjecture that customers' price sensitivity is a likely explanation for this result. Price-sensitive customers may have a low search cost and typically would purchase low-priced items, whereas price-insensitive customers would be more likely to have a high search cost and typically purchase high-priced items. Thus, price-sensitive customers are likely to spend more time than their price-insensitive counterparts to locate and purchase their preferred products (particularly when those items are placed on shelves far from the eye-level), making low-price products less sensitive and high-priced products more sensitive to the vertical location effect.

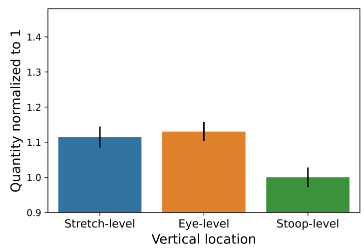
[5]For brevity, we present the results for *Quantity*. The results for *Revenue* are qualitatively similar, yet the magnitude of the heterogeneity is more pronounced, making the results presented here more conservative.



(a) High price



(b) Medium price

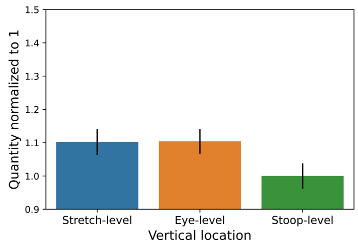


(c) Low price

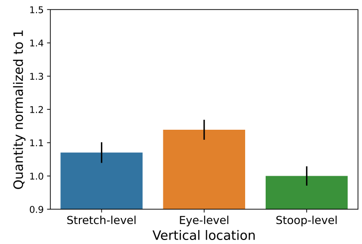
Figure 4.5.2: Heterogeneity of the vertical location effect with respect to price.

Second, regarding promotion, as shown in Figure 4.5.3, the vertical location effect diminishes as the monetary value of the discount offered for a product increases. This finding aligns with our earlier result on price and supports our conjecture on customer price sensitivity because a higher discount implies a lower price and vice versa.

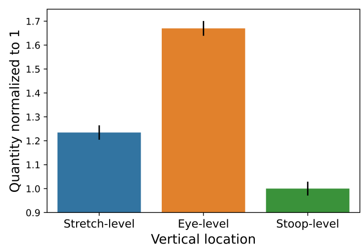
Previous research has demonstrated that broken assortments (because of stockouts), which are often attributed to low supplier replenishment performance (Grant and Fernie 2008), can have a negative impact on sales performance. Certain products can be characterized as chronically exposed to broken assortments if other products provided by suppliers with a low replenishment performance face frequent stockouts. Thus, as the third step, we investigate the moderating role of a broken assortment on the vertical location effect. As illustrated in Figure 4.5.4, in contrast to our main results,



(a) High promotion



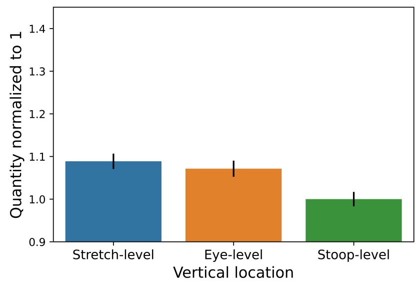
(b) Medium promotion



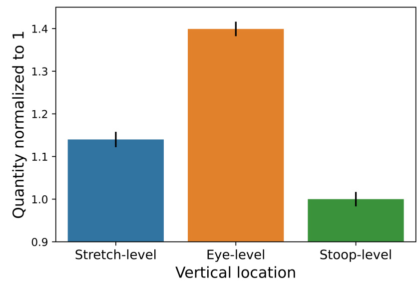
(c) Low promotion

Figure 4.5.3: Heterogeneity of the vertical location effect with respect to promotion discount.

we find that the vertical location effect is no longer statistically significant when a product is exposed to a broken assortment. The vertical location effect is present only when the assortment is complete. This finding further supports our conjecture on search cost. As the number of missing items in a planogram increases, the search cost decreases (Mantrala et al. 2009) and the likelihood of finding the preferred item (particularly those placed at shelves far from the eye-level) increases, thus attenuating the vertical location effect.



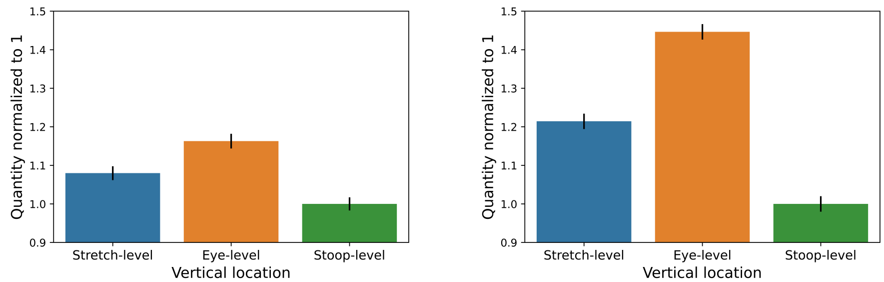
(a) Broken assortment



(b) Complete assortment

Figure 4.5.4: Heterogeneity of the vertical location effect with respect to a broken assortment.

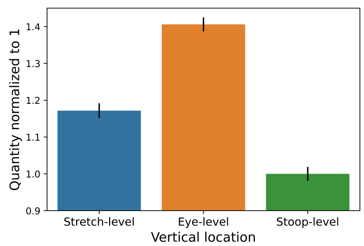
Fourth, we examine the heterogeneity of the vertical location effect regarding market concentration. Market concentration is defined as the degree to which a market is dominated by a small number of brands (Rhoades 1993). We operationalize market concentration by using the data from three months prior to the experiment date. In particular, following Rhoades (1993), we calculate the Herfindahl–Hirschman Index (HHI) to characterize the market concentration for each product category. As shown in Figure 4.5.5, we find that the vertical location effect is more pronounced for products in categories with a low market concentration. A market with a low concentration is not dominated by large players, so it is considered to be competitive (Karuna 2007); this often arises when customers do not have strong preferences among alternative options (Chaudhuri 1999). In these cases, the products placed at the eye-level are more likely to be purchased because exploring other shelves for alternative products for which customers are likely indifferent creates an additional search cost, hence amplifying the vertical location effect.



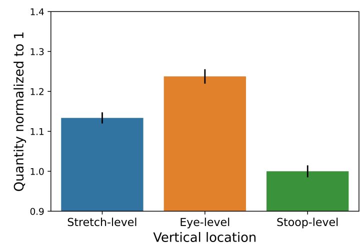
(a) High market concentration (b) Low market concentration

Figure 4.5.5: Heterogeneity of the vertical location effect with respect to market concentration.

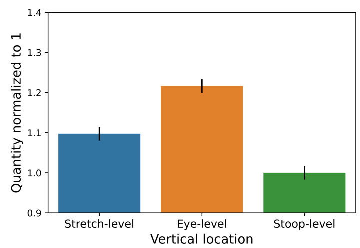
Fifth, we assess how the vertical location effect changes with the number of substitute products. For a given product, we determine its substitutable products based on a set of product attributes including brand, size, flavor, alcohol level, diet/zero, and weight. Using these attributes, we calculate the Gower's distance between each pair of products in the same category and normalize the distance value to between 0 and 1. We then consider a product to be a substitute for another product if the distance value falls below a certain threshold. Specifically, we evaluate three threshold values (0.1, 0.15, and 0.2). Because our findings remain consistent across all three values, we present the results when using a threshold of 0.1. As shown in Figure 4.5.6, the vertical location effect becomes more pronounced as the number of substitute products increases. This finding is also consistent with our previous conjecture on search cost. Indeed, it is known that the search cost increases with the number of substitute products (Chernev and Hamilton 2009). Finally, as detailed in Appendix A3.4, we assess whether the vertical location effect is moderated by the package size and the flashiness level, finding no evidence for such moderation.



(a) High number of substitutes



(b) Medium number of substitutes



(c) Low number of substitutes

Figure 4.5.6: Heterogeneity of the vertical location effect with respect to the number of substitute products.

In summary, our analysis shows that the vertical location effect (i) is amplified with the price and the number of substitute products and (ii) is attenuated with promotional discounts, stockouts, and market concentration. Our results suggest that retailers can enhance their revenue by displaying products with high prices, with low discounts, from a complete assortment, in a category with a low market concentration and with a large number of substitute products at the eye-level.

4.6 Economic Implications

Our study establishes that the vertical location effect depends on how the products on the other shelves are reorganized simultaneously. Combined with the difference in profit margins, this suggests that retailers can earn higher profits by optimizing their planograms holistically while considering the interplay among products at all vertical location levels rather than myopically focusing on the eye-level effect. In addition, the vertical location effect is heterogeneous across products, which, if incorporated into the planogram optimization, can further enhance the profitability of the holistic approach. To assess the economic significance of our empirical results, we conduct a counterfactual analysis in two stages.

First, we (mistakenly) assume that the vertical location effect is independent of the product characteristics. Using this homogeneous vertical location effect, we optimize the planogram on each fridge and calculate the potential profit change compared with the business-as-usual practice at the focal retailer. Second, we solve the same planogram optimization by allowing the vertical location effect to be a function of the product characteristics and calculate the additional gain in profit from incorporating the heterogeneity in the vertical location effect into the planogram optimization.

For both counterfactual studies, we consider the following linear integer programming model that serves as the foundational base for several planogram optimization problems in the literature (e.g., Yang and Chen 1999, Lim et al. 2004, Bianchi-Aguiar et al. 2018, Bianchi-Aguiar et al. 2021):

$$\begin{aligned} Max \quad & \sum_{i=1,...,N} \sum_{k=1,2,3} d_i \times p_i \times a_k \times y_{ik} \\ s.t. \quad & \sum_{k=1,2,3} y_{ik} = 1 \quad \forall i \in \{1,...,N\} \\ & \sum_{i=1,...,N} m_i \times y_{ik} \leq C_k \quad \forall k \in \{1,2,3\} \\ & y_{ik} \in \{0,1\}, \end{aligned} \tag{4.6.1}$$

where (i) d_i is the base demand/sales of product i when product i is placed at the stoop-level (i.e., the vertical location that results in the lowest sales), (ii) p_i is the profit margin of product i, (iii) k represents the vertical location where 1 denotes the stretch-level, 2 denotes the eye-level, and 3 denotes the stoop-level, (iv) a_k is the multiplier for the sales change when a product is located in vertical location k (i.e., the vertical location effect on sales quantity), where a_3 is normalized to 1, (v) m_i is the number of facings for product i, and (vi) C_k is the capacity of shelf k in terms of number of facings. The decision variable, y_{ik}, is a binary variable indicating whether product i is displayed on shelf k. The optimization formulation in (4.6.1) can be used to maximize the profit in a fridge, which is subject to two constraints: (i) each product must be placed in only one vertical location, and (ii) the number of products on a shelf must not exceed the shelf capacity. Here, the planogram optimization model in Equation (4.6.1) assumes that the vertical location effect, a_k, is homogeneous (i.e., independent of the products). To obtain a model that considers a heterogeneous vertical location effect, we replace a_k in Equation (4.6.1) with a_{ik}, where a_{ik} represents the vertical location effect for product i when placed in vertical location k and a_{i3} is normalized to 1.

In our counterfactual analysis, we set m_i and C_k to values that match the practice at the focal retailer. This enables us to quantify the impact of changing only the vertical locations of products on profitability. We obtain data on profit margin (i.e., p_i) for each product. To operationalize a_k, using post-intervention data from the treated stores, we estimate the following regression model:

$$Quantity_{ijt} = \beta_0 + \beta_1 Stretch_{ijt} + \beta_2 Eye_{ijt} + TVControls_{ijt} + Seasonality_t + u_i + \epsilon_{ijt}, \tag{4.6.2}$$

where $Stretch_{ijt}$ and Eye_{ijt} are binary variables indicating whether product i in store j on day t is placed on the stretch-level and eye-level, respectively (with the stoop-level being the reference). Using the estimated model in Equation (4.6.2), we can operationalize a_1 as $1 + \hat{\beta}_1$ and a_2 as $1 + \hat{\beta}_2$.

To operationalize d_i and a_{ik}, using the same data, we estimate the following random-slope model, in which the vertical location effect varies across

Table 4.6.1: Estimation results for linear and random-slope models.

	Linear model	Random-slope model
Stretch	0.0539***	0.0589***
	(5.83)	(6.06)
Eye	0.116***	0.0996***
	(11.53)	(9.76)
TVControls	Yes	Yes
$\sigma_{\zeta_{0i}}$	-	0.111
$\sigma_{\zeta_{1i}}$	-	0.126
$\sigma_{\zeta_{2i}}$	-	0.312
Wald χ^2	-	839.99***
LR test χ^2	-	9,431.79***
Sample size (N)	36,876	36,876

Standard errors are clustered at the product level and presented in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

products:

$$\begin{aligned} Quantity_{ijt} =& \beta_0 + \beta_1 Stretch_{ijt} + \beta_2 Eye_{ijt} + TVControls_{ijt} + Seasonality_t \\ &+ \zeta_{0i} + \zeta_{1i} Stretch_{ijt} + \zeta_{2i} Eye_{ijt} + u_i + \epsilon_{ijt}. \end{aligned} \tag{4.6.3}$$

In Equation (4.6.3), (i) ζ_{0i} denotes the deviation of product i's intercept from the mean intercept β_0, (ii) ζ_{1i} denotes the deviation of product i's slope for the stretch-level effect from the mean slope β_1 for the stretch-level effect, and (iii) ζ_{2i} denotes the deviation of product i's slope for the eye-level effect from the mean slope β_2. Thus, for product i, the eye-level effect and the stretch-level effect (compared with the stoop-level effect) are estimated to be $\beta_2 + \zeta_{2i}$ and $\beta_1 + \zeta_{1i}$, respectively. Using the estimated model in Equation (4.6.3), we operationalize (i) a_{i1} as $1 + \hat{\beta}_1 + \hat{\zeta_{1i}}$, (ii) a_2 as $1 + \hat{\beta}_2 + \hat{\zeta_{2i}}$, and (iii) d_i as the average of the predicted outcome variable (obtained after the vertical location is set to the stoop-level) across all observations of product i.

Table 4.6.1 shows the estimation results of the models specified in Equations (4.6.2) and (4.6.3). Consistent with our heterogeneity analysis, the significant log-likelihood ratio test statistic (i.e., the LR test χ^2) in the random-slope model indicates that the vertical location effect is indeed product specific, implying that the random-slope model specification is the correct specification. In addition, the estimated eye-level and stretch-level effects in both models are similar to those obtained using the DiD model in Section 4.5.3, thus providing a sanity check for our counterfactual analysis and serving as another robustness test for the DiD specification.

For our counterfactual analysis, we consider three scenarios. The first scenario represents the current practice. For this, we recode the vertical location variables $Stretch_{ijt}$ and Eye_{ijt} to represent the business-as-usual planogram (i.e., planogram $P1$ in Figure 4.3.3). Using the estimated random-slope model from Equation (4.6.3) along with these recoded data, we make an out-of-sample prediction to obtain the predicted sales ($Quantity_{ijt}$). Multiplying the predicted sales with the profit margin p_i results in the estimated profit for each observation. The retailer's estimated overall profit under the current practice is then equal to the summation of the estimated profit across all observations.

The second scenario represents the counterfactual case in which the retailer optimizes its planograms using the homogeneous vertical location effects. For this scenario, using a_k estimated in the linear model from Equation (4.6.2), we solve the planogram optimization model in Equation (4.6.1) and obtain the values of y_{ik} (i.e., the vertical location decision for each product in a fridge). We recode $Stretch_{ijt}$ and Eye_{ijt} in the data to represent the values of y_{ik}. We then repeat the same out-of-sample prediction steps as in the first scenario to obtain the estimated overall profit (i.e., the counterfactual profit in the second scenario). Finally, the third scenario represents the counterfactual case in which the retailer optimizes its planograms while using the heterogeneous vertical location effects. The steps to obtain the counterfactual profit in this scenario are similar to those in the second scenario, with the exception of solving the planogram optimization model using a_{ik} in Equation (4.6.1) (instead of a_k), as estimated in the random-slope model.

We quantify the counterfactual profit improvement by comparing the estimated profits across the three scenarios. We find that, relative to the current practice, the profit could have been 2.2% higher if the retailer had optimized its planograms using homogeneous vertical location effects. Given the notoriously low single-digit profit margins in the retail industry, this result suggests that considering the interplay among products (arising from simultaneously changing vertical locations) significantly increases profit. In addition, the profit could have been 3% higher if the retailer had optimized its planograms using heterogeneous vertical location effects. These results testify to the economic and practical importance of the empirical findings in the present paper.

4.7 Conclusion

The belief that "eye-level is buy level" forms the basis for planogram optimization for retailers and contractual agreements between retailers and manufacturers. Although several studies have investigated the effect of eye-level placements on customer attention and sales, it is not clear (i) how this effect influences the products placed on other shelves, and thus, the overall sales, and (ii) whether that effect varies across products on the same shelving unit. To address these gaps, we conducted a novel two-stage field experiment in collaboration with a North American convenience store chain. Our study contributes to the literature by conducting the first field experiment to identify the interplay among a set of products for which the vertical locations change simultaneously.

Theoretically, our study contributes by providing evidence that the effect of changing the vertical location of a product depends not only on the level at which that product is moved, but also on how the vertical locations of the other products change as a result of that move. Subsequently, we find that the same vertical location change, which can be achieved by two different planograms, can result in two different sales outcomes if the vertical locations of the other products change differently between these two planograms. When we average these asymmetric vertical location change effects, we find that the stoop-level placement results in the lowest sales, whereas the eye-level placement results in the highest sales. Second, our study contributes by demonstrating that the incremental sales for products displayed at the eye-level just compensates for the sales loss for products displayed at the stretch- and stoop-levels, leading to no change in the overall sales. This implies that changing vertical locations through a change in the planogram results in a pure substitution pattern among products. As a third theoretical contribution, our study demonstrates that the vertical location effect is heterogeneous across products in the same shelving unit. Because products have different margins and prices, this implies that, despite the pure substitution among products, retailers can still improve their overall profitability by carefully selecting the products to display at the eye-level versus at other levels. Our counterfactual analysis shows that optimizing planograms by considering the average interplay among products (arising from simultaneous vertical changes) increases profit by 2.2%. Our results indicate that products with high prices, and low discounts, that are displayed in a complete assortment, in a category with low market concentration, and with a large number of substitute products benefit more from being displayed at the eye-level. Incorporating this heterogeneity into planogram optimization can improve the performance of planogram optimization by 36.4% (i.e., from 2.2% to 3%) for the focal retailer.

There are several implications of our research for managers. First, retailers should carefully assess the heterogeneity in the vertical location effect across products and optimize their planograms accordingly. Retailers can identify such heterogeneity by replicating our two-stage experimental design within a reasonable time frame and budget. Second, our study can be used to improve the contractual agreements for slotting fees between retailers and manufacturers. The existing contracts in practice often arise from an unbalanced negotiation because determining the optimal value for such fees is challenging due to the information asymmetry and misaligned incentives (Caro et al. 2020, Rao and Mahi 2003, Cachon and Kök 2010). Our results on heterogeneous vertical location effects can enable retailers and manufacturers to overcome this challenge. In particular, with a clear understanding of the product characteristics associated with greater eye-level effects (e.g., high price, low discount, low stockout), retailers and manufacturers can have better-informed negotiations. For instance, if a contractual agreement necessitates the retailer to display at the eye-level a product that is unlikely to benefit from such a location, thus requiring the retailer to deviate from the optimal vertical location placement, the retailer could negotiate a slotting fee that is high enough to at least mitigate the loss because of this suboptimal placement. Likewise, a manufacturer operating in a competitive category can be more determined and willing to pay a high slotting fee to reserve the eye-level location because the additional sales at the eye-level will likely offset the slotting fee, whereas a manufacturer selling a niche product (that is less exposed to competition) may not necessarily be interested in securing the eye-level location. Third, our results suggest that retailers and manufacturers should consider the vertical location effect when making promotion decisions because high promotions diminish the vertical location effect. For instance, reserving the eye-level location may not go hand in hand with a frequent promotion strategy for manufacturers. Finally, procurement teams can leverage our results in the context of their sourcing and category management strategies. In particular, given the scarcity of available eye-level locations in a physical store, retail managers can be more inclined to reserve such locations for manufacturers with a predictable and consistent pricing strategy (leading to no promotions), a reliable replenishment capability (leading to a complete assortment), and a high product variety (leading to more competition).

Bibliography

Alexander, D. (2003). High-stakes shelf games, *Chicago Tribune, December* .
URL: *https://www.chicagotribune.com/news/ct-xpm-2003-12-14-0312140364-story.html*

Arslan, H. A., Tereyağoğlu, N. and Yılmaz, Ö. (2022). Scoring a touchdown with variable pricing: Evidence from a quasi-experiment in the nfl ticket markets. Management Science.

Atalay, A. S., Bodur, H. O. and Rasolofoarison, D. (2012). Shining in the center: Central gaze cascade effect on product choice, *Journal of Consumer Research* **39**(4): 848–866.

Atan, Z., Honhon, D. and Pan, X. A. (2023). Displaying and discounting perishables: Impact on retail profits and waste. Available at SSRN 4369956.

Ausick, P. (2017). A map of how wal-mart lays out its stores to lift sales, *Yahoo Finance* .
URL: *https://finance.yahoo.com/news/map-wal-mart-lays-stores-185011900.html*

Bandi, N., Cohen, M. C. and Ray, S. (2022). Incentivizing healthy food choices using add-on bundling: A field experiment. Available at SSRN.

Bianchi-Aguiar, T., Hübner, A., Carravilla, M. A. and Oliveira, J. F. (2021). Retail shelf space planning problems: A comprehensive review and classification framework, *European Journal of Operational Research* **289**(1): 1–16.

Bianchi-Aguiar, T., Silva, E., Guimarães, L., Carravilla, M. A. and Oliveira, J. F. (2018). Allocating products on shelves under merchandising rules: Multi-level product families with display directions, *Omega* **76**: 47–62.

Bianchi-Aguiar, T., Silva, E., Guimarães, L., Carravilla, M. A., Oliveira, J. F., Amaral, J. G., Liz, J. and Lapela, S. (2016). Using analytics to enhance a food retailer's shelf-space management, *Interfaces* **46**(5): 424–444.

Cachon, G. P. and Kök, A. G. (2010). Competing manufacturers in a retail supply chain: On contractual form and coordination, *Management science* **56**(3): 571–589.

Card, D. and Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania, *The American Economic Review* **84**(4): 772–793.

Caro, F., Kök, A. G. and Martínez-de Albéniz, V. (2020). The future of retail operations, *Manufacturing & Service Operations Management* **22**(1): 47–58.

Chandon, P., Hutchinson, J. W., Bradlow, E. T. and Young, S. H. (2009). Does in-store marketing work? effects of the number and position of shelf facings on brand attention and evaluation at the point of purchase, *Journal of Marketing* **73**(6): 1–17.

Chaudhuri, A. (1999). Does brand loyalty mediate brand equity outcomes?, *Journal of Marketing Theory and Practice* **7**(2): 136–146.

Chelstad, E. (2018). Know your slotting fees, *Observan* .
URL: *https://www.observanow.com/know-slotting-fees/*

Chen, M., Burke, R. R., Hui, S. K. and Leykin, A. (2021). Understanding lateral and vertical biases in consumer attention: an in-store ambulatory eye-tracking study, *Journal of Marketing Research* **58**(6): 1120–1141.

Chernev, A. and Hamilton, R. (2009). Assortment size and option attractiveness in consumer choice among retailers, *Journal of Marketing Research* **46**(3): 410–420.

Cohen, M. C., Jacquillat, A., Serpa, J. C. and Benborhoum, M. (2023). Managing airfares under competition: Insights from a field experiment. Management Science.

Cui, R., Lu, Z., Sun, T. and Golden, J. (2020). Sooner or later? promising delivery speed in online retail. Working paper.

Dreze, X., Hoch, S. J. and Purk, M. E. (1994). Shelf management and space elasticity, *Journal of Retailing* **70**(4): 301–326.

Ebster, C. (2011). *Store design and visual merchandising: Creating store space that encourages buying*, Business Expert Press.

Federal Trade Commission (2001). Report on the federal trade commission workshop on slotting allowances and other marketing practices in the grocery industry.

Frank, R. E. and Massy, W. F. (1970). Shelf position and space effects on sales, *Journal of Marketing Research* **7**(1): 59–66.

Gallino, S. and Moreno, A. (2018). The value of fit information in online retail: Evidence from a randomized field experiment, *Manufacturing & Service Operations Management* **20**(4): 767–787.

Gencosman, B. C. and Begen, M. A. (2022). Exact optimization and decomposition approaches for shelf space allocation, *European Journal of Operational Research* **299**(2): 432–447.

Grant, D. B. and Fernie, J. (2008). Research note: Exploring out-of-stock and on-shelf availability in non-grocery, high street retailing. International Journal of Retail & Distribution Management.

Hansen, J. M., Raut, S. and Swami, S. (2010). Retail shelf allocation: A comparative analysis of heuristic and meta-heuristic approaches, *Journal of Retailing* **86**(1): 94–105.

Hübner, A., Düsterhöft, T. and Ostermeier, M. (2021). Shelf space dimensioning and product allocation in retail stores, *European Journal of Operational Research* **292**(1): 155–171.

Karuna, C. (2007). Industry product market competition and managerial incentives, *Journal of Accounting and Economics* **43**(2-3): 275–297.

Lim, A., Rodrigues, B. and Zhang, X. (2004). Metaheuristics with local search techniques for retail shelf-space optimization, *Management Science* **50**(1): 117–131.

Logie, J. (2022). Grocery store layout: Marketing tricks & strategy that lead you to spend more, *Regained Wellness* .
URL: *https://www.regainedwellness.com/supermarket-psychology/*

Lopez Mateos, D., Cohen, M. C. and Pyron, N. (2022). Field experiments for testing revenue strategies in the hospitality industry, *Cornell Hospitality Quarterly* **63**(2): 247–256.

Mantrala, M. K., Levy, M., Kahn, B. E., Fox, E. J., Gaidarev, P., Dankworth, B. and Shah, D. (2009). Why is assortment planning so difficult for retailers? a framework and research agenda, *Journal of Retailing* **85**(1): 71–83.

Meyersohn, N. (2022). Why stores always put candy and soda near the cash register. CNN Business.

Parker, J. R. and Koschmann, A. R. (2018). Shelf layout and consumer preferences, *Handbook of Research on Retailing*, Edward Elgar Publishing, pp. 251–270.

Rao, A. R. and Mahi, H. (2003). The price of launching a new product: Empirical evidence on factors affecting the relative magnitude of slotting allowances, *Marketing Science* **22**(2): 246–268.

Rhoades, S. A. (1993). The herfindahl-hirschman index, *Federal Reserve Bulletin* **79**: 188.

Rivlin, G. (2016). *Rigged: Supermarket shelves for sale*, Center for Science in the Public Interest.

Russell, R. A. and Urban, T. L. (2010). The location and allocation of products and product families on retail shelves, *Annals of Operations Research* **179**(1): 131–147.

Smirnov, D. and Huchzermeier, A. (2019). Shelf-space management under stockout-based substitution and merchandising constraints. Available at SSRN 3413256.

Sudhir, K. and Rao, V. R. (2006). Do slotting allowances enhance efficiency or hinder competition?, *Journal of Marketing Research* **43**(2): 137–155.

Touche, D. ., on New Product Introductions, J. I. T. F. and Grocery Manufacturers of America, I. (1990). *Managing the Process of Introducing and Deleting Products in the Grocery and Drug Industry.*

Valenzuela, A. and Raghubir, P. (2015). Are consumers aware of top–bottom but not of left–right inferences? implications for shelf space positions., *Journal of Experimental Psychology: Applied* **21**(3): 224.

Valenzuela, A., Raghubir, P. and Mitakakis, C. (2013). Shelf space schemas: Myth or reality?, *Journal of Business Research* **66**(7): 881–888.

Van Nierop, E., Fok, D. and Franses, P. H. (2008). Interaction between shelf layout and marketing effectiveness and its impact on optimizing shelf arrangements, *Marketing Science* **27**(6): 1065–1082.

Yang, M.-H. and Chen, W.-C. (1999). A study on shelf space allocation and management, *International Journal of Production Economics* **60**: 309–317.

A3 Appendix to Article 3

A3.1 Correlation Matrices for Experiments

Table A3.1: Correlation matrix for the data used in A/A test.

	1	2	3	4	5	6	7	8
1. Quantity	1							
2. Revenue	0.87	1						
3. Price	-0.27	0.03	1					
4. SetPromotion	0.16	0.04	-0.39	1				
5. FridgePromotion	0.03	-0.05	-0.28	0.51	1			
6. SetStockout	0.13	0.14	0.03	0.03	-0.00	1		
7. FridgeStockout	0.15	0.19	0.01	-0.02	0.02	0.53	1	
8. NumFaces	-0.02	-0.11	-0.34	0.14	0.01	-0.41	-0.24	1
Sample Size				6,245				

Table A3.2: Correlation matrix for the data used in the pilot experiment.

	1	2	3	4	5	6	7	8
1. Quantity	1							
2. Revenue	0.86	1						
3. Price	-0.28	0.08	1					
4. SetPromotion	0.15	0.01	-0.40	1				
5. FridgePromotion	0.07	-0.04	-0.28	0.65	1			
6. SetStockout	0.09	0.10	0.04	0.03	-0.02	1		
7. FridgeStockout	0.10	0.17	0.13	-0.00	0.03	0.53	1	
8. NumFaces	-0.02	-0.16	-0.37	0.14	0.04	-0.44	-0.33	1
Sample Size				9,822				

Table A3.3: Correlation matrix for the data used in the main experiment.

	1	2	3	4	5	6	7	8
1. Quantity	1							
2. Revenue	0.87	1						
3. Price	-0.32	-0.00	1					
4. SetPromotion	0.22	0.09	-0.44	1				
5. FridgePromotion	0.13	0.04	-0.32	0.62	1			
6. SetStockout	0.14	0.12	-0.01	0.02	-0.02	1		
7. FridgeStockout	0.17	0.16	-0.03	0.03	0.02	0.52	1	
8. NumFaces	-0.03	-0.11	-0.32	0.12	0.05	-0.50	-0.35	1
Sample Size				20,291				

A3.2 Robustness Check for Testing the Insensitivity Assumption in the Pilot Experiment

In the pilot experiment, we test the insensitivity assumption by comparing the sales for product sets whose vertical locations are kept constant between Round 0 and Round 1. One can say that across the three treatment stores, the product sets whose vertical locations are kept constant between Round 0 and Round 1 are not only different products (i.e., product set B in Treatment Store 1, product set C in Treatment Store 2, and product set A in Treatment Store 3), but also displayed at different vertical locations (i.e., at the eye-level in Treatment Store 1, at the stoop-level in Treatment Store 2, and at the stretch-level in Treatment Store 3). Hence, the estimated indifference in sales can also be attributed to the differences in vertical locations at which a product set is kept constant. To address this concern, we consider the data from Rounds 1 and 2 in our experiment. As illustrated in Figure 4.3.3, between Rounds 1 and 2, the product sets remain the same always at the eye-level in the three treatment stores. Hence, a comparison of the sales for products sets at the eye-level between Round 1 and Round 2 enables us to test the insensitivity assumption while controlling for the vertical locations of those products.

We have a total of 9,041 observations for our analysis by focusing on the data from the products placed at eye-level. Table A3.4 provides the descriptive statistics for the data used in this robustness test. We estimate Equation (4.4.2) with only one modification: $Post_d$ now represents a binary variable that equals 1 for time periods in Round 2 and 0 for time periods in Round 1. We also replace $VerLoc_g$ with $ProductSet_g$, indicating whether the product set placed at the eye-level is product set A, B, or C.

Table A3.4: Descriptive statistics of the data sample in our pilot experiment.

	Round 1				Round 2			
	Treated stores		Control stores		Treated stores		Control stores	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Quantity	0.39	0.37	0.55	0.39	0.43	0.39	0.55	0.39
Revenue	0.90	0.74	1.26	0.78	0.98	0.77	1.27	0.78
Price	5.99	4.00	5.71	3.99	5.90	3.96	5.86	4.00
SetPromotion	1.29	1.28	1.46	1.27	1.32	1.26	1.46	1.21
FridgePromotion	1.10	1.02	1.21	1.03	1.16	1.04	1.23	0.99
SetStockout	0.76	0.88	1.03	0.99	0.65	0.84	0.76	0.99
FridgeStockout	0.83	0.57	1.10	0.61	0.83	0.63	0.82	0.75
NumFaces	5.58	1.59	5.54	1.55	5.73	1.72	5.77	1.76
Number of product sets	49		123		47		119	
Sample Size	9,041							

Table A3.5 provides the results. We find that the DiD estimators are not statistically significant in any specifications. Furthermore, in all specifications, we compare the DiD estimators using a Wald test and find that they are not statistically different from each other ($F = 2.1$, $p = 0.125$). The results establish that controlling for the vertical locations, the sales of a product set at a vertical location is not sensitive to swapping the product sets in the other two vertical locations, ruling out the alternative explanation for the insensitivity assumption test.

Table A3.5: Estimation results for our pilot experiment.

	(1) Quantity	(2) Quantity	(3) Revenue	(4) Revenue
$Post \times Treatment$	0.0338	0.00683	0.0712	0.0175
	(0.0298)	(0.0292)	(0.0769)	(0.0765)
$ProductSet(A) \times Post \times$	-0.0796	-0.0381	-0.183	-0.0988
$Treatment$	(0.0441)	(0.0463)	(0.104)	(0.113)
$ProductSet(B) \times Post \times$	0.0522	0.0587	0.138	0.151
$Treatment$	(0.0573)	(0.0579)	(0.121)	(0.124)
$TVControls$	No	Yes	No	Yes
Sample size	9,041	9,041	9,041	9,041
Adj. R^2	0.068	0.086	0.064	0.074

Standard errors are clustered at the product set-level and presented in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

A3.3 Main Experiment: Tables

Table A3.6: Estimates for Equation (4.5.1).

	(1) Quantity	(2) Quantity	(3) Revenue	(4) Revenue
$Intervention^1$	0.114***	0.0827***	0.229***	0.182**
	(0.0243)	(0.0243)	(0.0561)	(0.0676)
$Treated^1 \times Intervention^1$	0.0558	0.0527	0.112	0.0933
	(0.0365)	(0.0344)	(0.0736)	(0.0700)
$Treated^2 \times Intervention^1$	-0.0164	-0.0290	0.0352	0.0137
	(0.0291)	(0.0282)	(0.0667)	(0.0661)
$Treated^3 \times Intervention^1$	0.0661*	0.00890	0.167*	0.0860
	(0.0332)	(0.0310)	(0.0722)	(0.0653)
$Intervention^2$	0.159***	0.124***	0.342***	0.288***
	(0.0240)	(0.0249)	(0.0499)	(0.0632)
$Treated^1 \times Intervention^2$	0.143***	0.132***	0.316***	0.283***
	(0.0381)	(0.0358)	(0.0722)	(0.0702)
$Treated^2 \times Intervention^2$	0.0712*	0.0579	0.301***	0.284**
	(0.0351)	(0.0340)	(0.0875)	(0.0875)
$Treated^3 \times Intervention^2$	0.0762*	0.0201	0.193**	0.114
	(0.0343)	(0.0339)	(0.0720)	(0.0711)
$Intervention^3$	0.164***	0.138***	0.360***	0.318***
	(0.0254)	(0.0255)	(0.0595)	(0.0713)
$Treated^1 \times Intervention^3$	0.101*	0.0866*	0.221*	0.181*
	(0.0426)	(0.0379)	(0.0911)	(0.0855)
$Treated^2 \times Intervention^3$	0.0577	0.0323	0.183*	0.145
	(0.0365)	(0.0349)	(0.0892)	(0.0895)
$Treated^3 \times Intervention^3$	0.106*	0.0263	0.246**	0.120
	(0.0412)	(0.0385)	(0.0828)	(0.0770)
$TVControls$	No	Yes	No	Yes
Sample size	20,291	20,291	20,291	20,291
Adj. R^2	0.139	0.163	0.139	0.149

Standard errors are clustered at the product-set level and presented in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A3.7 Estimates for Equation (4.5.2).

	(1)	(2)	(3)	(4)
	Quantity	Quantity	Revenue	Revenue
$Intervention^1$	0.00176	-0.0316	0.00339	-0.0380
	(0.0460)	(0.0436)	(0.119)	(0.127)
$ProductSet^A \times Intervention^1$	0.113**	0.118**	0.223*	0.227*
	(0.0548)	(0.0525)	(0.128)	(0.126)
$ProductSet^B \times Intervention^1$	0.233***	0.233***	0.472***	0.450***
	(0.0601)	(0.0551)	(0.150)	(0.141)
$Treated^1 \times Intervention^1$	0.0960*	0.103*	0.261*	0.253*
	(0.0580)	(0.0545)	(0.149)	(0.144)
$Treated^2 \times Intervention^1$	0.100*	0.0830	0.360***	0.324**
	(0.0544)	(0.0512)	(0.137)	(0.130)
$Treated^3 \times Intervention^1$	0.156**	0.115*	0.425***	0.363**
	(0.0624)	(0.0598)	(0.149)	(0.142)
$ProductSet^A \times Treated^1 \times Intervention^1$	0.0329	0.00605	-0.0605	-0.0895
	(0.0857)	(0.0808)	(0.176)	(0.174)
$ProductSet^A \times Treated^2 \times Intervention^1$	-0.129*	-0.114*	-0.408***	-0.384**
	(0.0677)	(0.0659)	(0.153)	(0.149)
$ProductSet^A \times Treated^3 \times Intervention^1$	-0.0781	-0.106	-0.240	-0.270
	(0.0797)	(0.0780)	(0.171)	(0.170)
$ProductSet^B \times Treated^1 \times Intervention^1$	-0.170**	-0.166**	-0.412**	-0.401**
	(0.0799)	(0.0761)	(0.189)	(0.183)
$ProductSet^B \times Treated^2 \times Intervention^1$	-0.230***	-0.236***	-0.585***	-0.572***
	(0.0727)	(0.0678)	(0.176)	(0.169)
$ProductSet^B \times Treated^3 \times Intervention^1$	-0.202**	-0.222***	-0.556***	-0.580***
	(0.0817)	(0.0739)	(0.190)	(0.178)
$Intervention^2$	0.136**	0.0956*	0.335***	0.281**
	(0.0525)	(0.0531)	(0.116)	(0.127)
$ProductSet^A \times Intervention^2$	0.0232	0.0308	-0.0570	-0.0473
	(0.0617)	(0.0608)	(0.125)	(0.126)
$ProductSet^B \times Intervention^2$	0.0454	0.0520	0.0744	0.0648
	(0.0629)	(0.0610)	(0.137)	(0.134)
$Treated^1 \times Intervention^2$	0.148**	0.140**	0.367**	0.330**
	(0.0709)	(0.0676)	(0.147)	(0.145)

Continued on next page

Table A3.7 Estimates for Equation (4.5.2).

	(1) Quantity	(2) Quantity	(3) Revenue	(4) Revenue
$Treated^2 \times Intervention^2$	0.233***	0.214***	0.720***	0.701***
	(0.0675)	(0.0658)	(0.178)	(0.178)
$Treated^3 \times Intervention^2$	0.0296	-0.000798	0.0114	-0.0315
	(0.0710)	(0.0688)	(0.147)	(0.142)
$ProductSet^A \times Treated^1 \times Intervention^2$	0.114	0.0958	0.195	0.182
	(0.0911)	(0.0867)	(0.172)	(0.171)
$ProductSet^A \times Treated^2 \times Intervention^2$	-0.239***	-0.219***	-0.604***	-0.597***
	(0.0815)	(0.0769)	(0.194)	(0.190)
$ProductSet^A \times Treated^3 \times Intervention^2$	0.0564	0.0167	0.254	0.206
	(0.0883)	(0.0841)	(0.168)	(0.161)
$ProductSet^B \times Treated^1 \times Intervention^2$	-0.136	-0.123	-0.352**	-0.319*
	(0.0887)	(0.0834)	(0.176)	(0.171)
$ProductSet^B \times Treated^2 \times Intervention^2$	-0.180**	-0.187**	-0.539**	-0.552***
	(0.0885)	(0.0850)	(0.213)	(0.212)
$ProductSet^B \times Treated^3 \times Intervention^2$	0.0742	0.0394	0.266	0.207
	(0.0858)	(0.0849)	(0.187)	(0.187)
$Intervention^3$	0.0376	0.0113	0.108	0.0751
	(0.0448)	(0.0437)	(0.123)	(0.134)
$ProductSet^A \times Intervention^3$	0.151***	0.146***	0.317**	0.306**
	(0.0547)	(0.0514)	(0.133)	(0.130)
$ProductSet^B \times Intervention^3$	0.239***	0.242***	0.457***	0.438***
	(0.0622)	(0.0580)	(0.160)	(0.153)
$Treated^1 \times Intervention^3$	0.275***	0.259***	0.738***	0.688***
	(0.0579)	(0.0556)	(0.171)	(0.171)
$Treated^2 \times Intervention^3$	0.0993	0.0678	0.165	0.117
	(0.0680)	(0.0622)	(0.158)	(0.152)
$Treated^3 \times Intervention^3$	0.225***	0.163**	0.518***	0.418***
	(0.0766)	(0.0732)	(0.164)	(0.152)
$ProductSet^A \times Treated^1 \times Intervention^3$	-0.191**	-0.187**	-0.648***	-0.628***
	(0.0928)	(0.0842)	(0.203)	(0.201)
$ProductSet^A \times Treated^2 \times Intervention^3$	-0.0749	-0.0502	-0.0677	-0.0469
	(0.0826)	(0.0766)	(0.174)	(0.166)
$ProductSet^A \times Treated^3 \times Intervention^3$	-0.0627	-0.0899	-0.215	-0.233

Continued on next page

Table A3.7 Estimates for Equation (4.5.2)**.**

	(1) Quantity	(2) Quantity	(3) Revenue	(4) Revenue
	(0.0986)	(0.0937)	(0.193)	(0.188)
$ProductSet^B \times Treated^1 \times Intervention^3$	-0.333***	-0.330***	-0.889***	-0.864***
	(0.0963)	(0.0872)	(0.228)	(0.221)
$ProductSet^B \times Treated^2 \times Intervention^3$	-0.0520	-0.0626	0.128	0.131
	(0.0867)	(0.0816)	(0.222)	(0.219)
$ProductSet^B \times Treated^3 \times Intervention^3$	-0.308***	-0.329***	-0.623***	-0.677***
	(0.0976)	(0.0891)	(0.210)	(0.193)
$TVControls$	No	Yes	No	Yes
Sample size	20,291	20,291	20,291	20,291
Adj. R^2	0.158	0.181	0.164	0.172

Standard errors are clustered at the product set-level and presented in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A3.8: Estimates of the vertical location effect on the overall sales.

	(1) Quantity	(2) Quantity	(3) Revenue	(4) Revenue
$Intervention^1$	0.108***	0.0646*	0.217**	0.155
	(0.0276)	(0.0324)	(0.0753)	(0.103)
$Treated^1 \times Intervention^1$	0.0648	0.0608	0.127	0.117
	(0.0440)	(0.0431)	(0.0944)	(0.0973)
$Treated^2 \times Intervention^1$	-0.00535	-0.0113	0.0556	0.0391
	(0.0325)	(0.0295)	(0.0844)	(0.0801)
$Treated^3 \times Intervention^1$	0.0828*	0.0415	0.213*	0.172*
	(0.0416)	(0.0386)	(0.0939)	(0.0852)
$Intervention^2$	0.193***	0.151***	0.540***	0.490***
	(0.0257)	(0.0321)	(0.0675)	(0.0962)
$Treated^1 \times Intervention^2$	0.114**	0.100*	0.0675	0.0615
	(0.0397)	(0.0382)	(0.0817)	(0.0790)
$Treated^2 \times Intervention^2$	0.0472	0.0388	0.198*	0.165
	(0.0414)	(0.0364)	(0.0952)	(0.0955)
$Treated^3 \times Intervention^2$	0.0562	0.00394	0.0649	0.00181
	(0.0378)	(0.0386)	(0.0842)	(0.0945)
$Intervention^3$	0.186***	0.150***	0.463***	0.418***
	(0.0287)	(0.0334)	(0.0870)	(0.116)
$Treated^1 \times Intervention^3$	0.0687	0.0510	0.0642	0.0423
	(0.0543)	(0.0485)	(0.112)	(0.108)
$Treated^2 \times Intervention^3$	0.0611	0.0442	0.158	0.120
	(0.0425)	(0.0394)	(0.112)	(0.112)
$Treated^3 \times Intervention^3$	0.0872	0.0127	0.160	0.0585
	(0.0484)	(0.0410)	(0.104)	(0.0931)
$TVControls$	No	Yes	No	Yes
Sample size	7,753	7,753	7753	7,753
Adj. R^2	0.239	0.271	0.275	0.302

Standard errors are clustered at the product set-level and presented in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

A3.4 Heterogeneous Vertical Location Effect: Details and Tables

With the heterogeneity analysis, our intention is to identify the types of products that benefit the most from being placed at the eye-level locations, rather than to identify the asymmetric effect of the change in vertical location. Therefore, for this analysis, we use data only from the post-intervention period (i.e., from Rounds 2, 3, and 4) at the treatment stores. Because the experiment is randomized, we expect that the estimated vertical location effects from this analysis would be comparable to those obtained using the DiD specification in Section 4.5.3 and reported in Figure 4.5.1. Thus, a linear regression on post-intervention data only from treated stores serves as an alternative specification (compared to our DiD specification) for the estimation of the vertical location effects. To verify this, we first specify the following linear regression model:

$$Y_{ijt} = \beta_0 + \beta_1 Stretch_{ijt} + \beta_2 Eye_{ijt} + TVControls_{ijt} + Seasonality_t + \epsilon_{ijt}, \quad \text{(A3.1)}$$

where $Stretch_{ijt}$ and Eye_{ijt} are binary variables indicating whether product set i in store j on day t is placed at the stretch- and the eye-levels, respectively. Therefore, β_1 and β_2 capture the vertical location effect when a product set is placed at the stretch-level and at the eye-level, respectively, compared to at the stoop-level.

Table A3.9 shows the estimation results from this specification. The estimated eye-level and stretch-level effects are consistent with those obtained using the DiD model in Section 4.5.3, serving as a robustness test for our DiD identification strategy and providing a sanity check for our specification for the heterogeneity analysis.

Table A3.9: Estimation results for cross-group comparison specification.

	(1) Quantity	(2) Quantity	(3) Revenue	(4) Revenue
Stretch	0.0560**	0.0495***	0.132***	0.128***
	(0.0168)	(0.0141)	(0.0320)	(0.0301)
Eye	0.0965***	0.0912***	0.283***	0.275***
	(0.0183)	(0.0169)	(0.0498)	(0.0490)
TVControls	No	Yes	No	Yes
Sample size	8,155	8,155	8,155	8,155
Adj. R^2	0.025	0.216	0.042	0.085

Standard errors are clustered at the product set-level and presented in parentheses.
$^+$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Because in the heterogeneity analysis, we aim to identify the product characteristics that are associated with more eye-level effects, we construct product-level data using observations from the post-intervention period at treatments stores. To estimate the heterogeneity, we specify the following linear regression model with the interaction terms as follows:

$$Y_{ijt} = \beta_0 + \beta_1 Stretch_{ijt} + \beta_2 Eye_{ijt} + \beta_3 Stretch_{ijt} \times ProductFeature_i + \beta_4 Eye_{ijt} \times ProductFeature_i + ProductFeature_i + TVControls_{ijt} + Seasonality_t + \epsilon_{ijt}, \tag{A3.2}$$

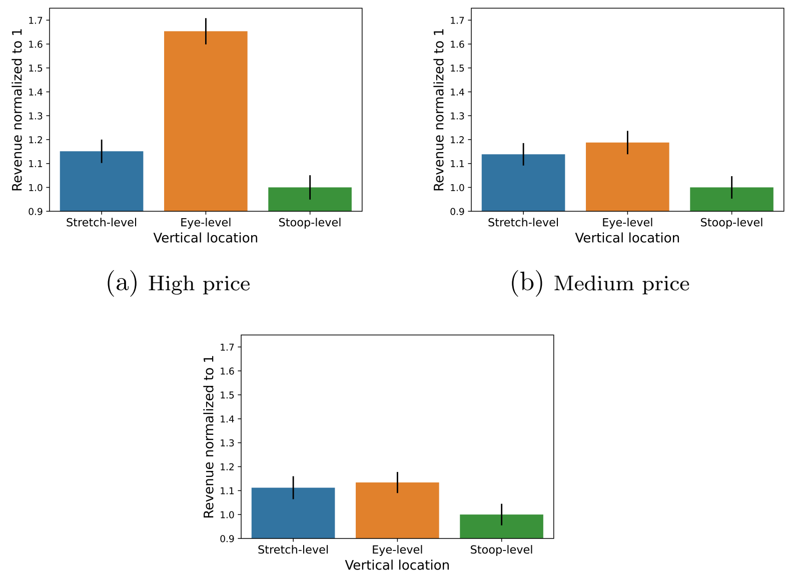
where β_3 and β_4 capture the moderating effect of a specific product feature (i.e., $ProductFeature$) on the stretch-level and eye-level effects (compared to the stoop-level), respectively. The product features we evaluate include price, promotions, broken assortment, market concentration, number of substitute products, package size, and flashiness.

Price: We create a categorical $ProductFeature$ variable by dividing the products into three groups based on the price value: low, medium, and high. We use the 33^{rd} and 66^{th} percentiles of the price distribution to create these groups. We set the low price level to be the base level and estimate Equation A3.2. The results are reported in Table A3.10 and Figure A3.1.

Table A3.10: Estimation results for heterogeneous vertical location effect (price).

	(1) Quantity	(2) Quantity	(3) Revenue	(4) Revenue
Stretch	0.0549**	0.0585**	0.0952**	0.103**
	(0.0203)	(0.0196)	(0.0333)	(0.0318)
Eye	0.0621**	0.0657**	0.116***	0.123***
	(0.0212)	(0.0209)	(0.0340)	(0.0335)
Price(Medium)	-0.0916*	-0.0740+	-0.0313	-0.0241
	(0.0379)	(0.0385)	(0.0621)	(0.0636)
Price(High)	-0.342***	-0.205***	-0.361***	-0.310***
	(0.0334)	(0.0444)	(0.0587)	(0.0781)
Stretch×	0.00902	0.00424	0.0325	0.0214
Price(Medium)	(0.0301)	(0.0295)	(0.0521)	(0.0511)
Stretch×	-0.0141	-0.0259	0.00318	-0.0104
Price(High)	(0.0244)	(0.0235)	(0.0463)	(0.0455)
Eye×	0.0357	0.0257	0.0657	0.0453
Price(Medium)	(0.0319)	(0.0316)	(0.0548)	(0.0546)
Eye×	0.0837**	0.0673*	0.296***	0.275***
Price(High)	(0.0272)	(0.0271)	(0.0573)	(0.0574)
TVControls	No	Yes	No	Yes
Sample size	36,876	36,876	36,876	36,876
Adj. R^2	0.062	0.078	0.032	0.041

Standard errors are clustered at the product level and presented in parentheses $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$



(a) High price

(b) Medium price

(c) Low price

Figure A3.1: Heterogeneity of the vertical location effect with respect to price (revenue).

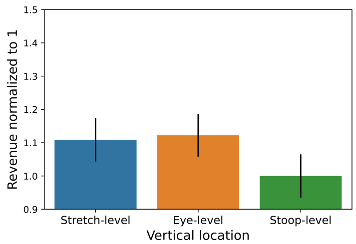
Promotion: Similarly, we create a categorical $ProductFeature$ variable by dividing the products into three groups based on the monetary value of the promotions using the 33^{rd} and 66^{th} percentiles of the promotions distribution. We set the low promotion level to be the base level and estimate Equation A3.2. The results are reported in Table A3.11 and Figure A3.2.

Table A3.11: Estimation results for heterogeneous vertical location effect (promotion).

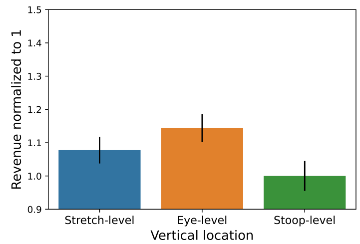
	(1) Quantity	(2) Quantity	(3) Revenue	(4) Revenue
$Stretch$	0.0495^{**}	0.0489^{**}	0.111^{***}	0.107^{***}
	(0.0167)	(0.0152)	(0.0326)	(0.0321)
Eye	0.143^{***}	0.140^{***}	0.375^{***}	0.367^{***}
	(0.0170)	(0.0163)	(0.0403)	(0.0399)
$Promotion(Medium)$	0.326^{***}	0.317^{***}	0.438^{***}	0.543^{***}
	(0.0343)	(0.0423)	(0.0564)	(0.0702)
$Promotion(High)$	0.235^{***}	0.388^{***}	0.331^{***}	0.675^{***}
	(0.0310)	(0.0613)	(0.0563)	(0.104)
$Stretch\times$	-0.0145	-0.0118	-0.0434	-0.0314
$Promotion(Medium)$	(0.0318)	(0.0304)	(0.0564)	(0.0554)
$Stretch\times$	0.000928	0.0113	-0.00384	0.0143
$Promotion(High)$	(0.0286)	(0.0274)	(0.0520)	(0.0509)
$Eye\times$	-0.0744^{*}	-0.0673^{+}	-0.239^{***}	-0.226^{***}
$Promotion(Medium)$	(0.0357)	(0.0347)	(0.0651)	(0.0646)
$Eye\times$	-0.0878^{**}	-0.0779^{**}	-0.244^{***}	-0.231^{***}
$Promotion(High)$	(0.0298)	(0.0293)	(0.0598)	(0.0594)
$TVControls$	No	Yes	No	Yes
Sample size	36,876	36,876	36,876	36,876
Adj. R^2	0.055	0.087	0.036	0.049

Standard errors are clustered at the product level and presented in parentheses $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

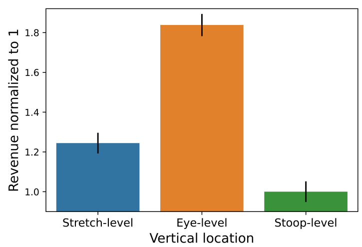
Broken Assortment: We define a product on a shelf as having a broken assortment (i.e., $ProductFeature = 1$) if at least one item within that shelf is out of stock and not having a broken assortment otherwise (i.e., $ProductFeature = 0$). The



(a) High promotion



(b) Medium promotion



(c) Low promotion

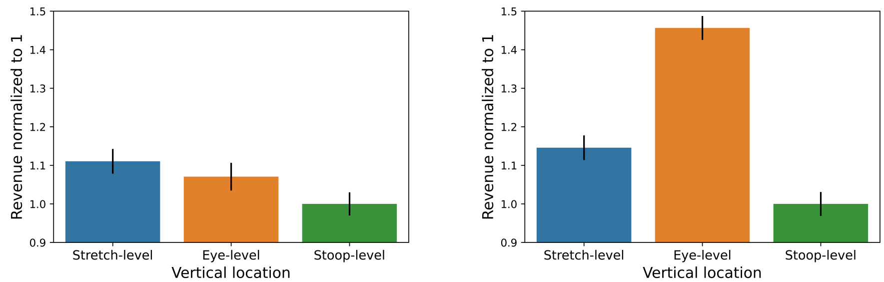
Figure A3.2: Heterogeneity of the vertical location effect with respect to promotion.

results from the estimation of Equation A3.2 with this variable are demonstrated in Table A3.12 and Figure A3.3.

Table A3.12: Estimation results for heterogeneous vertical location effect (broken assortment).

	(1) Quantity	(2) Quantity	(3) Revenue	(4) Revenue
Stretch	0.0636***	0.0609***	0.117***	0.121***
	(0.0163)	(0.0148)	(0.0291)	(0.0280)
Eye	0.183***	0.174***	0.382***	0.379***
	(0.0170)	(0.0162)	(0.0330)	(0.0324)
BrokenAssortment	0.0604**	-0.0298	0.101**	-0.0523
	(0.0217)	(0.0225)	(0.0374)	(0.0398)
Stretch×	-0.0318	-0.0249	-0.0309	-0.0357
BrokenAssortment	(0.0283)	(0.0250)	(0.0503)	(0.0469)
Eye×	-0.195***	-0.174***	-0.340***	-0.334***
BrokenAssortment	(0.0273)	(0.0263)	(0.0507)	(0.0502)
TVControls	No	Yes	No	Yes
Sample size	36,876	36,876	36,876	36,876
Adj. R^2	0.019	0.078	0.022	0.043

Standard errors are clustered at the product level and presented in parentheses $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$



(a) Broken assortment (b) Complete assortment

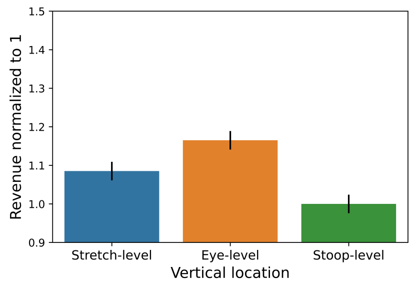
Figure A3.3: Heterogeneity of the vertical location effect with respect to a broken assortment.

Market Concentration: We create a binary $ProductFeature$ variable by dividing the products into two groups based on the HHI using the 50^{th} percentile of the HHI distribution. We also explored alternative split points using the mean, yielding a similar finding. We set the low market concentration level to be the base level and estimate Equation A3.2. The results are reported in Table A3.13 and Figure A3.4.

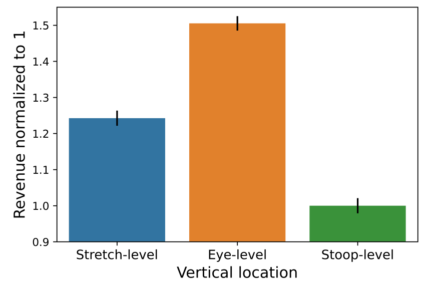
Table A3.13: Estimation results for heterogeneous vertical location effect (market concentration).

	(1) Quantity	(2) Quantity	(3) Revenue	(4) Revenue
Stretch	0.0474***	0.0490***	0.0933***	0.0950***
	(0.0108)	(0.0108)	(0.0211)	(0.0211)
Eye	0.101***	0.100***	0.186***	0.184***
	(0.0109)	(0.0109)	(0.0212)	(0.0212)
MarketConcentration(Low)	-0.347***	-0.333***	-0.568***	-0.546***
	(0.0197)	(0.0198)	(0.0384)	(0.0386)
Stretch×	0.0134	0.0105	0.0464^{+}	0.0426
MarketConcentration(Low)	(0.0141)	(0.0142)	(0.0276)	(0.0277)
Eye×	0.0252^{+}	0.0245^{+}	0.143***	0.143***
MarketConcentration(Low)	(0.0141)	(0.0141)	(0.0274)	(0.0276)
TVControls	No	Yes	No	Yes
Sample size	36,876	36,876	36,876	36,876
Adj. R^2	0.011	0.012	0.010	0.011

Standard errors are clustered at the product level and presented in parentheses $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$



(a) High market concentration



(b) Low market concentration

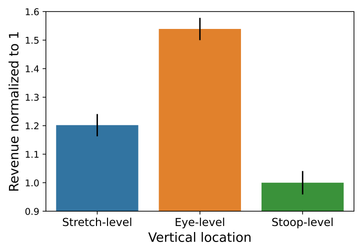
Figure A3.4: Heterogeneity of the vertical location effect with respect to market concentration.

Number of Substitute Products: We create a categorical $ProductFeature$ variable by dividing the products into three groups based on the number of substitute products using the 33^{rd} and 66^{th} percentiles of the number of substitute products distribution. We set the low substitution level to be the base level and estimate Equation A3.2. The results are reported in Table A3.14 and Figure A3.5.

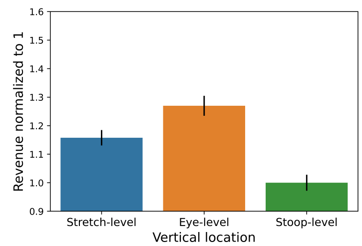
Table A3.14: Estimation results for heterogeneous vertical location effect (number of substitute products).

	(1) Quantity	(2) Quantity	(3) Revenue	(4) Revenue
$Stretch$	0.0463**	0.0452**	0.102***	0.100***
	(0.0155)	(0.0151)	(0.0284)	(0.0277)
Eye	0.101***	0.100***	0.227***	0.226***
	(0.0152)	(0.0150)	(0.0307)	(0.0304)
$SubstituteProducts(Medium)$	-0.0758**	-0.0572*	-0.115*	-0.0840+
	(0.0277)	(0.0272)	(0.0491)	(0.0485)
$SubstituteProducts(High)$	-0.129***	-0.0943**	-0.257***	-0.198**
	(0.0343)	(0.0341)	(0.0663)	(0.0670)
$Stretch\times$	0.00590	0.00870	0.0207	0.0247
$SubstituteProducts(Medium)$	(0.0220)	(0.0216)	(0.0417)	(0.0410)
$Stretch\times$	0.0160	0.0177	0.0339	0.0362
$SubstituteProducts(High)$	(0.0228)	(0.0224)	(0.0430)	(0.0423)
$Eye\times$	-0.00577	-0.00514	-0.0135	-0.0128
$SubstituteProducts(Medium)$	(0.0253)	(0.0252)	(0.0509)	(0.0506)
$Eye\times$	0.0510*	0.0481*	0.145**	0.140**
$SubstituteProducts(High)$	(0.0234)	(0.0233)	(0.0523)	(0.0523)
$TVControls$	No	Yes	No	Yes
Sample size	36,876	36,876	36,876	36,876
Adj. R^2	0.025	0.027	0.027	0.028

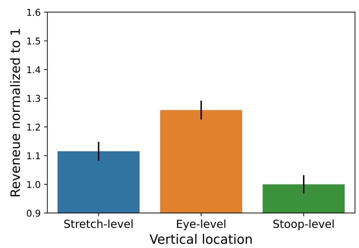
Standard errors are clustered at the product level and presented in parentheses
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$



(a) High number of substitutes



(b) Medium number of substitutes



(c) Low number of substitutes

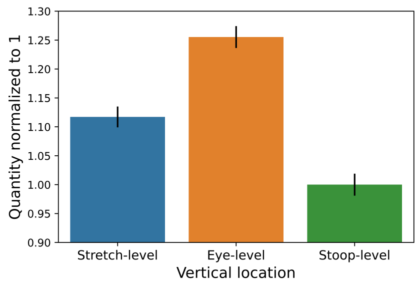
Figure A3.5: Heterogeneity of the vertical location effect with respect to the number of substitute products.

Package Size: We create a binary $ProductFeature$ variable by dividing the products into two groups based on the package size using the 50^{th} percentiles of the package size distribution. We also explored alternatives split points using the mean, yielding a similar finding. We set the small package size level to be the base level and estimate Equation A3.2. The results are reported in Table A3.15 and Figures A3.6–A3.7.

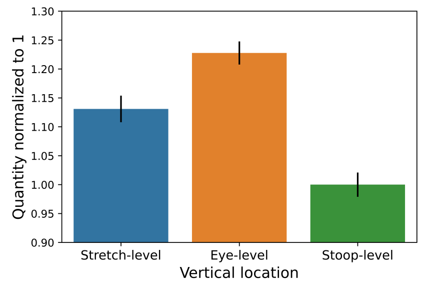
Table A3.15: Estimation results for heterogeneous vertical location effect (package size).

	(1)	(2)	(3)	(4)
	Quantity	Quantity	Revenue	Revenue
Stretch	0.0401**	0.0389**	0.0830**	0.0802**
	(0.0152)	(0.0147)	(0.0271)	(0.0267)
Eye	0.0884***	0.0845***	0.181***	0.170***
	(0.0173)	(0.0174)	(0.0321)	(0.0326)
PackageSize(large)	0.0613*	0.194***	0.148**	0.261***
	(0.0311)	(0.0303)	(0.0521)	(0.0532)
Stretch×	0.0289	0.0298	0.0591	0.0624
PackageSize(large)	(0.0241)	(0.0217)	(0.0427)	(0.0406)
Eye	0.0224	0.0245	0.118	0.128^{+}
PackageSize(large)	(0.0252)	(0.0719)	(0.0496)	(0.0724)
TVControls	No	Yes	No	Yes
Sample size	36,876	36,876	36,876	36,876
Adj. R^2	0.018	0.092	0.026	0.050

Standard errors are clustered at the product level and presented in parentheses $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

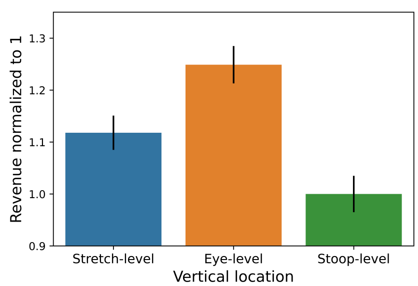


(a) Small package size

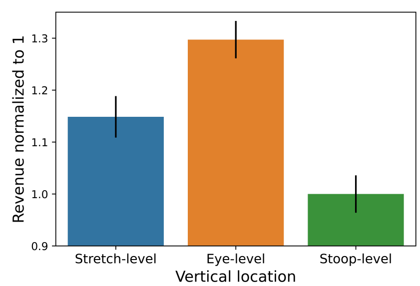


(b) Large package size

Figure A3.6: Heterogeneity of the vertical location effect with respect to package size.



(a) Small package size



(b) Large package size

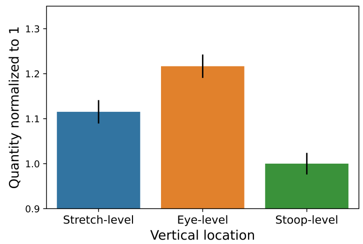
Figure A3.7: Heterogeneity of the vertical location effect with respect to package size.

Flashiness: In our experiments, we keep images of all shelves during store visits to archive these images for future reference. Additionally, the marketing team responsible for designing the planogram possesses images of all the products. These images were used to calculate the flashiness level of the different products. More specifically, to measure the flashiness, we employ two different metrics. The first metric is based on the hue, color saturation, and color value of the product package, as proposed by Gorn et al. (1997). The second metric of flashiness is visual saliency, which refers to the extent to which an object stands out from its surrounding objects (Gidlöf et al. 2017). We use the algorithm proposed by Itti et al. (1998) to calculate the visual saliency of each product on the shelf, assigning a saliency factor between 0 and 1 based on elements that are known to attract visual attention, such as color, intensity, contrast, and edge orientation. We then create a categorical $ProductFeature$ variable by dividing the products into three groups based on their flashiness levels using the 33^{rd} and 66^{th} percentiles of the flashiness distribution. We set the low flashiness level to be the base level and estimate Equation A3.2. The results are presented in Table A3.16 and Figures A3.8–A3.9.

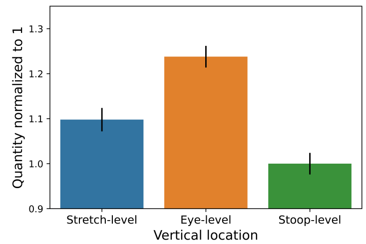
Table A3.16: Estimation results for heterogeneous vertical location effect (flashiness).

	(1) Quantity	(2) Quantity	(3) Revenue	(4) Revenue
Stretch	0.0606**	0.0583**	0.113**	0.112**
	(0.0205)	(0.0189)	(0.0380)	(0.0369)
Eye	0.106***	0.101***	0.244***	0.233***
	(0.0214)	(0.0210)	(0.0435)	(0.0435)
Flashiness(Medium)	0.0315	0.0214	0.0220	0.0190
	(0.0362)	(0.0324)	(0.0630)	(0.0590)
Flashiness(High)	0.0261	0.0264	-0.00459	-0.00721
	(0.0367)	(0.0328)	(0.0631)	(0.0597)
Stretch×	-0.0165	-0.0164	-0.0241	-0.0279
Flashiness(Medium)	(0.0297)	(0.0277)	(0.0529)	(0.0512)
Stretch×	-0.00829	-0.00803	0.00264	0.00487
Flashiness(High)	(0.0282)	(0.0257)	(0.0511)	(0.0496)
Eye×	-0.00291	0.00128	-0.00426	0.00359
Flashiness(Medium)	(0.0308)	(0.0301)	(0.0616)	(0.0616)
Eye×	-0.0156	-0.0170	-0.0216	-0.0172
Flashiness(High)	(0.0307)	(0.0297)	(0.0609)	(0.0606)
TVControls	No	Yes	No	Yes
Sample size	36,876	36,876	36,876	36,876
Adj. R^2	0.015	0.070	0.018	0.035

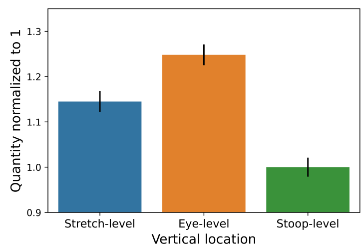
Standard errors are clustered at the product level and presented in parentheses $^+ p < 0.10$, $^* p < 0.05$, $^{**} p < 0.01$, $^{***} p < 0.001$



(a) High level of flashiness

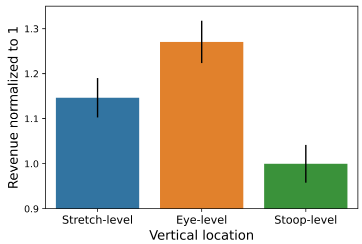


(b) Medium level of flashiness

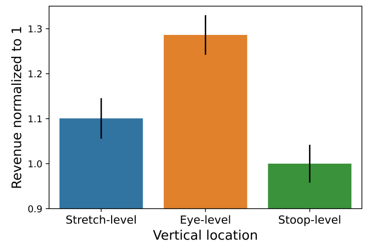


(c) Low level of flashiness

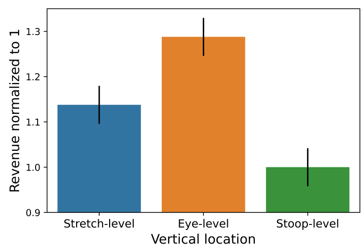
Figure A3.8: Heterogeneity of the vertical location effect with respect to product flashiness.



(a) High level of flashiness



(b) Medium level of flashiness



(c) Low level of flashiness

Figure A3.9: Heterogeneity of the vertical location effect with respect to product flashiness.

References

Gorn GJ, Chattopadhyay A, Yi T, Dahl DW (1997) Effects of color as an executional cue in advertising: They're in the shade. *Management Science* 43(10):1387–1400.

Gidlof K, Anikin A, Lingonblad M, Wallin A (2017) Looking is buying. how visual attention and choice are affected by consumer preferences and properties of the supermarket shelf. *Appetite* 116:29–38.

Itti L, Koch C, Niebur E (1998) model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(11):1254–1259.

5

Conclusions

In this thesis, the exploration into the realm of data analytics in healthcare and retail operations management unfolds through three distinct essays, each addressing pivotal issues in these domains. The primary objective of these projects is to leverage data science to enhance the efficiency of healthcare and retail operations.

In the first essay, the primary focus is the critical issue of crowding in emergency departments (EDs). Drawing upon data from 500,000 patients across eight hospitals, I employed a multimethodology approach to investigate the reasons behind the EDs' crowding issue in Quebec. Initially, my analysis identified boarding congestion as one of the reasons behind ED overcrowding in the studied hospitals. Further examination revealed an inverted-U shaped relationship between boarding congestion and treatment time in EDs.

Utilizing econometric analysis, I demonstrated that the increased workload on ED resources due to boarding patients, coupled with hospitalist visits triggered by boarding congestion, constitutes the primary cause of this observed phenomenon. Subsequently, using an analytical framework, I introduced two operational interventions aimed at mitigating the negative impact of boarding congestion on treatment time.

Simulation results indicate that implementing these two interventions together can result in a 68% reduction in the impact of boarding congestion, providing valuable insights for informed decision-making in ED management.

In the second essay, I redirected my attention to the retail sector, focusing on the growing trend of pickup partnerships. These partnerships empower online retailers to provide in-store pickup services by collaborating with physical stores. Within this study, I assessed two prevalent policies adopted by online retailers in these partnerships: the fixed fee policy and the coupon policy.

Through the creation of a stylized model, I identified circumstances guiding the choice between fixed fee and coupon policies. Additionally, the analysis highlighted potential inefficiencies stemming from misaligned incentives and proposed a new policy to enhance the overall efficiency of pickup partnerships.

In the third essay, I delved into the impact of sales amid simultaneous changes in the vertical locations of multiple products within retail stores. I conducted a novel field experiment spanning 20 weeks across six retail stores, revealing that the effect of vertical location changes is contingent on product reorganization and heterogeneous across products. The study quantifies the sales boost at eye-level shelves and identifies product profiles benefiting the most from this strategic placement.

Moreover, the research underscores the importance of a meticulous selection of products at eye level and offers insights into optimizing planograms. These optimizations have the potential to lead to profit increases of up to 3%.

In conclusion, the incorporation of data science into healthcare and retail operations management marks a fundamental shift in the operational landscape of these industries. The capability to leverage data for informed decision-making, personalized experiences, and enhanced operational efficiency is re-defining the trajectory of healthcare and retail, presenting a promising outlook for improved outcomes for businesses and the individuals they serve. As these industries persist in embracing and adapting to the data-driven era, the potential for innovation and positive impact is limitless. Throughout this thesis, I demonstrate the tangible ways in which data science contributes to the realms of healthcare and retail operations management.