IE UNIVERSIDAD

TESIS DOCTORAL/DOCTORAL DISSERTATION

ENSAYOS SOBRE OPERACIONES MINORISTAS DE COMESTIBLES

ESSAYS ON GROCERY RETAIL OPERATIONS

BENGÜ NUR ÖZDEMİR

SEGOVIA, 2024

IE UNIVERSIDAD

TESIS DOCTORAL/DOCTORAL DISSERTATION

Ensayos sobre Operaciones Minoristas de Comestibles

Essays on Grocery Retail Operations

Bengü Nur Özdemir

Doctoral Thesis Advisor: Antti Tenhiälä

Abstract

This dissertation investigates various issues in grocery retail operations by analyzing field data with econometric techniques.

The first two chapters examine inventory replenishment in a supermarket chain from a behavioral perspective. Particularly, I focus on the potential impact of behavioral biases when store managers receive order proposals from an automatic store replenishment system and decide on order quantities. In Chapter 1, I study a paradox where store managers deviate downward from algorithmic proposals after a stockout. I argue that this counterintuitive behavior is explained by censorship bias, and when this bias triggers such deviation decisions, it has negative performance implications.

In Chapter 2, I use the same data to investigate why supermarket managers order more than the algorithmic proposals. In answering this, I empirically show that store managers are susceptible to a novel bias: newsvendor double counting. My analyses show that order increases triggered by newsvendor double-counting bias do not lead to additional sales. To sum up, these two chapters contribute to the literature by testing behavioral biases with field data, showing that they can explain deviations from algorithms, and by proposing a novel bias.

In Chapter 3, I move the focus to consumer behavior. Retailers are transforming their stores to attract more customers to brick-and-mortars; one way to do so is offering experiential services. I utilize data from a supermarket chain that introduced a taproom service to some stores as an experience-enhancing service. I examine the spillover effects of this

service by focusing on measures related to the retailer's main products. Being the first empirical study of the spillover effects of experiential services, the findings of this study have significant managerial implications for retailers.

Together, these three chapters provide new scholarly insights regarding contemporary and significant issues to retailers, while generating actionable policy implications that can improve management practice.

Resumen

Esta disertación investiga diversos problemas en las operaciones minoristas de comestibles mediante el análisis de datos de campo con técnicas econométricas.

Los dos primeros capítulos examinan la reposición de inventarios en una cadena de supermercados desde una perspectiva conductual. En particular, me concentro en el impacto potencial de los sesgos de comportamiento cuando los gerentes de tienda reciben propuestas de pedidos de un sistema automático de reabastecimiento de tiendas y deciden las cantidades de los pedidos. En el Capítulo 1, estudio una paradoja en la que los gerentes de tienda se desvían hacia abajo de las propuestas algorítmicas después de un desabastecimiento. Sostengo que este comportamiento contraintuitivo se explica por el sesgo de censura, y cuando este sesgo desencadena tales decisiones de desviación, tiene implicaciones negativas para el desempeño. En el capítulo 2, utilizo los mismos datos para investigar por qué los gerentes de supermercados piden más que las propuestas algorítmicas. Al responder a esto, demuestro empíricamente que los gerentes de las tiendas son susceptibles a un sesgo novedoso: la doble contabilización de los vendedores de periódicos. Mis análisis muestran que los aumentos de pedidos provocados por el sesgo de doble conteo de los vendedores de periódicos no generan ventas adicionales. En resumen, estos dos capítulos contribuyen a la literatura al probar los sesgos conductuales con datos de campo, mostrando que pueden explicar las desviaciones de los algoritmos y proponiendo un sesgo novedoso.

En el capítulo 3, me centraré en el comportamiento del consumidor. Los minoristas

están transformando sus tiendas para atraer más clientes a las tiendas físicas; una forma de hacerlo es ofreciendo servicios experienciales. Utilizo datos de una cadena de supermercados que introdujo un servicio de taberna en algunas tiendas como un servicio para mejorar la experiencia. Examino los efectos indirectos de este servicio centrándome en medidas relacionadas con los principales productos del minorista. Al ser el primer estudio empírico de los efectos indirectos de los servicios experienciales, los hallazgos de este estudio tienen importantes implicaciones de gestión para los minoristas.

En conjunto, estos tres capítulos brindan nuevos conocimientos académicos sobre temas contemporáneos e importantes para los minoristas, al tiempo que generan implicaciones políticas procesables que pueden mejorar la práctica de gestión.

Acknowledgements

It takes a village to complete a PhD! I have been very fortunate over the last five years to have my village filled with people whose support has been instrumental in enabling me to finish my dissertation. I first want to give my most sincere gratitude to my advisor, Dr. Antti Tenhiälä. From day one, he treated me as a colleague, valued my ideas, and provided significant guidance along the way. How lucky I was to have an advisor who is honest, considerate, engaged, and committed to the highest standards. Thank you, Antti, for opening many doors for me, and for showing me what kind of advisor I should be one day.

I am also especially grateful to Dr. Stanley Lim. When I emailed him a year ago asking about potential avenues for us to collaborate, he had no idea who I was. Yet, he gave me the greatest gifts of trust and opportunity. Thanks to him, I got the chance to spend three months at Eli Broad Business School, which added a joyful note to the culmination of my PhD journey. Working with him has seen my research skills flourish tremendously, and I cannot thank him enough for that!

Next, I want to take this opportunity to thank my doctoral committee, formed of Dr. Jan Fransoo, Dr. Daniel Corsten, Dr. María Ibáñez and Dr. Antoaneta Momcheva. I am incredibly lucky to have such world-class researchers in my committee, whose invaluable comments and questions improved my papers and the way I think as a researcher.

I would not be able to write any of the following many pages without the indefatigable support and endless encouragement of my parents, Bayram Ali Özdemir and Handan

Özdemir, and my sister Burcu Nur Özdemir. Thank you, my family, for always being there. Each of you have brought me a great deal of comfort in the most challenging times. Had my mom and dad not taught me the importance of hard work and perseverance, I would not have made it this far. And Burcu, thank you for listening to me, sometimes for hours, venting about an analysis. I am sure it was not fun! And thank you for taking care of Canavar whenever I needed to travel, which was much needed for the completion of this dissertation. Sizi çok seviyorum!

And top among my supporters, my husband, Anthony Clarke. Words cannot fully express how grateful I am for him. He tirelessly read my papers many times, listened to my talks before conferences, and flew thousands of miles to support me before big presentations. I do not know who travels to Phoenix from London just to stand by their partner's side, but he did! With his love and belief in me, this demanding process has significantly ($p<0.05$) become much more enjoyable. Thank you, Anthony, for bringing calmness when I was in panic, for showing me compassion when I had self-doubt, and for being my rock.

I cannot finish this section without giving a special mention to my cat, Canavar. She attended every Zoom class and online talk with me over the last five years, for which I believe she deserves at least an honorary Master's. I love you, my baby, my most important collaborator.

List of Tables

List of Figures

Contents

Introduction

The retail sector plays a key role in global economic growth by providing goods that consumers demand, jobs within local communities, and stimulating domestic and international trade. From independent corner shops to multinational supermarket chains, grocery retailers bear the importance of keeping people fed. Despite providing essential goods, change is constant; hence, grocery retailers have to innovate and be flexible and future-ready to survive and thrive during global economic and industry challenges. Important and contemporary topics for grocery retailers include digitalization and retail transformation, which are the main topics of this dissertation.

Within the digitalization context, my first two chapters investigate how automatic store replenishment systems are utilized by supermarkets. Particularly, in these two chapters, I analyze why store managers of a supermarket chain deviate from the algorithmic system's suggestions in inventory replenishment decisions. In the first chapter, I study a paradox in which store managers of a supermarket chain deviate downward from order proposals of an automatic store replenishment system after a stockout when replenishing their inventories. I argue that this counterintuitive ordering behavior can be explained by censorship bias, and when this bias triggers such deviation decisions, it has negative performance implications. To understand the relative importance of censorship bias, I add anchoring bias in my models as a benchmark. I tackle the endogeneity of deviation decisions by estimating a probit model with sample selection, an endogenous switching regression model, and a recursive bivariate probit model with exclusion restrictions. The results from

these various models suggest that store managers' probability of deviating downward increases after a stockout, an effect in line with censorship bias. When this happens, the probability of a further stockout increases, showing that censorship bias explains a portion of uninformed deviations from algorithmic suggestions. To suggest actionable policies, I collect more data to test the idea of blocking downward deviations when censorship bias is suspected. With this additional data analysis, I show that by blocking the downward deviations that are susceptible to censorship bias, retail managers can reduce self-inflicted stockouts. To understand this policy's impact more holistically, I predict a demand pattern and calculate the inventory holding cost implications of the policy. I also test the implications of two other policies that are built on the empirical results suggesting that large deviations increase the likelihood of future stockouts. The changes in the self-inflicted stockouts and inventory holding costs favor the policy of blocking deviations susceptible to censorship bias over the two benchmark policies.

In the second chapter, I analyze the augmentation decisions of store managers of the same supermarket chain in inventory replenishment decisions. Specifically, these store managers are empowered to augment ASR orders to ensure that demand is met and stockouts are prevented. Yet, I empirically show that store managers are susceptible to a novel bias when augmenting ASR orders before finalizing orders: newsvendor double-counting bias. This bias suggests that decision makers, especially practitioners, may apply the newsvendor logic mentally when they encounter an algorithmic suggestion as if this suggestion refers to a demand forecast. Yet, in the context of the usage of automated store replenishment systems, this would be a double-counting bias because the newsvendor optimization logic is applied to forecasts by the algorithm before those forecasts are converted into order proposals. I analyze ordering decisions of store managers of our collaborator supermarket chain by a probit model with sample selection model, a Heckman selection model, and an endogenous treatment regression model. All these models include exclusion restrictions adopted from the literature to tackle the endogeneity. I also

control for anchoring and supply line underweighting as benchmark biases to understand the relative significance of newsvendor double-counting bias. These analyses show that augmentation decisions triggered by newsvendor double-counting bias do not lead to additional sales. This shows that the discretionary power of the users to augment orders may lead to increased inventory holding costs and waste.

In the third chapter, my focus is on retail transformation, another contemporary topic for retailers. With the ubiquity of online shopping nowadays, many products are one click away, so retailers are looking for new ways to transform their brick-and-mortars to attract consumers back to physical stores. A way to do so is offering experiential services, such as featuring bars and cafes in grocery stores (e.g., Target), offering cooking classes in supermarkets (e.g., Whole Foods), or free workout classes in active wear shops (e.g., Lululemon). I am particularly interested in the spillover effects of such services to increase our understanding of the impact on retailers' main products. By collaborating with a supermarket chain from the United States, and combining their transaction-level data with point of interest data from SafeGraph, I investigate the effect of featuring a taproom service in grocery stores on sales revenue, sales quantity, number of transactions, and assortment size at the store level. To understand the mechanisms behind the impact I observe at the store level, I follow up with additional analyses on basket-level changes, heterogeneous department effects, sales changes based on specific product characteristics, and various customer-reach measures. Being the first empirical study on the spillover effects of offering an experiential service, this study generates important insights to the retailers who consider offering such services.

Introducción

El sector minorista desempeña un papel clave en el crecimiento económico mundial al proporcionar bienes que los consumidores demandan: empleos dentro de las comunidades locales y estimular el comercio nacional e internacional. Desde tiendas de barrio independientes hasta cadenas de supermercados multinacionales, los minoristas de comestibles tienen la importancia de mantener alimentada a la gente. A pesar de proporcionar bienes esenciales, el cambio es constante; por lo tanto, los minoristas de comestibles deben innovar, ser flexibles y estar preparados para el futuro para sobrevivir y prosperar durante los desafíos económicos e industriales globales. Los temas importantes y contemporáneos para los minoristas de comestibles incluyen la digitalización y la transformación del comercio minorista, que son los temas principales de esta disertación.

Dentro del contexto de la digitalización, mis dos primeros capítulos investigan cómo los supermercados utilizan los sistemas automáticos de reabastecimiento en tienda. En particular, en estos dos capítulos analizo por qué los gerentes de tienda de una cadena de supermercados se desvían de las sugerencias del sistema algorítmico en las decisiones de reabastecimiento de inventario. En el primer capítulo, estudio una paradoja en la que los gerentes de tienda de una cadena de supermercados se desvían hacia abajo de las propuestas de pedidos de un sistema de reabastecimiento automático de tiendas después de un desabastecimiento al reponer sus inventarios. Sostengo que este comportamiento de ordenamiento contraintuitivo puede explicarse por un sesgo de censura, y cuando este sesgo desencadena tales decisiones de desviación, tiene implicaciones

negativas para el desempeño. Para comprender la importancia relativa del sesgo de censura, agrego el sesgo de anclaje en mis modelos como punto de referencia. Abordo la endogeneidad de las decisiones de desviación estimando un modelo probit con selección de muestra, un modelo de regresión de conmutación endógena y un modelo probit bivariado recursivo con restricciones de exclusión. Los resultados de estos diversos modelos sugieren que la probabilidad de que los gerentes de tienda se desvíen a la baja aumenta después de un desabastecimiento, un efecto acorde con el sesgo de censura. Cuando esto sucede, la probabilidad de un mayor desabastecimiento aumenta, lo que demuestra que el sesgo de censura explica una parte de las desviaciones desinformadas de las sugerencias algorítmicas. Para sugerir políticas viables, recopilo más datos para probar la idea de bloquear las desviaciones a la baja cuando se sospecha un sesgo de censura. Con este análisis de datos adicional, muestro que al bloquear las desviaciones a la baja que son susceptibles al sesgo de censura, los gerentes minoristas pueden reducir los desabastecimientos autoinfligidos. Para comprender el impacto de esta política de manera más integral, predigo un patrón de demanda y calculo las implicaciones de la política en los costos de mantenimiento de inventario. También pruebo las implicaciones de otras dos políticas que se basan en resultados empíricos que sugieren que las grandes desviaciones aumentan la probabilidad de futuros desabastecimientos. Los cambios en los desabastecimientos autoinfligidos y en los costos de mantenimiento de inventarios favorecen la política de bloquear las desviaciones susceptibles de sesgo de censura sobre las dos políticas de referencia.

En el segundo capítulo, analizo las decisiones de aumento de los gerentes de tienda de una misma cadena de supermercados en decisiones de reabastecimiento de inventario. Específicamente, estos gerentes de tienda están facultados para aumentar los pedidos de ASR para garantizar que se satisfaga la demanda y se evite el desabastecimiento. Sin embargo, demuestro empíricamente que los gerentes de las tiendas son susceptibles a un sesgo novedoso cuando aumentan los pedidos de ASR antes de fi-

nalizarlos: el sesgo de doble conteo del vendedor de periódicos. Este sesgo sugiere que quienes toman decisiones, especialmente los profesionales, pueden aplicar mentalmente la lógica del vendedor de periódicos cuando encuentran una sugerencia algorítmica como si esta sugerencia se refiriera a un pronóstico de demanda. Sin embargo, en el contexto del uso de sistemas automatizados de reabastecimiento de tiendas, esto sería un sesgo de doble conteo porque el algoritmo aplica la lógica de optimización del vendedor de periódicos a los pronósticos antes de que esos pronósticos se conviertan en propuestas de pedido. Analizo las decisiones de pedido de los gerentes de tienda de nuestra cadena de supermercados colaboradora mediante un modelo probit con un modelo de selección de muestra, un modelo de selección de Heckman y un modelo de regresión de tratamiento endógeno. Todos estos modelos incluyen restricciones de exclusión adoptadas de la literatura para abordar la endogeneidad. También controlo la infraponderación del anclaje y de la línea de suministro como sesgos de referencia para comprender la importancia relativa del sesgo de doble conteo de los vendedores de periódicos. Estos análisis muestran que las decisiones de aumento provocadas por el sesgo de doble conteo de los vendedores de noticias no generan ventas adicionales. Esto muestra que el poder discrecional de los usuarios para aumentar los pedidos puede generar mayores costos de mantenimiento de inventario y desperdicio.

En el tercer capítulo, me centraré en la transformación del comercio minorista, otro tema contemporáneo para los minoristas. Con la ubicuidad de las compras en línea hoy en día, muchos productos están a un clic de distancia, por lo que los minoristas están buscando nuevas formas de transformar sus tiendas físicas para atraer consumidores a estas. Una forma de hacerlo es ofrecer servicios experienciales, como ofrecer bares y cafeterías en las tiendas de comestibles (por ejemplo, Target), ofrecer clases de cocina en supermercados (por ejemplo, Whole Foods) o clases de ejercicio gratuitas en tiendas de ropa deportiva (por ejemplo, Lululemon). Estoy particularmente interesada en los efectos indirectos de dichos servicios para aumentar nuestra comprensión del impacto en los

principales productos de los minoristas. Al colaborar con una cadena de supermercados de Estados Unidos y combinar sus datos a nivel de transacciones con datos de puntos de interés de SafeGraph, investigo el efecto de presentar un servicio de taberna en las tiendas de comestibles sobre los ingresos por ventas, la cantidad de ventas, el número de transacciones y tamaño del surtido a nivel de tienda. Para comprender los mecanismos detrás del impacto que observo a nivel de tienda, hago un seguimiento con análisis adicionales sobre cambios a nivel de canasta, efectos heterogéneos en los departamentos, cambios en las ventas basados en características específicas del producto y varias medidas de alcance al cliente. Al ser el primer estudio empírico sobre los efectos indirectos de ofrecer un servicio experiencial, este estudio genera conocimientos importantes para los minoristas que consideran ofrecer dichos servicios.

Chapter 1

Discretion in Automated Supermarket Replenishment: Censorship Bias and Self-inflicted Stockouts

1.1 Introduction

Automation of grocery replenishment processes is spearheading digital transformation in the retail sector (McKinsey 2020). The backbone technology for the use of artificial intelligence in replenishment decisions is an automated store replenishment (ASR) system designed to optimize the tradeoff between shelf availability and inventory holding costs (Angerer 2007). Despite the demonstrated effectiveness of ASR systems (Avlijas et al. 2015), human decision-makers typically have the freedom to override and deviate from the proposed algorithmic solutions. Allowing this discretion aims to utilize human decision-makers' private knowledge and insight that cannot be taken into account algorithmically. In the seminal study on ASR system use, van Donselaar et al. (2010) report that deviations from the proposed orders may result in performance improvement. However, discretion may also have unintended consequences because of the cognitive limi-

tations of decision-makers. Specifically, Sun et al. (2021) propose that deviations from algorithmic suggestions can be classified as either information deviations or complexity deviations. The former are triggered by human decision-makers' private information and insight, whereas the latter are triggered by their inability to understand algorithmic prescriptions and are, therefore, susceptible to decision-making biases and fallacies. If organizations were able to distinguish information deviations from complexity deviations, they would benefit from human insight and private information while minimizing the detrimental effect of decision-making biases and fallacies. In this paper, we propose a way to distinguish one form of uninformed deviation in the grocery replenishment context.

In the inventory replenishment context, deviations from the ASR proposals can be either upward or downward. The main reasons for upward deviations include expectations of abnormal peaks in demand as well as logistical reasons to advance replenishment shipments (van Donselaar et al. 2010). Meanwhile, downward deviations are made to avoid excessive inventory and spoilage when abnormal slumps in demand are expected. Aside from the different reasons, the risks associated with upward and downward deviations are different. Upward deviations may cause high levels of spoilage and inventory holding costs, whereas downward deviations may lead to stockouts. Despite these differences in reasons and risks linked to upward and downward deviations, only the former have been studied thus far (van Donselaar et al. 2010). Given how critical product availability is for customer satisfaction, revenue, and profitability in the retail sector (Gruen et al. 2002, Anderson et al. 2006), decisions to increase stockout risk by deviating downward from the ASR proposals constitute an intriguing phenomenon. This gets even more interesting for the perishable products because of the tradeoff between stockout and waste risks. User discretion in inventory decisions of perishable products has not been studied so far, although these products account for a considerable part of grocers' revenue. Hence, this study focuses on the antecedents of downward deviations from ASR proposals for perishable products and the performance implications of deviations, particularly when they are

marked by decision-making biases.

Since the valid reasons to deviate from ASR proposals are idiosyncratic and based on private information that is not recorded in the ASR system, our approach to studying the deviation phenomenon is to identify drivers of invalid, biased decisions to deviate from ASR proposals and then measure their detrimental performance implications. Our focus is on censorship bias (Feiler et al. 2013), which has thus far received little attention in the behavioral operations management literature. Censorship bias refers to a situation in which a decision-maker interprets sales as demand after a stockout, failing to consider the demand that was not fulfilled due to the stockout, i.e., lost sales. Such misinterpretation creates downward-biased demand expectations, potentially leading to situations in which, after a stockout, ASR system users perceive ASR proposals to be excessive and decide to order less than proposed in an attempt to avoid unnecessary inventory. Censorship bias has been examined in lab studies where subjects, on average, order significantly less under the censored demand condition (Feiler et al. 2013, Rudi and Drake 2014, Tong et al. 2018). Ordering less after a stockout is curious behavior since, logically, it could be expected that one would order more after a stockout. In the absence of field research, it is unclear whether behavioral biases exist in the actual orders of practitioners (Sachs et al. 2022). For example, one might expect that censorship bias is not so common in actual retail practice where the stakes are much higher than they are for lab subjects. Therefore, as an initial check of its existence, we turn to our data in Figure 1. The left part of the bar chart shows the percentages for upward and downward deviations in the absence of a recent stockout (within a week from the order placement). The right part of the chart illustrates the same when there has been a recent stockout. We observe that downward deviations are six percentage points (40%) more common after a stockout than when there has been no recent stockout. These numbers demonstrate the possibility that censorship bias transcends the conditions of lab studies. However, something else could be going on as well. One alternative explanation to ordering less than what the ASR

system proposes after a stockout is anchoring bias (Tversky and Kahneman 1974), which in contrast to censorship bias, has been extensively studied in the literature. Since it was first examined in the inventory replenishment context by Schweitzer and Cachon (2000), anchoring bias has been consistently observed in behavioral operations management research (e.g., Bostian et al. 2008, Bolton et al. 2012, Rudi and Drake 2014). We include anchoring bias in our analyses to test whether censorship bias has a distinct effect and to compare the effects of these two biases on downward deviations from ASR proposals and stockouts.

In addition to examining whether censorship bias exists in practice and how it compares to anchoring bias in triggering deviations from algorithmic suggestions, we also analyze the performance implications of the two biases. In particular, when ASR system users order less than the system's proposals after a stockout, do they do so to the extent that it causes operational harm in the form of subsequent stockouts? Although censorship bias has been observed in lab experiments (Feiler et al. 2013, Tong et al. 2018), field research is needed to understand whether it has any economically relevant performance implications. For example, even if censorship bias occurred in practice, the practitioners could be more cautious than the lab subjects and react to it so subtly that it might not cause operational harm. If it does cause stockouts, then how does that harm compare to the harm caused by the well-established anchoring bias?

To study these questions, we use data from a sample of stores of an upmarket European supermarket chain. Since we are interested in the stockout implications of downward deviations from ASR proposals and since the decisions to deviate are evidently endogenous, our primary econometric strategy relies on a probit model with sample selection (Van de Ven and Van Praag 1981), built on the Heckman selection model (1979). We develop the selection model to predict downward deviations based on the findings of van Donselaar et al. (2010) and Khosrowabadi et al. (2022). We compare the adverse effects of censorship bias and anchoring bias by comparing how the likelihood of a

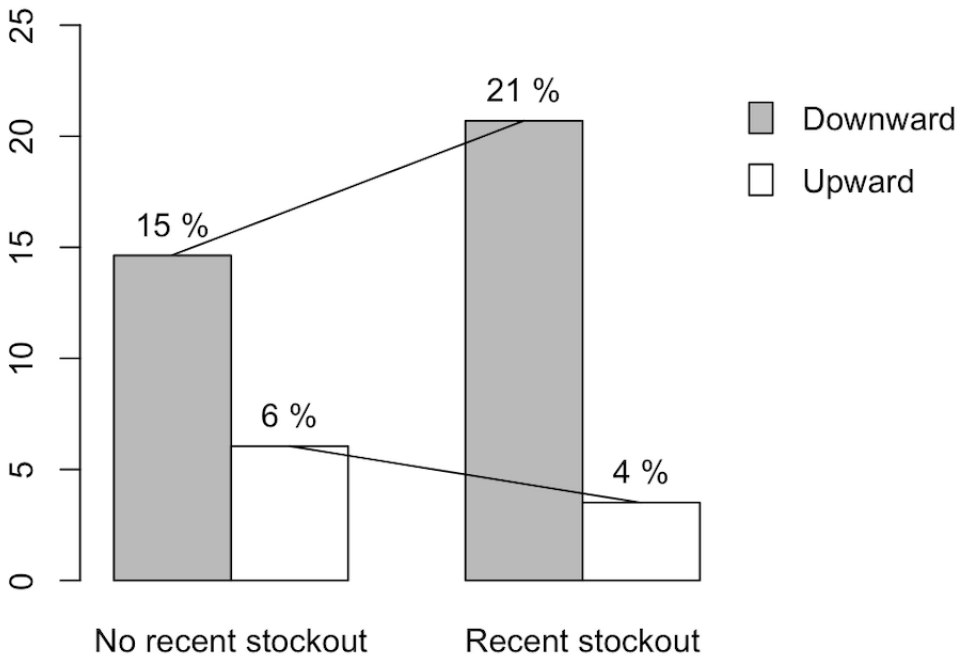

Figure 1.1: Deviation Frequencies

stockout changes when the deviation decisions are susceptible to these biases. Since it is possible that the driver of censorship bias, namely, a recent stockout, may directly increase the likelihood of a new stockout regardless of downward deviations, and since the probit model with sample selection uses only the subsample where downward deviations have been made in its second stage, we complement our analysis with an endogenous switching regression model (Lokshin and Sajaia 2011, Marra et al. 2022) and a recursive bivariate probit model (Maddala 1986) that use all of the cases in their second stage. In these analyses, we observe that the stockout effect is significantly greater when the ordering decision is aligned with censorship bias and a downward deviation is made after a recent stockout. Together, these models indicate that there is a causal link between recent stockouts, downward deviations, and subsequent stockouts and that these links can be explained by censorship bias. More specifically, the findings show that censorship bias explains downward deviations from ASR proposals, but its effect is not as strong as the effect of anchoring bias on the likelihood of a downward deviation. Regarding the

performance implications of downward deviations, our models show that the likelihood of a stockout increases more when a downward deviation is explained by censorship bias than when it is explained by anchoring bias. To explore the practical relevance of our findings, we collect additional data to test the performance implications of a policy of blocking the deviations susceptible to censorship bias. Our simulations in this additional dataset suggest that blocking downward deviations made after a recent stockout would help grocery stores decrease self-inflicted stockouts while minimally increasing inventory holding costs. This practical insight can be employed to further develop ASR software and to train ASR system users.

1.2 Literature Review

Organizations exploit the advancements in artificial intelligence, big data, and machine learning to assist human decision making in various contexts, from ordering (van Donselaar et al. 2010) to hiring (Kuncel et al. 2014). However, algorithmic assistance has limitations since its prescriptions sometimes lack the information that practitioners have (Campbell and Frei 2011). Therefore, decision-makers typically have the freedom to override algorithmic solutions so that they can incorporate their private information. For example, Elmaghraby et al. (2015) report that salespeople of grocery product distributors take into account a wide spectrum of factors when deviating from the price changes recommended by their algorithm. Campbell and Frei (2011) report that local managers at a large retail bank override the recommendations of a centralized capacity planning system to take market-specific customer sensitivities to service times into consideration. Phillips et al. (2015) estimate that local salespeople's adjustments to the headquarters' price list in auto lending pricing lead to profit improvements. However, discretion may have unintended consequences given the bounded rationality of decision-makers. For example, Caro and Saez de Tejada Cuenca (2022) report a negative revenue impact when users

deviate from the recommendations of a pricing support system in fashion retail. In radiological diagnoses, Ibanez et al. (2018) report a negative productivity effect for doctors' discretion over their task sequences. In the automobile industry, Kesavan and Kushwaha (2020) find that merchants' discretionary power results in subsequent reductions in profitability. Given that deviations from decision support systems are triggered by different mechanisms, Sun et al. (2021) propose a classification scheme for discretionary behavior, making a distinction between information deviations and complexity deviations. The former comprise the cases where decision-makers deviate because they have information that the algorithm has not utilized, thus more likely improving the decision outcomes, whereas the latter are caused by decision-makers' failure to understand algorithmic recommendations, resulting more often in worse outcomes.

Humans' hesitancy to follow algorithms (Meehl 1954) may be one explanation for complexity deviations. Research shows that such hesitancy does not hinge on the context; instead, decision-makers exhibit it in financial forecasts (Önkal et al. 2009), legal decisions (Eastwood et al. 2012), operating room management in hospitals (Prahl and Van Swol 2017), employee selection (Kuncel et al. 2014), and even in predicting the funniness of a joke (Yeomans et al. 2019). This tendency to prefer human judgment over decision support system recommendations has been dubbed algorithm aversion by Dietvorst et al. (2015). One mechanism triggering algorithm aversion is that human decision-makers typically perceive an algorithm as a "black box" (Yeomans et al. 2019). Since they cannot completely comprehend processes inside algorithms, the perceived lack of control makes them uncomfortable. The resulting reluctance to follow algorithms has been observed to be stronger among practitioners than student subjects (Logg et al. 2019), making the phenomenon relevant for field research on the use of commercial decision support systems whose algorithms are black boxes due to intellectual property reasons.

While most studies on algorithm aversion have focused on fairly straightforward forecasting tasks (Önkal et al. 2009, Dietvorst et al. 2015, Prahl and Van Swol 2017, Dietvorst

et al. 2018), there has been a recent interest in more comprehensive tasks (Snyder et al. 2022). For example, Castelo et al. (2019) investigate algorithm aversion when subjects receive dating advice versus financial advice from a human decision-maker versus from an algorithm. Germann and Merkle (2020) examine algorithm aversion in financial decision making by analyzing subjects' tendency to follow a human fund manager or an investment algorithm. In the retail context, an important activity that expands on the plain forecasting task is replenishment ordering, which in itself is a widely studied subject – albeit typically not in conjunction with the algorithm aversion phenomenon. Instead of adherence to any algorithm, the core interest in the existing behavioral research on replenishment ordering has been in decision-makers' adherence to the optimal order quantities of the newsvendor model.

Behavioral issues that may affect ordering decisions in the newsvendor setting were first examined by Schweitzer and Cachon (2000). In that study, subjects demonstrated a strong anchoring bias (Tversky and Kahneman 1974), and although they adjusted away from the anchor toward the optimal order quantities, their adjustments were insufficient, and the resulting decisions were suboptimal (Schweitzer and Cachon 2000). That study and a long stream of subsequent investigations have concentrated on two anchor points for the ordering decisions: the mean demand (giving rise to the term "pull-to-center bias," e.g., Bostian et al. 2008) and the previously placed replenishment order (Lau and Bearden 2013). These are logical anchor points in lab studies, where they are typically the most visible pieces of information to the subjects and where the demand is often assumed to follow a uniform distribution (e.g., Lau and Bearden 2013). In retail practice, however, the demand is seldom stationary (Ali and Gürlek 2020), full demand information is not necessarily available (but instead can be censored, Feiler et al. 2013), and the daily demand figures do not correspond directly to any replenishment order quantities due to variability in order coverage periods (i.e., days between consecutive replenishments). While these specific complexities may discourage anchoring on past demand, it has also been ob-

served that when faced with complexity, decision-makers in an inventory ordering task are inclined to select an anchor closer to the optimal solution (Gavirneni and Xia 2009). In the grocery replenishment context, such an anchor might be past purchase orders, to which newsvendor logic has already been applied. However, even for that anchor, practical complications arise from the retail sector's prevalent weekday seasonality patterns (van Donselaar et al. 2010), potentially diminishing the temptation to anchor directly on the previous order, which may have been placed to cater to the demand of different days of the week. In such situations, decision-makers may develop a sense of what has been their average or usual order size over time and use it to anchor their decisions (Wansink et al. 1998). In the grocery replenishment context, a weekly average of past purchase orders could work for that purpose, as it would smooth over the weekday seasonality effects.

Apart from the less evident anchor points, another difference between retail practice and most experimental research on newsvendor decision making is the availability of ASR system-based decision support. In behavioral newsvendor experiments, subjects typically do not receive algorithmic suggestions; thus, the notion of a deviation in those studies is the difference between the placed order and a newsvendor model solution that is unknown to the subjects. One exception to this is Lee and Siemsen (2017), who provide subjects with algorithm-based normative order quantities. They conclude that anchoring occurs even in the presence of such suggestions. Behavioral newsvendor literature, just like algorithm aversion literature, has explored potential differences between practitioners and student subjects, ultimately concluding that both are equally susceptible to anchoring bias (Bolton et al. 2012). These observations mean that the anchoring bias must be controlled for, and that it provides a benchmark for the assessment of the magnitude of the effect of other biases.

In this paper, we complement van Donselaar and others' (2010) study on the antecedents of order advancements of ASR proposals with an investigation into the an-

tecedents of downward deviations. We also extend the investigation to the performance implications of deviations, which enables us to distinguish between the antecedents of informed and uninformed deviations. In this sense, we build on the work of Sun et al. (2021), who studied informed and uninformed deviations, calling them information and complexity deviations. We shift the perspective from both earlier studies' product-related antecedents to antecedents derived from the behavioral decision-making literature. Using anchoring bias as our benchmark, we focus on censorship bias, whose existence and effects have not yet been tested in actual retail practice and in the presence of algorithmic suggestions.

1.3 Hypotheses

1.3.1 Censorship Bias in Grocery Retail Replenishment

Censorship bias refers to a tendency "to form beliefs about the underlying population that are biased in the direction of the observed censored sample" (Feiler et al. 2013, p. 574). A classic example of censoring is the relationship between demand and sales. Typically, decision-makers cannot observe lost sales after a stockout, leading them to infer demand from sales figures that have been constrained by unavailable inventory (Tong et al. 2018). This means that stockouts today may distort demand expectations for tomorrow. Rudi and Drake (2014) used the term *observation bias* to discuss this type of distorted inference under censored demand. In laboratory experiments, subjects show downward-biased demand beliefs when demand data are censored by a stockout, leading to lower inventory order decisions (Feiler et al. 2013, Rudi and Drake 2014, Tong et al. 2018). Ordering less after a recent stockout is counterintuitive behavior since one might expect decision-makers to order more after a stockout to avoid future stockouts. However, interpreting sales up until the stockout as all of the demand that had existed would be in line with

the boundedly-rational human inclination to take mental shortcuts (Bazerman and Moore 2012). Censorship bias would thus be an outcome of an unintentional mental shortcut to avoid the cognitive burden of trying to account for the unknown lost sales that followed the stockout.

Despite the results of the lab experiments, it is not obvious that censorship bias exists in retail practice. On the one hand, practitioners can be expected to be more cognizant of the lost sales concept (Bell and Luddington 2006) and its many detrimental performance effects than the student subjects of lab studies. Additionally, due to the actual performance effects, the stakes are evidently higher for practitioners, making them possibly more cautious in their decision making. On the other hand, censorship bias may well exist because practitioners have to deal with hundreds of stock-keeping units (SKUs) on a daily basis and, thus, bear a much greater cognitive burden than lab subjects. Research has shown that decision-makers dealing with many SKUs are more susceptible to biases in the newsvendor setting (Chen and Li 2018). This type of effect can be pronounced in retail practice because ASR systems should not suffer from censorship bias, at least not as much as their human users do, since drastic progress has been made in the methods for forecasting and ordering under censored demand (Shi et al. 2016). The basic principle underlying these methods is to forecast lost sales by extrapolating the demand rate that preceded the stockout. When the ASR system does that, the resulting proposals may be perceived as excessive by the users, who are susceptible to censorship bias. To study whether censorship bias triggers downward deviations, we make use of the condition under which censorship bias can exist, that is, a recent stockout, and we hypothesize:

Hypothesis 1: The likelihood of a downward deviation from an ASR proposal is higher when an SKU has had a recent stockout.

1.3.2 Performance Implications of Censorship Bias: Self-inflicted Stockouts

Retailers allow discretion for ASR system users to benefit from their private information, yet this discretionary power brings the risk of uninformed decisions that are marked by biases. We propose that censorship bias offers a way to detect one category of uninformed downward deviations. To test this idea, we are interested in the performance implications of deviations that are susceptible to censorship bias. The underlying assumption is that uninformed deviations are more likely to result in negative outcomes: stockouts in the case of downward deviations from ASR proposals.

Despite the fact that any downward deviation from an ASR proposal increases the stockout risk, some of these deviations may be informed. In fact, in forecasting tasks, research has shown that negative adjustments, where managers decrease the quantity of algorithmic forecasts, tend to increase the forecasting accuracy (Fildes et al. 2009). This is because downward deviations are, on average, more often informed, as they must overcome human risk aversion and overoptimism (Fildes et al. 2009, Hewage et al. 2022). Similarly as in the forecasting context, in packaging tasks, Sun et al. (2021) interpret downward deviations (switching to smaller boxes) as informed deviations. Therefore, in the replenishment ordering context, a significant part of downward deviations from ASR proposals may be informed and beneficial. To identify uninformed decisions, it makes sense to focus on the information that boundedly rational decision-makers are missing (Gavetti et al. 2012). The information they lack after a stockout is lost sales, which is a crucial piece of information because its absence distorts the baseline for future forecasts (Gruen and Corsten 2007). We thus believe that focusing on recent stockouts allows us to detect a portion of the uninformed downward deviations from ASR proposals that are driven by boundedly rational human nature and thus likely to be detrimental to performance. When a downward deviation occurs after a recent stockout, which is the condition

in which censorship bias exists, we expect it to have a higher likelihood of future stockouts.

Hypothesis 2: The likelihood of a stockout is higher when a downward deviation from an ASR proposal is made after a recent stockout.

1.4 Empirical Context and Methods

1.4.1 Setting and Data

We examine our research questions by collaborating with a retail optimization software provider that is recognized as the industry leader by software vendor rankings (JCMR 2022). We collected two datasets from one of this software provider's customers, an upmarket European supermarket chain. Following the approach of van Donselaar et al. (2010), we used the first dataset to test our hypotheses and the second to test a policy based on the results of the hypothesis testing. Two data collection periods are separated by one year, and each dataset involves four months of ordering decisions. The studied supermarket chain had implemented the ASR system more than three years prior to the first data collection period. Like in van Donselaar et al. (2010), our first dataset is from three store locations of the studied supermarket chain, while the second dataset is from two other locations. All five stores are in different cities, which are the five largest cities of the primary market of the studied supermarket chain. In this chain, store personnel's interactions with customers provide them with private information on demand signals that may make discretionary power useful. If these interactions signal a decreasing demand for an SKU, store personnel can deviate downwards from ASR proposals, particularly in perishable product categories, despite the increasing risk of stockouts. Of course, a stockout may occur even without a downward deviation. However, if a stockout occurs after a downward deviation, it presumably occurs earlier than it would have otherwise occurred and worsens the lost sales. This is what we call a *self-inflicted stockout*. Although

inventory and spoilage minimization are essential for operational efficiency, stockouts are a source of great concern for grocery retailers because they lead to customer complaints (Bell and Luddington 2006), loss of intended purchases (Gruen et al. 2002), and reduced profits (Anderson et al. 2006), potentially in the long term, as customers are driven to try competitors' products (Gruen and Corsten 2007). The negative implications of stockouts are an even greater concern for upmarket retailers for which customer satisfaction is vital. Thus, it is imperative to discern downward deviations tainted by biases from those triggered by decision-makers' private information.

The phenomenon of interest of this study should manifest itself most prominently in perishable products because our interactions with grocery retail professionals indicate that, due to spoilage concerns, users of ASR systems are particularly alert to conducting downward deviations in the case of perishable SKUs. Additionally, past research has associated product perishability with suboptimal ordering decisions (Bloomfield and Kulp 2013). In addition, perishable products account for a considerable part of grocers' revenue, yet previous field research has concentrated only on nonperishables (e.g., van Donselaar et al. 2010). We focus on two perishable product categories: packaged meat products and ready-made meals that are not frozen. This choice stems from several reasons. First, the SKUs of these product categories are handled in integer piece quantities, whereas the SKUs of many other perishable product categories are handled in weight units (e.g., fruits, fresh meats, and fish). Records in integer piece quantities make the operationalization of stockouts unambiguous. Second, we also wanted to avoid promotional effects in this investigation and collected a sample that contains no SKUs on promotional campaigns. Avoiding campaigns would have been harder in many other perishable product categories, such as dairy products. Third, these two product groups happened to be of great interest to the supermarket chain with which we collaborated. From these product categories across the three locations in the data used for hypothesis testing, we have 472 SKUs, leading to 1,416 different SKU-location combinations. These data provide us with

SKUs that are different in many characteristics, such as margin, demand volatility, substitutability, forecasting error, and on-hand inventory. Hence, we are able to disentangle the impact of product characteristics from that of censorship bias and to control for previously reported antecedents of deviations from ASR proposals (van Donselaar et al. 2010).

We collected data from two periods that are separated by one year, each period containing 12 weeks of replenishment decisions, including 57 days when ASR proposals were created. On average, an SKU has 2.5 such days per week, ranging from 1 to 5. The units of analysis are the daily decisions per SKU with a nonzero ASR proposal in each store. We focus on nonzero ASR proposals, given that we are interested in downward deviations; yet the robustness of the results will be tested with data including zero proposals. In total, the data consist of 11,434 nonzero ASR proposals. The performance outcome we are interested in is whether a stockout occurs in the period between the arrival of the ordered replenishment and the next possible replenishment, the so-called order coverage period. This period ranges from 1 to 7 days in the data. The lead times from placing the order to replenishment are known and reliable, ranging from 2 to 4 days. The expected lead time and actual lead time are different for only 267 out of 11,434 cases, suggesting that unavailability of supply could not be an alternative explanation to the studied phenomenon. We planned the data collection to exclude all major public holidays to avoid abnormal patterns, yet it was impossible to avoid one minor (Saturday) holiday. We control for its potential impact in our models.

In each store location, there is one designated ASR system user for each product group. The ASR system produces proposals based on time-series forecasting and newsvendor optimization algorithms taking into account the current inventory level, both of which are a black box to the user, which is typically the case in studies on the use of advanced decision support systems (e.g., Sun et al. 2021). Each designated user views the ASR proposals for all SKUs within their product group and either adjusts them or confirms them as they are. A downward deviation is recorded when the final purchase

order quantity is less than the ASR proposal. It is worth mentioning that the screen layout uses the term “sales”, not “demand,” when referring to the past sales; hence, if users are susceptible to censorship bias, it is not due to the design of the ASR system. The users in our setting are not constrained by any budget restrictions when adjusting and confirming their orders. Neither do they receive any individual performance-based incentives. They receive a fixed salary, and the only variable component of their pay is a bonus based on annual store profits. Thus, they are incentivized to achieve profit maximization, which is in line with the incentives of the supermarket chain. One limitation of this dataset is that the individual characteristics of the designated ASR system users are unknown. However, by controlling for the fixed effects of store locations and product groups, as well as their interactions, we alleviate the potential impacts of this limitation. Since there is one designated user for each product group in every location, the controls capture the individual traits that may affect the deviation behavior, such as risk preferences or cognitive abilities. Another limitation is that, like other studies on worker discretion in the use of decision support systems (e.g., Elmaghraby et al. 2015, Kesavan and Kushwaha 2020), we cannot report the proprietary algorithms employed by the ASR system due to their confidentiality. Although these algorithms are unknown to us, our discussions with the software provider indicate that this ASR system has no bias towards understocking or overstocking; hence, it does not over penalize spoilage over stockouts, or vice versa.

1.4.2 Measures

Dependent Variables: The main dependent variables of this study are downward deviation decisions and stockouts. *Downward Deviation* is a binary variable indicating that the final purchase order quantity is less than the ASR proposal. *Stockout* is a binary variable indicating that the inventory count goes to zero during the order coverage period (i.e., the interval between the arrival of the ordered replenishment and the next possible replenishment).

Independent Variables: Since censorship bias can only exist after a stockout, we analyze it with a binary variable called *Recent Stockout* that takes the value of 1 if the inventory of an SKU has gone to 0 in the same location in the previous seven days (alternative time windows are used in Section 1.6.1). For both hypotheses, the expected coefficient of this variable is positive.

To evaluate the relative prevalence and implications of censorship bias, we compare its impact with that of anchoring bias (Tversky and Kahneman 1974). We create several alternative operationalizations to capture anchoring bias, including variables based on past mean sales. The operationalization used in the main analysis is based on past purchase orders because of their saliency to the ASR system users and because the effect of that operationalization is the strongest. In this way, we obtain the most conservative estimate for the relative significance of censorship bias. Like *Recent Stockout* serving as a driver of censorship bias, we operationalize a driver of anchoring bias in the form of *Relative Proposal Size* to the mean of past seven (alternative time windows are used in Section 1.6.1) days' purchase orders for the SKU in the same location:

$$\frac{ASROrderProposal_{kj}}{MeanPastPurchaseOrders_{kj}} - 1$$

where k and j represent the SKU and location, respectively. We winsorize this variable at the 1% tails. The expected coefficient of this variable in predicting downward deviations is positive: when a proposal is greater than the past purchase orders, anchoring bias will lure the decision-maker to consider it excessively high. Notably, this may happen after a stockout if the ASR system, after estimating the lost sales, considers the past order level insufficient and produces a higher proposal. Therefore, it is important to control for this effect to be able to appropriately estimate the effect of censorship bias. In predicting a stockout with *Relative Proposal Size*, we also expect a positive coefficient because downward deviations, which are made when the driver of anchoring bias is strong, are less likely to be informed and, thus, more likely to cause self-inflicted stockouts. We note

that much of the earlier inventory management literature on anchoring bias has studied it as part of a heuristic that comprises first anchoring and then insufficient adjustment toward the optimum (e.g., Schweitzer and Cachon 2000), yet we ignore the insufficient adjustment part, as it would only rescale the variable and produce substantively the same coefficient estimates.

Controls: The control variables we use are *Proposal Size* and *Deviation Size*. We include *Proposal Size*, winsorized at the 1% tails, to predict downward deviations and stockouts. We expect a positive coefficient in predicting downward deviations because as an ASR proposal gets larger, users' inclination to deviate downward would increase. It is important to note that the effect of *Proposal Size* is different from that of *Relative Proposal Size* (associated with anchoring bias), which reflects the reaction to a large or small ASR proposal relative to past purchase orders. By controlling for *Proposal Size*, we capture the reaction to seeing a large number as an ASR proposal and its saliency. The expected coefficient of *Proposal Size* is negative in predicting a stockout since a large ASR proposal leads to a large purchase order quantity, as long as there is no downward deviation, whose effect we capture with *Deviation Size*. In addition to capturing user's reaction to seeing a large or small ASR proposal, *Proposal Size* accounts for the impact of previous deviations. For example, there could be a carryover effect, such that after deviating upward from an ASR proposal, users may tend to deviate downward in the next decision (i.e., order advancement of van Donselaar et al., 2010). Although this will be further examined in Section 1.6.2, controlling for *Proposal Size* already mitigates this concern because the inventory level caused by the previous deviation decision is considered by the ASR system when it creates the next proposal.

Deviation Size is the absolute value of the difference between the ASR proposal and the purchase order. It only exists when there is a downward deviation and is, therefore, only used to predict the likelihood of a stockout. We control for *Deviation Size* because

we are interested in the stockout effect of the reason that drives downward deviations, namely, censorship bias. Controlling for *Deviation Size* allows us to obtain a more purified effect of *Recent Stockout*. Another reason to control for *Deviation Size* is to eliminate the confounding effects of our exclusion restrictions. As will be explained in Section 1.4.3, the probit model with sample selection requires us to find variables that have an impact on downward deviation decisions and that do not correlate with the error term of the outcome equation. By controlling for *Deviation Size* in the second stage, we are able to account for the effects of the exclusion restrictions on the likelihood of a stockout. This is because as the effects of valid exclusion restrictions intensify, it is not only ASR system users' inclination to deviate downward that increases but also their inclination to make larger deviations. Thus, the effect of the exclusion restrictions on the likelihood of a stockout is carried by the effect of *Deviation Size*. The expected coefficient of *Deviation Size*, winsorized at the 1% tails, is positive.

Adding *Proposal Size* and *Deviation Size* together as control variables also allows us to control for the demand uncertainty of an SKU, and decision-makers' underestimation of this uncertainty. Specifically, if an SKU has a highly volatile demand, the ASR system suggests a higher *Proposal Size*. Capturing demand fluctuations' impact is important also to account for the effects of our exclusion restrictions on the likelihood of a stockout. For example, though *Hot Day* and *Cold Day* would affect *Deviation Size*, they may affect stockout through demand fluctuations. By adding *Proposal Size*, we are able to alleviate such concerns and to be more confident in how we tackle endogeneity. As for users' possible underestimation of variance around demand, we argue that *Deviation Size* gets higher as this behavioral pattern gets more prevalent.

We use fixed effects to control for locations, product groups, the interactions of locations and product groups, proposal weekdays, and weeks. The reason for adding interactions of locations and product groups is to capture the ASR system users' characteristics. As explained before, there is one designated user for each product group in each location.

Therefore, the interactions of these two fixed effects capture the individual traits of each designated user. We control for seasonality and other time-based effects with the fixed effects for proposal weekdays and weeks.

Exclusion Restrictions: In the examination of our hypotheses, the ASR system users' decisions to deviate downward are endogenous. To account for this endogeneity when estimating the performance effects of the deviations, we make use of the probit model with sample selection (Van de Ven and Van Praag 1981). The exclusion restriction of the selection model requires us to include at least one additional variable in the selection equation where we predict downward deviations; therefore, we turn to the literature predicting discretionary behavior in ordering and forecasting decisions. In particular, we use the variables studied by van Donselaar et al. (2010) and Khosrowabadi et al. (2022).

We measure *Case Pack Coverage* the same way as van Donselaar et al. (2010): the ratio of an SKU's case pack size to the average weekly sales of that SKU. We winsorize this variable at the 1% tails. The expected effect is negative because the larger the case pack coverage is, the more abrupt are the implications of any deviation. Since downward deviations entail a stockout risk, large case pack coverage may have a discouraging effect on making any deviation. Our measure of *On-hand Inventory*, winsorized at the 1% tails, is the ratio of the decision day's ending inventory of an SKU to the weekly average of the daily end inventory balances of that SKU. It is effectively a reversed measure of the net shelf space studied by van Donselaar et al. (2010). We could not use precisely the same measure since our data do not contain fixed shelf space allocations for the SKUs. The expected effect is positive, meaning that extensive on-hand inventory encourages downward deviations. Our measure of *Item Size* is the area defined by the width and height of the package since that was how the size was defined in our data, instead of the three-dimensional volume studied in van Donselaar et al. (2010). Since van Donselaar et al. (2010) hypothesized and showed that store managers prefer to receive

larger items earlier, we expect size to be negatively related to downward deviations. We measure *Margin* similarly as van Donselaar et al. (2010) as the absolute profit margin of the SKU, but we also tested that all the effects would remain the same had we used the percentage unit profit margin instead. The expected effect is negative, meaning that decision-makers are less willing to accept the stockout risk associated with downward deviations when the SKU has a higher profit margin. We winsorize this variable at the 1% tails. The measure of *Variety* is the number of other SKUs in the same subgroup of products, and the expected effect is negative since wide variety is typically a response to lower substitution (van Donselaar et al. 2010), leading to higher lost sales in the case of a stockout and hence reducing ASR system users' inclination toward downward deviations. To analyze the effect of demand uncertainty, van Donselaar et al. (2010) studied two variables: *Seasonality Error* and *Forecast Dispersion*. We replicate their approach by measuring the former as the root mean squared error of the differences between the last seven days' daily demands and the seasonality pattern predicted by the past demand. The latter is measured as the weekly standard deviation of the daily forecast errors from a trend-based forecast model divided by average sales. As uncertainty encourages buffering, we expect a negative effect for both variables. We winsorize *Seasonality Error* and *Forecast Dispersion* at the 1% tails.

From Khosrowabadi et al.'s (2022) exploration of forecast adjustments, we employ *Hot Day*, *Cold Day*, *Sales*, *Price*, *Discount*, *Day Before Holiday*, and *Holiday*. We note that the authors did not speculate on the expected directions of these variables in predicting deviations from algorithmic forecasts; therefore, we also only report the operationalizations of the variables here. *Hot Day* and *Cold Day* are binary variables indicating whether any day within the order coverage period belongs to the 10% hottest or the 10% coldest days of the past 30 years. *Sales* is calculated as the sales of the decision day relative to the mean of the last week's sales. *Price* is the price of the SKU in the corresponding store on the decision day. We winsorize *Sales* and *Price* at the 1% tails. *Discount* is a binary

variable taking the value of 1 if an SKU will be discounted during the order coverage period. We have only one holiday in the data by design. To control for its impact on deviation decisions, we create two variables following the approach of Khosrowabadi et al. (2022): *Day Before Holiday* and *Holiday*. *Day Before Holiday* is a binary variable taking the value of 1 if the order coverage period includes the day before the holiday. *Holiday* is a binary variable as well, taking the value of 1 if the coverage period includes the holiday.

1.4.3 Empirical Strategy

ASR system users' decisions to deviate downward are endogenous. Specifically, decisions to deviate downward create a selection bias in the examination of stockouts. Additionally, some unobserved factors may affect both downward deviation decisions and stockouts (causing omitted variable bias). Since we need to mitigate the endogeneity concerns and to control for Deviation Size as explained in Section 1.4.2, we employ a probit model with sample selection (Van de Ven and Van Praag 1981) that is built on the work of Heckman (1979) and uses full information maximum likelihood estimation (Maddala 1986). The Heckman selection model has been recently used by Caro and Saez de Tejada Cuenca (2022) in the examination of deviations from algorithmic recommendations in retail pricing. We complement our analysis with other models, including endogenous switching regression model (Amemiya 1984) and recursive bivariate probit model (Maddala 1986). Our probit model with sample selection includes two equations: one for the selection (downward deviation) and one for the outcome (stockouts). The decision to deviate downward from the ASR proposal is binary. The probit model to estimate it takes the following form:

Table 1.1: Descriptive Statistics and Correlation Table

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
mean	0.157	0.057	0.177	1.011	9.942	4.415	0.91	15.488	0.371	6.49	0.152	1.213	1.234	2.633	0.002	0.152	0.356	0.06	0.039	7.67
sd	0.364	0.232	0.382	1.334	8.305	4.988	0.729	14.102	0.309	5.276	0.141	1.461	2.209	1.065	0.04	0.359	0.479	0.237	0.194	4.257
min	0	0	0	-0.909	2	0.024	0	5	-0.146	0	0	0	0	0.857	0	0	0	0	0	1
max	1	1	1	5	40	36	5.529	60	1.425	21	0.847	11.757	20.801	6.93	1	1	1	1	1	16
1 Downward Deviation																				
2 Stockout	0.324 *																			
3 Recent Stockout	0.063 *	0.173 *																		
4 Relative Proposal Size	0.235 *	0.101 *	0.052 *																	
5 Proposal Size	0.115 *	-0.059 *	-0.159 *	0.005																
6 Case Pack Coverage	0.073 *	0.044 *	0.035 *	0.174 *	-0.161 *															
7 On-hand Inventory	-0.039 *	-0.055 *	0.107 *	-0.228 *	-0.072 *	-0.046 *														
8 Item Size	0.004	-0.072 *	-0.160 *	-0.114 *	0.239 *	0.153 *	-0.001													
9 Margin	-0.029 *	-0.072 *	-0.169 *	-0.161 *	0.180 *	-0.032 *	-0.006	0.350 *												
10 Variety	-0.071 *	-0.026 *	-0.039 *	-0.025 *	-0.028 *	-0.003	-0.011	0.152 *	0.061 *											
11 Seasonality Error	0.018	0.061 *	0.085 *	0.093 *	-0.157 *	0.065 *	0.048 *	-0.070 *	-0.079 *	0.098 *										
12 Forecast Dispersion	-0.003	-0.019 *	0.003	0.041 *	-0.058 *	0.035 *	-0.001	-0.053 *	-0.055 *	-0.050 *	0.002									
13 Sales	-0.016	0.010	-0.004	0.116 *	-0.076 *	0.112 *	0.067 *	-0.048 *	-0.056 *	-0.006	0.097 *	0.018								
14 Price	0.046 *	-0.012	-0.023 *	0.004	-0.073 *	0.148 *	-0.023 *	0.243 *	0.581 *	0.025 *	-0.009	-0.025 *	0.010							
15 Discount	0.043 *	0.028 *	-0.001	0.003	0.004	0.031 *	0.021 *	0.000	0.020 *	-0.010	0.001	-0.005	0.002	-0.036 *						
16 Cold Day	0.033 *	0.047 *	0.082 *	0.037 *	-0.030 *	0.041 *	0.021 *	-0.005	-0.040 *	-0.033 *	0.133 *	-0.018	0.042 *	-0.004	-0.005					
17 Hot Day	0.034 *	0.002	-0.020 *	-0.044 *	0.089 *	0.040 *	-0.039 *	0.075 *	0.085 *	0.008	-0.012	0.017	-0.044 *	0.041 *	0.007	-0.246 *				
18 Day Before Holiday	-0.048 *	-0.032 *	0.009	-0.014	0.032 *	0.026 *	-0.043 *	0.007	0.008	0.004	-0.064 *	0.082 *	-0.029 *	0.006	-0.001	-0.107 *	0.157 *			
19 Holiday	-0.021 *	-0.028 *	0.002	-0.026 *	0.054 *	0.047 *	-0.026 *	0.027 *	0.030 *	-0.004	-0.062 *	0.051 *	-0.004	0.016	0.003	-0.085 *	0.271 *	0.800 *		
20 Deviation Size	NA	-0.030	-0.123 *	0.077 *	0.751 *	-0.053 *	-0.035	0.243 *	0.126 *	-0.044	-0.072 *	0.016	-0.058 *	-0.166 *	-0.012	-0.008	0.070 *	0.045	0.059 *	

Note:

* $p < 0.05$

Descriptive statistics and correlation vector of Deviation Size variable are calculated for the sample where downward deviation is 1.

$$
\begin{aligned}
DownwardDeviation_i^* = \beta_0 &+ \beta_1 ProposalSize_i \\
&+ \beta_2 CasePackCoverage_i + \beta_3 OnhandInventory_i \\
&+ \beta_4 ItemSize_i + \beta_5 Margin_i + \beta_6 Variety_i \\
&+ \beta_7 ForecastDispersion_i + \beta_8 SeasonalityError_i \\
&+ \beta_9 Sales_i + \beta_{10} Price_i + \beta_{11} Discount_i \\
&+ \beta_{12} ColdDay_i + \beta_{13} HotDay_i \\
&+ \beta_{14} DayBeforeHoliday_i + \beta_{15} Holiday_i \\
&+ \beta_{16} RelativeProposalSize_i + \beta_{17} RecentStockout_i \\
&+ \boldsymbol{B X_i} + \epsilon_i \\
DownwardDeviation_i = \mathbb{I}&[DownwardDeviation_i^* > 0]
\end{aligned}
$$

In this first stage, the effects of the exclusion restrictions from van Donselaar et al. (2010) and from Khosrowabadi et al. (2022) are represented from β_2 to β_8, and from β_9 to β_{15}, respectively. In the second stage, the outcome variable is stockout, also a binary variable. We model it through a probit model defined by a latent variable, which takes the following form:

$$
\begin{aligned}
Stockout_i^* = \gamma_0 &+ \gamma_1 ProposalSize_i + \gamma_2 DeviationSize_i \\
&+ \gamma_3 RelativeProposalSize_i + \gamma_4 RecentStockout_i \\
&+ \boldsymbol{\Gamma X_i} + \xi_i \\
Stockout_i = \mathbb{I}&[Stockout_i^* > 0]
\end{aligned}
$$

where i represents each ASR proprosal, $DownwardDeviation_i^*$ and $Stockout_i^*$ are continuous latent variables for downward deviation and stockouts ($DownwardDeviation_i$ and $Stockout_i$), and $\mathbb{I}[.]$ is the indicator function. To account for endogeneity in the downward deviation decision ($DownwardDeviation_i$), the probit model with sample selection allows ξ_i to be correlated with the unobservable factors affecting a downward deviation (ϵ_i in the

selection equation) with a correlation coefficient ρ by assuming (ϵ_i, ξ_i) are standard bivariate normal. When ρ is not 0, in other words, when endogeneity is suspected, this model provides consistent and asymptotically efficient estimates for all the parameters (Maddala 1986). To obtain unbiased estimates, we include exclusion restrictions in the selection equation. These restrictions are variables akin to instruments (Certo et al. 2016), affecting the downward deviation decision (i.e., relevance criterion) but not having an effect on ξ_i (i.e., exclusion criterion).Thus, these variables do not appear in the second stage. As described in Section 1.4.2, to identify valid exclusion restrictions, we rely on the works of van Donselaar et al. (2010) and Khosrowabadi et al. (2022) that examine adjustments to automated ordering and forecasting decisions. Finally, both equations include locations, product groups, proposal weekdays, weeks, and the interactions of locations and product groups fixed effects, represented by $\boldsymbol{X_i}$. The interactions of locations and product groups' fixed effects alleviate the concerns that may emanate from the limitation that we do not have individual-level data of the users. The supermarket chain indicated that designated store personnel rarely change, so we are confident that any individual tendencies to deviate are controlled by these fixed effects.

1.5 Results

1.5.1 Determinants of Deviations

To investigate whether *Recent Stockout* drives downward deviations from ASR proposals, Table 1.2 presents the estimation results of the first stage with robust standard errors. We also report the pseudo-R^2 as it is useful to assess the strength of exclusion restrictions in models based on the Heckman selection approach (Certo et al. 2016). We observe that not all exclusion restrictions borrowed from van Donselaar et al. (2010) and Khosrowabadi et al. (2022) are significant in predicting downward deviations. When we remove

the insignificant exclusion restrictions from the specification, the significance level and the sign of the estimated coefficient of *Recent Stockout* (censorship bias) do not change in either stage. Therefore, we report the estimations from the model with all exclusion restrictions. From van Donselaar et al. (2010), *On-hand Inventory*, *Item Size*, *Margin*, and *Variety* are significant and in the expected direction when predicting downward deviations. From Khosrowabadi et al. (2022), *Price*, *Discount*, *Hot Day*, *Day Before Holiday*, and *Holiday* are significant in predicting downward deviations. The main control variable, *Proposal Size*, is positive and significant ($p<0.001$), showing that ASR system users are more likely to deviate downward when they see a large proposal. The main independent variable of interest, *Recent Stockout*, has a positive and significant coefficient ($p<0.01$), supporting Hypothesis 1. This coefficient estimate means that when an SKU has a recent stockout, its designated ASR system user counterintuitively orders a smaller quantity of that SKU than the ASR system proposes, which is a behavior aligned with censorship bias. Additionally, the estimated effect of *Relative Proposal Size* is positive and significant ($p<0.001$), indicating that the studied decision-makers' behaviors are also aligned with the anchoring bias.

The average marginal effects (AMEs) show that when there has been a recent stockout, the likelihood of a downward deviation from the ASR proposal increases by 2.2 percentage points. Given the 15.7 percent base rate of deviations (Table 1.1), this is not a trivial difference (14.0% increase). However, the effect of the *Relative Proposal Size* is greater at 3.5 percentage points, indicating that anchoring bias is a stronger predictor than censorship bias for the downward deviations from the ASR proposals.

Table 1.2: Effect of Censorship Bias on Downward Deviations

	Coefficients	AME
Constant	-2.838^{***}	
	(0.206)	
Proposal Size	0.032^{***}	0.0059
	(0.002)	
Case Pack Coverage	-0.002	-0.0003
	(0.003)	
On-hand Inventory	0.066^{**}	0.012
	(0.023)	
Item Size	-0.006^{***}	-0.0011
	(0.002)	
Margin	-0.184^{*}	-0.0336
	(0.074)	
Variety	-0.014^{***}	-0.0026
	(0.003)	
Forecast Dispersion	-0.004	-0.0008
	(0.011)	
Seasonality Error	0.278	0.0508
	(0.158)	
Sales	-0.012	-0.0022
	(0.007)	
Price	0.137^{***}	0.0249
	(0.019)	
Discount	1.198^{***}	0.3125
	(0.330)	
Cold Day	-0.097	-0.0173
	(0.057)	
Hot Day	0.259^{***}	0.0488
	(0.047)	
Day Before Holiday	-0.670^{**}	-0.0955
	(0.227)	
Holiday	0.438^{*}	0.0926
	(0.222)	
Relative Proposal Size	0.196^{***}	0.0357
	(0.013)	
Recent Stockout	0.118^{**}	0.0223
	(0.042)	
Pseudo R^2	0.32	
N (ASR Order Proposals)	11421	

$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$
Specification includes fixed effects to control for locations, product groups, proposal weekdays, weeks, and the interactions of locations and product groups.

1.5.2 Performance Implications of Downward Deviations and Censorship Bias

To assess the performance implications of downward deviations from ASR proposals, we present the estimation results of the second stage with robust standard errors in Table 1.3. *Proposal Size* and *Deviation Size* have significant ($p<0.001$) effects in the expected directions. The positive and significant coefficient ($p<0.001$) of *Recent Stockout* supports Hypothesis 2, indicating that when a downward deviation is susceptible to censorship bias, it is likely to lead to a stockout. Meanwhile, the coefficient of *Relative Proposal Size* is not significant, suggesting that even though this driver of anchoring bias is a significant trigger of downward deviations, it does not predict whether the downward deviations lead to stockouts (We test the robustness of this outcome to different operationalizations in Section 1.6). This is in stark contrast to the estimated AME of 24.8 percentage points of *Recent Stockout*, which is highly consequential given the 5.7 percent base rate of

Table 1.3: Effect of Censorship Bias on Stockouts

	Dependent variable: Stockout	
	Coefficients	AME
Constant	-1.240	
	(0.767)	
Proposal Size	-0.058^{***}	-0.0162
	(0.010)	
Deviation Size	0.098^{***}	0.0275
	(0.018)	
Relative Proposal Size	0.051	0.0144
	(0.043)	
Recent Stockout	0.773^{***}	0.2481
	(0.091)	
Pseudo R^2	0.24	
N (Downward Deviations)	1795	

$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$
Specification includes fixed effects to control for locations, product groups, proposal weekdays, weeks, and the interactions of locations and product groups.

stockouts in the data. Thus, we conclude that even though censorship bias is not as strong as anchoring bias in predicting deviations from ASR proposals, when censorship bias does trigger a downward deviation, the consequences are very likely to be severe.

1.6 Robustness Checks and Alternative Explanations

1.6.1 Robustness Checks

We employ various alternative model specifications to test the robustness of our results. First, we note that in our main empirical strategy, the moderation relationship between censorship bias and downward deviations is constructed via the selection mechanism. It shows that when there is a downward deviation from an ASR proposal, *Recent Stockout* effectively predicts whether a new stockout will occur. However, it does not take into account the possibility that recent stockouts might lead to new stockouts, irrespective of the downward deviations. In other words, the probit model with sample selection does not rule out the possibility that *Recent Stockout* causes a stockout even without a *Downward Deviation*. To test this alternative explanation to the observed result, we use the endogenous switching regression model, also known as Tobit type 5 (Amemiya 1984), which has been used in earlier decision-making research by Freeman et al. (2021) and Liu et al. (2015). In contrast to the probit model with sample selection, the endogenous switching regression has two second-stage outcome equations, enabling us to estimate the coefficients under both situations: when there is a downward deviation and when there is no deviation. The following two structural equations are jointly estimated via the Newton–Raphson maximization method by assuming that ξ_i, ζ_i and ϵ_i have a trivariate normal distribution (Lokshin and Sajaia 2011, Marra et al. 2022).

$$
\begin{aligned}
Stockout_i^* &= \gamma_0 + \gamma_1 ProposalSize_i + \gamma_2 DeviationSize_i \\
&\quad + \gamma_3 RelativeProposalSize_i + \gamma_4 RecentStockout_i \\
&\quad + \boldsymbol{\Gamma X_i} + \xi_i \\
Stockout_i &= \mathbb{I}[Stockout_i^* > 0] \qquad\qquad if\ DownwardDeviation_i = 1
\end{aligned}
$$

and

$$
\begin{aligned}
Stockout_i^* &= \gamma_5 + \gamma_6 ProposalSize_i \\
&\quad + \gamma_7 RelativeProposalSize_i + \gamma_8 RecentStockout_i \\
&\quad + \boldsymbol{\Gamma X_i} + \zeta_i \\
Stockout_i &= \mathbb{I}[Stockout_i^* > 0] \qquad\qquad if\ DownwardDeviation_i = 0
\end{aligned}
$$

where

$$
\begin{aligned}
DownwardDeviation_i^* &= \beta_0 + \beta_1 ProposalSize_i \\
&\quad + \beta_2 CasePackCoverage_i + \beta_3 OnhandInventory_i \\
&\quad + \beta_4 ItemSize_i + \beta_5 Margin_i + \beta_6 Variety_i \\
&\quad + \beta_7 ForecastDispersion_i + \beta_8 SeasonalityError_i \\
&\quad + \beta_9 Sales_i + \beta_{10} Price_i + \beta_{11} Discount_i \\
&\quad + \beta_{12} ColdDay_i + \beta_{13} HotDay_i \\
&\quad + \beta_{14} DayBeforeHoliday_i + \beta_{15} Holiday_i \\
&\quad + \beta_{16} RelativeProposalSize_i + \beta_{17} RecentStockout_i \\
&\quad + \boldsymbol{B X_i} + \epsilon_i \\
DownwardDeviation_i &= \mathbb{I}[DownwardDeviation_i^* > 0]
\end{aligned}
$$

Table 1.4 shows the estimation results of the endogenous switching regression model's second stage (The first stage is substantively similar to Table 1.2). The positive and significant coefficient ($p<0.001$) of *Recent Stockout* for no deviation cases signals that recent stockouts indeed lead to new stockouts irrespective of the downward deviations.

Table 1.4: Endogenous Switching Regression

	Dependent variable: Stockout	
	No Deviation	Downward Deviation
Constant	-1.711^{***}	-1.372
	(0.209)	(0.873)
Proposal Size	-0.032^{***}	-0.057^{***}
	(0.009)	(0.010)
Deviation Size		0.098^{***}
		(0.015)
Relative Proposal Size	-0.065	0.059
	(0.039)	(0.047)
Recent Stockout	0.446^{***}	0.782^{***}
	(0.064)	(0.092)
N	8986	1795

*p<0.05; **p<0.01; ***p<0.001
Specification includes fixed effects to control for locations, product groups, proposal weekdays, weeks, and the interactions of locations and product groups.

To compare this effect and the effects of the other predictors under each condition (i.e., no deviation and downward deviation), Figure 2 shows the predicted probability changes of a future stockout. For continuous variables, we take their 10% and 90% quantiles to calculate the probability changes and the confidence intervals. For binary variables, the figure shows the change when they move from 0 to 1. We observe that the likelihood of a stockout increases by 3.6 percentage points after a recent stockout even if no deviation is made. However, when a downward deviation is made after a recent stockout, the likelihood of a stockout increases by 24.7 percentage points. This substantive increase in the estimated effect further supports Hypothesis 2.

We complement our analysis with the recursive bivariate probit model (Maddala 1986) as an alternative way to tackle the possibility that the driver of censorship bias, a recent stockout, may cause future stockouts regardless of the deviations. Like our main model, this model includes two equations: one for the treatment (i.e., *Downward Deviation*) and one for the final outcome (i.e., *Stockout*). The specification of the first stage is the same as in the main model. For the second stage, since this model uses all the data in the

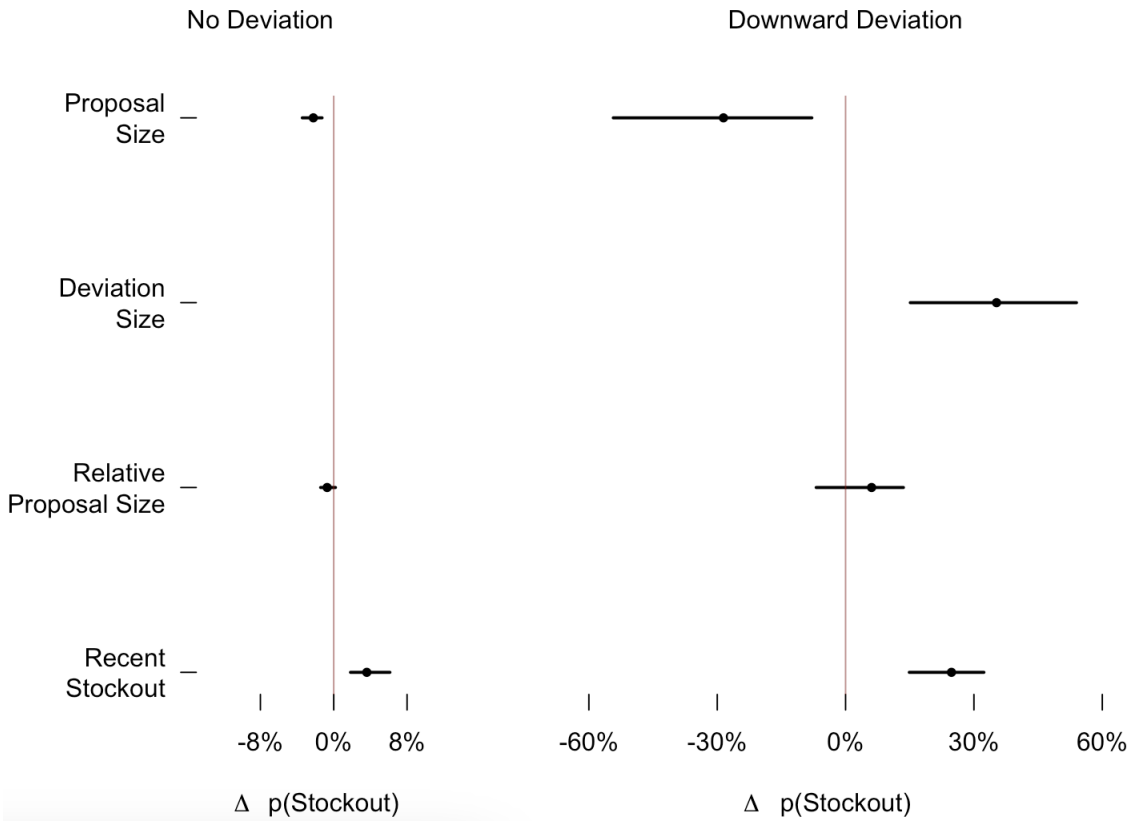

Figure 1.2: Endogenous Switching Regression

estimation of the outcome equation, we had to construct an interaction relationship between *Recent Stockout* and *Downward Deviation*. As our benchmark bias is anchoring, we also include the interaction of *Relative Proposal Size* and *Downward Deviation* in the second stage. Recursive bivariate probit model estimates both equations simultaneously via full maximum likelihood estimation (Greene 2003). To account for the endogeneity in deviation decisions, the model allows the error terms to be correlated with a correlation coefficient of ρ by assuming that the error terms follow a standard bivariate normal distribution. The recursive bivariate probit model has been recently used by Kim et al. (2015) to examine the effect of intensive care unit admission decisions on mortality, by Freeman et al. (2017) to examine the effect of epidural decisions on referral decisions in a maternity hospital, and by Liu et al. (2019) to examine the effects of rescheduling on patients' no-show behaviors. The results of the first stage are substantively similar

Table 1.5: Recursive Bivariate Probit Regression

	Dependent variable: Stockout
	Coefficients
Constant	-1.657^{***}
	(0.184)
Proposal Size	-0.051^{***}
	(0.006)
Deviation Size	0.076^{***}
	(0.012)
Downward Deviation	1.498^{***}
	(0.374)
Relative Proposal Size	-0.118^{***}
	(0.027)
Recent Stockout	0.393^{***}
	(0.064)
Downward Deviation * Relative Proposal Size	0.120^{**}
	(0.044)
Downward Deviation * Recent Stockout	0.287^{*}
	(0.117)
N	11421

*p$<$0.05; **p$<$0.01; ***p$<$0.001
Specification includes fixed effects to control for locations, product groups, proposal weekdays, weeks, and the interactions of locations and product groups.

to the results shown in Table 1.2. The positive and significant (p$<$0.01) coefficients for anchoring and censorship biases demonstrate that both biases trigger downward deviations from ASR proposals. The results of the outcome equation, in Table 1.5, show that when a downward deviation is susceptible to censorship bias, the probability of *Stockout* increases significantly (p$<$0.05).

Another robustness check relates to the operationalization of censorship bias. As mentioned in Section 1.4, to study censorship bias, we focus on the condition under which the bias can exist: *Recent Stockout*. It takes the value of 1 if the inventory goes to 0 in the previous seven days. Since "recent" could be defined in several ways, we change the time range to check if the results hold. In particular, we create *Recent Stockout* by considering the previous one, two, three, four, five, six, eight, nine, ten, eleven, twelve, thirteen, and

fourteen days. When we use one or two days, *Recent Stockout* is not significant in the first stage; however, the signs and significance of the estimated coefficients in the second stage remain the same. This is unsurprising since experiencing a recent stockout a day or two days ago would make the stockout incident very fresh in the minds of decision-makers. Thus, deviating downward from the ASR proposal would unlikely happen. Nevertheless, our results hold in both stages with the remaining operationalizations (three to fourteen days), providing further support for our hypotheses. Another possible approach to operationalize censorship bias is to focus on stockout occurrences in the most recent order coverage period. Particularly, *Recent Stockout* can be a binary variable taking the value of 1 if inventory goes to 0 in the most recently observed past order coverage period that matches the same weekdays as the order coverage period of the ASR proposal at hand. We run our main analysis with this operationalization of censorship bias and observe that *Recent Stockout* has significant and positive coefficient estimates in both stages. This provides further support for both our hypotheses.

We also test the robustness of the results to the operationalization of anchoring bias, our benchmark bias. In the main analysis, we used the mean of last seven days' purchase orders to calculate the tendency to anchor. To be sure that our results do not change depending on the time window, we also created this variable by using the mean purchase orders of the previous 14 and 21 days. The use of these alternative operationalizations did not change the insignificance of the coefficient estimate of *Relative Proposal Size* in the second stage. In the first stage, we observed that the effect decreased as the time window increased yet remaining significant in the same direction. Although we have argued that past purchase orders are more salient than past demand to anchor on, we are aware that most of past research has used the latter to operationalize the anchoring bias. Therefore, we use the mean of past sales as an alternative operationalization. We estimate the same probit model with the sample selection specification, and observe that this operationalization does not change the results. Specifically, the tendency to anchor

on the mean of past sales is significant in predicting downward deviations, yet it is not significant in predicting stockouts. Importantly, the signs and significance levels of *Recent Stockout* do not change in either stage, providing further support for both hypotheses.

Apart from anchoring bias, the newsvendor ordering literature has examined another biased behavior: demand chasing, which is a heuristic where decision-makers anchor on the previous order quantity and adjust orders toward the last realized demand (Kremer et al. 2010, Becker-Peth et al. 2013). The reason why such heuristic is biased is based on an experiment design where each period's demand is considered independent from the previous period's demand. Such a premise is tricky for field research in retail operations, where forecasting models that incorporate information from previous periods, namely time series models, show consistently superior performance (Ali and Gürlek 2020). It is therefore unclear whether the demand-chasing heuristic should be considered a bias in this setting. Nevertheless, as a robustness check, we add a variable to our model to capture demand chasing to see if it influences the effect of censorship bias. We follow a regression-based approach to operationalize demand chasing (Kirshner and Moritz 2021). In particular, our operationalization is the slope of sales over time, capturing the *Sales Trend*. We add this variable to the selection equation to predict the likelihood of a downward deviation and to the outcome equation to predict the likelihood of a stockout. The results suggest that *Sales Trend* has a negative coefficient estimate, meaning that the likelihood of a downward deviation increases when the sales trend is decreasing (negative), which is consistent with demand-chasing behavior. However, this behavior does not have negative operational consequences in terms of stockouts, as *Sales Trend* is not significant in the second stage. We also observe that the significance levels and signs of other variables' coefficient estimates do not change. In particular, *Recent Stockout* is still significant and positive in both stages, showing further support for our two hypotheses.

The next robustness check relates to how we operationalize our main dependent variable. In the main analysis, a binary variable, *Stockout*, is used to assess the performance

implications of deviations susceptible to censorship bias. As an alternative dependent variable, we compute the number of days without any stock relative to the order coverage period of an SKU. This dependent variable captures the percent of the order coverage period without the SKU at hand. The results of a classic Heckman selection model in which the second stage is linear show support for both hypotheses: *Recent Stockout* is significant ($p<0.01$) and positive in both stages. The second stage results suggest that a downward deviation made after a recent stockout increases the percent of the order coverage period without the SKU at hand by 22%.

We then consider the dependent variable of the first stage: *Downward Deviation*. In the main analysis, this variable takes the value of 1 if the purchase order is less than the corresponding ASR proposal, meaning that an upward deviation or no deviation is coded as 0. Lumping those two cases together may affect the coefficient estimation. To alleviate any such concerns, we run a multinomial logit model where the dependent variable takes the value of 1, 2, and 3 for downward deviation, no deviation, and upward deviation cases, respectively. The results indicate that, when there is a recent stockout, a downward deviation is more likely relative to no deviation ($p<0.05$), but an upward deviation is not ($p=0.14$). These results are in line with censorship bias, supporting Hypothesis 1.

We next run two-stage least squares (2SLS) regression since the probit model with sample selection (Van de Ven and Van Praag 1981) uses the full information maximum likelihood estimator, and it has been argued that fixed effects (such as our dummy variables for the product groups, weeks, weekdays, locations, and interactions of product groups and locations) create a risk of making the maximum likelihood estimator inconsistent (Greene 2004). The 2SLS procedure alleviates the inconsistency concern since it is based on a linear model in both stages. In the first stage, predicting downward deviations, the estimated coefficients of *Recent Stockout* and *Relative Proposal Size* remain positive and significant ($p<0.05$), lending further support for the existence of both censorship bias (Hypothesis 1) and anchoring bias. For the second stage, it is important to note that 2SLS

uses all the data in the estimation of the outcome equation, whereas the probit model with sample selection only used the observations where the selection variable equals 1. Thus, we followed the same logic as in the recursive bivariate probit analysis by including the interaction terms. The estimated coefficient for the interaction of *Recent Stockout* and *Downward Deviation* is positive and significant ($p<0.001$), supporting Hypothesis 2. We also run another 2SLS model using *Deviation Size* as the endogenous regressor in the second stage, instead of the binary *Downward Deviation*. Again, the estimated coefficient of the interaction between *Deviation Size* and *Recent Stockout* is significant and positive, supporting Hypothesis 2.

The probit model with sample selection assumes bivariate normality, and under distributional misspecification, the estimates may be inconsistent (Wojtyś et al. 2016) or biased (Park et al. 2022). To alleviate these concerns, we run a copula approach that relaxes the joint normality assumption (Hasebe 2013). As there are several different copulas available, we run our model with 15 different copulas. The likelihood-ratio-based tests for model selection developed by Vuong (1989) and Clarke (2007) favor the Clayton copula, and thus here, we discuss the results produced by this copula. However, we note that all 15 copulas produce similar coefficient estimates, demonstrating the robustness of the results. The coefficient estimates of the selection equation with the Clayton copula show support for Hypothesis 1, as *Recent Stockout* is positive and significant ($p<0.01$). We also observe that relaxing the joint normality assumption of the error terms produces a significant ($p<0.01$) coefficient estimate for *Relative Proposal Size* (anchoring bias). *Recent Stockout* also has a positive and significant ($p<0.001$) effect in this model, supporting Hypothesis 2.

Next, we exclude a week around the one holiday we have in our data collection period. We already control for the possible holiday effect by two variables in our main analysis: *Holiday* and *Day Before Holiday*. Nevertheless, to ensure the robustness of our results, we run our main model by removing the data points a week around the holiday day. One

difference in this model from our main model is that we drop *Holiday* and *Day Before Holiday*. The results hold in terms of significance levels and the signs of the coefficient estimates for *Recent Stockout* in both stages.

Next, we investigate whether having a recent stockout of the products that are in the same product group as the focal product triggers a downward deviation. If our explanation of censorship bias is correct, we should not observe a significant effect of the other products' recent stockout. To test this idea, we create a binary variable taking the value of 1 if the products within the same product group with the focal product had a stockout in the last seven days. When we add this new variable instead of *Recent Stockout*, we observe that it is not significant ($p>0.05$).

Our next two robustness checks relate to how we utilize our data. First, as mentioned in Section 1.4, we focus on nonzero ASR proposals, given that we are interested in downward deviations and users cannot deviate downward if the ASR system suggests a zero order for an SKU. This may create a concern that excluding zero ASR proposals leads to overestimating the likelihood of downward deviations. To test if this is the case, we run our main analysis on the whole data that include the zero proposals (n= 21,121). The results show that *Recent Stockout* is significant ($p<0.001$) and positive in both stages, providing further support to our hypotheses. Second, as mentioned in Section 1.4, we have two datasets collected a year apart: one for hypothesis testing and another one for developing applicable policies to improve ASR system usage. The former involves 12 weeks of ordering decisions from three locations, whereas the latter involves 12 weeks of ordering decisions from an additional two locations. We run our main analysis, the probit model with sample selection, by combining these two datasets to ensure that the results hold regardless of store selection. We observe that *Recent Stockout* is positive and significant ($p<0.05$) in the estimation of the selection equation, indicating that Hypothesis 1 holds in the combined data. Regarding the performance implications, we observe that both *Recent Stockout* ($\beta = 0.768$, $p<0.001$, AME = 0.248) and *Relative Proposal Size* ($\beta = 0.077$,

$p<0.05$, AME = 0.021) are positive and significant, showing further support for Hypothesis 2, and also that the small second-stage effect of anchoring bias, which was previously in the expected direction but insignificant, becomes statistically significant when more data are used.

In the main analysis, robust standard errors are estimated. Since reliance on robust standard errors may have its own drawbacks (Freedman 2006, King and Roberts 2015), we note that the significance levels of the estimated coefficients do not change when regular standard errors are estimated.

Lastly, throughout our operationalizations, we winsorized the continuous variables to ensure that extreme values do not affect our results. As a robustness, we run our main analysis without winsorizing those variables, which gives coefficient estimates that significantly support our hypotheses.

1.6.2 Potential Confounders and Alternative Explanations

In this paper, we have argued that censorship bias is the reason for the phenomenon we observe, that is, ASR system users order less than algorithmic suggestions after a recent stockout, which leads to an increased probability of future stockouts. Yet, there could be alternative explanations, other than censorship bias, causing this behavioral pattern. One direct alternative explanation is anchoring bias: After a stockout, the ASR system considers the lost sales and produces a higher proposal. Users are then likely to deviate downward from the proposal since they mentally anchor on the past purchase orders (or past sales). To distinguish this effect from that of censorship bias, we control for anchoring bias in the main analysis as well as in all of our robustness checks. Yet, anchoring bias is only one alternative explanation. In this section, we investigate what else could explain the phenomenon we observe.

Safety Stock Ignorance: Since decision-makers generally aim to reduce the ex-post inventory error in inventory replenishment tasks (Schweitzer and Cachon 2000), ASR system users could be inclined to place their orders always to match exactly the demand that they expect. After a recent stockout, such approach to ordering would ignore that also the safety stock must be replenished. If ASR system users were indeed biased towards ordering expected demand while ignoring the safety stock requirements, they would be more likely to deviate downward after a stockout. Such tendency would be an alternative to censorship bias to explain the curious behavior of ordering less after a stockout. We note, however, that this effect should be captured by the operationalization of anchoring bias in our models. Both anchoring bias operationalizations (i.e., the anchor point being the mean purchase orders and mean sales) capture the tendency to "pull-to-center" while ignoring safety stock requirements after a stockout. It is therefore unlikely that safety stock ignorance, instead of censorship bias, explains the observed phenomenon. Yet, to further ensure that this possible alternative explanation does not confound censorship bias, we add *Expected Lead Time* as an additional independent variable in the first stage of our model to predict *Downward Deviation*. The behavioral model presented by Tong and Feiler (2017) predicts an ordering behavior that leads to having less safety stock for products with short lead times. This prediction means that when the expected lead time is shorter, ASR system users are more inclined to pursue lower safety stock levels, increasing the likelihood of deviating downward. Hence, by adding *Expected Lead Time*, we are able to check whether users' safety stock considerations cancel the impact of censorship bias. The results suggest that the ASR system users are indeed more likely to deviate downward as the lead time gets shorter, yet *Recent Stockout* remains significant and positive, showing support for our first hypothesis and the censorship bias explanation.

Previous Upward Deviations: The discretion of ASR system users is not limited to downward deviations; they may also deviate upward from the proposed orders. Thus, it

could be that users have a higher tendency to deviate downward after they have previously made an upward deviation decision. In the study of van Donselaar et al. (2010), this behavior is called order advancement: ordering more than the algorithmic suggestions on less busy days, and then ordering less on busier days. In our study, it could be that users deviate upward immediately after a stockout but then revise subsequent orders by deviating downward to avoid accumulating too much inventory. This could be an alternative explanation to the observed phenomenon of ordering less after a recent stockout. We think that the impact of a previous upward deviation should be carried in *Proposal Size*, since the inventory level caused by the previous deviation decision is considered by the ASR system when it produces the next proposal. In any case, we test this alternative explanation idea by adding a binary variable, *Previous Up*, in the first stage of our model. This variable takes the value of 1 if the previous decision was an upward deviation. The estimated coefficient of *Previous Up* is negative and insignificant (p¿0.05), meaning that we do not observe a behavioral pattern akin to order advancement of van Donselaar et al. (2010) in our setting. Nevertheless, even in this analysis, the positive and significant impact of *Recent Stockout* remains, so we can rule out the order advancement explanation.

Workload Leveling: One of the reasons for deviations from algorithmic suggestions in the study of van Donselaar et al. (2010) is that managers consider the labor capacity requirements for the handling of the arriving shipments when they decide on order quantities, whereas a typical ASR system does not take the handling capacity considerations into account when producing proposals. Therefore, users of ASR systems may be more inclined to deviate downward if lots of orders are about to arrive on the same day as the current proposal's projected arrival date. To test this idea, we include in our model delivery weekday fixed effects, which capture the average busyness of each weekday. Three of the estimated coefficients are significant: β_{Monday}=-0.46 (p$<$0.001), $\beta_{Wednesday}$=-0.20

($p<0.01$), and β_{Friday}=-0.14 ($p<0.05$), indicating that the likelihood of a downward deviation is indeed affected by the day when the delivery is going to arrive. This signals that ASR system users in our setting may be incorporating handling capacity considerations in their replenishment decisions. Yet, even after controlling for this effect, *Recent Stockout* remains positive and significant.

Users' Attention and ASR Proposals' Salience: Another thing that might potentially influence our results is some systematic factor that would draw ASR system users' attention to specific order proposals. It could be, for example, that the users would scrutinize more closely the order proposals of any SKUs where they have just had to scrap some items for spoilage. To check this idea, we run additional models where such events are controlled for with two alternative operationalizations: *Spoiled*, a binary variable that takes the value of 1 if the SKU had spoiled inventory on the decision day, and *Spoiled Quantity*, the spoiled inventory quantity of the SKU on the decision day. Turns out that neither variable is significant in the results, but *Recent Stockout* remains significant and positive in both stages. Alternatively, users' attention could be drawn to the order proposals of the contemporaneously highest-selling SKUs. In the main model, we already include *Sales*, which is the sales of the decision day relative to the mean of the previous week's sales, but in order to consider this possibility more completely, we run a model where we complement the relative measure with the absolute measure of *Sales Quantity*. Also, the effect of that measure turns out insignificant in the results, and its inclusion does not change the effects of *Recent Stockout* in either stage. Finally, one more factor that could affect ASR system users' attention could be their screen layout; for example, a user could sequence ASR proposals based on profit margins. While this would be captured with our control variable *Margin*, there are countless other ways to organize the screen layout of the ASR system. Since the layout, and thus the sequence of ASR proposals, is entirely up to the user and not recorded in the data, we are not able to control directly for its effect

in the analyses. However, since users' screen layout preferences may be habitual and not change so often over time, it is likely that the attention-influencing effect of the screen layout is captured by the user fixed effects (the fixed effects of product groups and locations) that we have incorporated in all our models.

1.7 Policy Implications of Censorship Bias

In alignment with censorship bias, the results showed that recent stockouts trigger downward deviations from ASR proposals, and when this happens, the likelihood of a stockout increases. These findings can help retailers make better use of their ASR systems. It is possible to configure an ASR system to prevent users' actions, or at least to alert them, if they are trying to deviate under prespecified conditions. Our results inspire the idea of blocking downward deviations susceptible to censorship bias, that is, when they are performed after a recent stockout (i.e., stockout in the last week). This would obviously not eliminate all bad deviation decisions, but it would address one form of uninformed decision without completely forbidding the users from making deviations when they have private information about a forthcoming slump in the demand. Of course, blocking any downward deviations will logically reduce stockouts, and our results indicate that blocking them after a recent stockout will do so very effectively, yet what the results do not show is the potential downside in terms of inventory costs, the reduction of which is the whole purpose of allowing ASR users to order less than proposed. To explore the performance effects of the blocking policy, we collected a second dataset from two additional stores of the same supermarket chain one year after the first data collection, as explained in Section 1.4. The new dataset includes 320 SKUs from the same product groups and a total of 3901 proposals with 842 downward deviations, 154 of which led to stockouts. We call these self-inflicted stockouts because even if a stockout would have occurred in any case, a downward deviation made it occur earlier and resulted in greater lost sales.

To evaluate the policy of blocking downward deviations that are susceptible to censorship bias (*Block Censored*), we also test the practical implications of two other policies that could be justified by our empirical results. In particular, the significant coefficient of *Deviation Size* (Table 1.3) suggests that it might be a good idea to block downward deviations when users are trying to make large deviations. Since there is no obvious best definition for a large deviation, we operationalize this policy in two ways: *Block Large 1*, if the deviation is larger than the mean of the deviation sizes (8 units in the second dataset), and *Block Large 2*, if the deviation is larger than one case pack of the SKU in question. In total, we block 34.8%, 19.8%, and 15.2% of downward deviations under the policies *Block Large 1*, *Block Large 2*, and *Block Censored*, respectively.

Block Large 1 would have led to blocking 293 downward deviations. To assess whether this would have been beneficial, we check whether the inventory count went to 0 during the order coverage period after each downward deviation. If it did, then blocking the downward deviation would have been a good policy. If the inventory count did not go to 0, the blocking should not have been done because the downward deviation actually saved some inventory costs without inflicting a stockout. Of the 293 blocked deviations under *Block Large 1*, the policy would have been good 43 times, constituting a 27.9% reduction in the self-inflicted stockouts. To evaluate the cost of this policy, we compute the inventory holding cost caused by blocking the downward deviations. We first estimate the sales quantity (SALESQ) that would have been sold during the coverage period had the ASR system user not been able to deviate downward. There are many ways to estimate SALESQ. To account for weekday seasonality, our primary approach is to equate SALESQ to the sales of the same days in the previous week. We use SALESQ to calculate the inventory that would have been left at the end of the order coverage period. When the blocking would have been effective, the inventory left at the end of the order coverage period is computed by subtracting SALESQ from the sum of the on-hand inventory at the beginning of the order coverage period, the purchase order, and the deviation size.

When the blocking should not have been done, we calculate the extra stock caused by the policy by subtracting the purchase order from the sum of the on-hand inventory at the beginning of the order coverage period and the ASR proposal. To estimate the inventory holding cost for the extra inventory caused by the blocking, we multiply the unit cost of the SKU with its spoilage rate and find that *Block Large 1* would have increased the inventory holding costs of the affected SKUs by 4.7%.

Block Large 2 would have led us to block 167 downward deviations. To assess whether blocking these deviations would have been beneficial, we follow the same logic as above. This time, the policy would have been good 31 times, resulting in a 20.1% reduction in self-inflicted stockouts. We also assess the inventory implications in the same way as above, finding that *Block Large 2* would have increased the total inventory holding costs of the affected SKUs by 4.0%.

Finally, the policy to address censorship bias, *Block Censored*, would have led us to block 128 downward deviations. To assess whether blocking these downward deviations would have been beneficial, we follow the same logic as earlier and find that the policy would have been good 51 times, resulting in a 33.1% reduction in self-inflicted stockouts. Employing the same approach as previously, we find that *Block Censored* would have increased the total inventory holding costs of the affected SKUs by 2.0%.

The changes in the self-inflicted stockouts and inventory holding costs favor *Block Censored* over *Block Large 1* and *Block Large 2* since blocking the downward deviations susceptible to censorship bias decreases the self-inflicted stockouts more effectively, while increasing the inventory holding costs less than the other two policies. This outcome is robust to the different approaches that we used to estimate SALESQ, including a simple daily average sales of the preceding seven days, for instance. *Block Censored* is favored over *Block Large 1* and *Block Large 2* with similar differences in the inventory holding costs.

1.8 Discussion

Retailers employ ASR systems because these systems improve store performance by balancing shelf availability and inventory holding costs. However, retailers may not be maximizing the effectiveness of these systems because the ASR system users are allowed to deviate from the proposals, and besides improving ordering decisions with their private insights, they may make deviations that are driven by cognitive biases. Using data from a supermarket chain, this study focused on censorship bias, which has been studied previously only in lab experiments (Feiler et al. 2013, Rudi and Drake 2014, Tong et al. 2018). The results showed that censorship bias fairly often explains ASR system users' decisions to deviate downward from the order proposals and that when the deviations are susceptible to censorship bias, they very often lead to self-inflicted stockouts. The results also showed that the effect of censorship bias on downward deviations is not as strong as that of anchoring bias, but the effect of censorship bias on self-inflicted stockouts is much greater than that of anchoring bias.

This study informs human–machine interaction research by empirically showing that censorship bias explains a portion of discretionary behavior in the inventory replenishment context. As inventory management is critical for the performance of retailers, the behavioral biases that affect inventory decisions should be addressed. To do so, retailers may restructure the process of inventory replenishment. For example, they may use our findings to proactively identify and block the downward deviations when they are susceptible to censorship bias. Alternatively, instead of blocking the deviations and risking that some informed deviations are prevented, it might be enough to alert the user when their deviation decisions are susceptible to censorship bias. This could potentially activate the System 2 processes of the mind, the slower, and more intentional, effortful, and logical reasoning processes (Bazerman and Moore 2012), which have been associated with better decision making in the newsvendor setting (Moritz et al. 2013). The findings can

also be used in the training of ASR users so that System 2 processes might be activated autonomously before any deviations are made.

Our results also contribute to the literature on behavioral newsvendor decisions by studying the existence and performance implications of two behavioral biases in practice. With regard to censorship bias, we complement the laboratory experiments (Feiler et al. 2013, Rudi and Drake 2014, Tong et al. 2018) by studying the bias with field data. As Tan and Staats (2020) have pointed out, complementing laboratory experiments with field research is needed to obtain greater external validity and an in-depth understanding of the effects analyzed and discussed in the behavioral operations management literature. In our analysis, we show that both censorship and anchoring biases lead to downward deviations from ASR proposals and that anchoring bias indeed explains a good portion of the deviations, yet only censorship bias has a significant performance effect. This empirical finding, combined with the fact that demand is often censored in practice, signals that future research could focus more on understanding and mitigating censorship bias. As for the anchoring bias, we complement the work of Sachs et al. (2022) by expanding the anchors that may drive decision making outside the laboratory settings. We believe that our alternative operationalizations of anchoring bias may inform future research by highlighting that the anchors used in the experimental research, such as the mean of past demand, might not be applicable in practice where the mean demand may not be stationary and demand data can be censored. On top of these two behavioral biases, our additional analyses presented in Section 1.6 also complement the laboratory experiments by showing support for demand chasing and safety stock ignorance behaviors.

Finally, this study has several limitations that may inspire future research. First, we do not have individual-level data on ASR system users. Thus, we are unable to directly control for the individual differences that may affect deviation decisions. By controlling for the interactions of product groups and locations, we mitigate the concern that this lack of data may have biased our results. However, these controls are not foolproof because the

designated users of the ASR system may have had some days off during our data collection period, even though the period did not include any major holidays, and if so, then the temporary change in the user would not have been captured by the location-product group interactions. Since the ASR software developer and the studied supermarket chain indicate that changes in the designated user are infrequent, we trust that this shortcoming is only causing noise in our analysis. However, future research should try to incorporate individual characteristics, such as risk aversion, user experience, or cognitive capabilities, to understand how these traits affect the extent to which decision-makers suffer from the biases examined in this study. Second, our data include only perishable product groups, potentially limiting the empirical generalizability of the results. Specifically, ASR system users might not decide to deviate downward as frequently in the case of nonperishable products, which do not have similar spoilage risk. This might reduce the effect of censorship bias in those product groups. However, in terms of theoretical generalizability, we argue that the overall effect is unlikely to be fundamentally different for nonperishable products because even in the absence of spoilage risk, there is always an inventory holding cost associated with excess stock, and as a cognitive bias, censorship bias is unlikely to depend on the nature of the products. Regardless of whether the product is a ready-made meal or a bottle of shampoo, decision-makers may misinterpret sales figures after a stockout as the demand figures and act accordingly.

Chapter 1 References

Ali ÖG, Gürlek R (2020) Automatic Interpretable Retail forecasting with promotional scenarios. *International Journal of Forecasting* 36(4):1389-1406.

Amemiya T (1984) Tobit models: A survey. *Journal of Econometrics* 24(1-2):3-61.

Anderson ET, Fitzsimons GJ, Simester D (2006) Measuring and mitigating the costs of stockouts. *Management Science* 52(11):1751-1763.

Angerer A (2007) The impact of automatic store replenishment on retail: technologies and concepts for the out-of-stocks problem (Springer Science & Business Media).

Avlijas G, Simicevic A, Avlijas R, Prodanovic M (2015) Measuring the impact of stock-keeping unit attributes on retail stock-out performance. *Operations Management Research* 8(3):131-141.

Bazerman MH, Moore DA (2012) Judgment in managerial decision making (John Wiley & Sons).

Becker-Peth M, Katok E, Thonemann UW (2013) Designing buyback contracts for irrational but predictable newsvendors. *Management Science* 59(8):1800-1816.

Bell SJ, Luddington JA (2006) Coping with customer complaints. *Journal of Service Research* 8(3):221-233.

Bloomfield RJ, Kulp SL (2013) Durability, transit lags, and optimality of inventory management decisions. *Production and Operations Management* 22(4):826-842.

Bolton GE, Ockenfels A, Thonemann UW (2012) Managers and students as newsvendors. *Management Science* 58(12):2225-2233.

Bostian AA, Holt CA, Smith AM (2008) Newsvendor "pull-to-center" effect: Adaptive learning in a laboratory experiment. *Manufacturing & Service Operations Management* 10(4):590-608.

Campbell D, Frei F (2011) Market heterogeneity and local capacity decisions in services. *Manufacturing & Service Operations Management* 13(1):2-19.

Caro F, Saez de Tejada Cuenca A (2022) Believing in analytics: Managers' adherence to price recommendations from a DSS. *Manufacturing & Service Operations Management*, 25(2):524-542.

Castelo N, Bos MW, Lehmann DR (2019) Task-dependent algorithm aversion. *Journal of Marketing Research* 56(5):809-825.

Certo ST, Busenbark JR, Woo Hs, Semadeni M (2016) Sample selection bias and Heckman models in strategic management research. *Strategic Management Journal* 37(13):2639-

2657.

Chen K-Y, Li S (2018) The behavioral traps in making multiple, simultaneous, newsvendor decisions. Simultaneous, Newsvendor Decisions (August 2, 2018).

Clarke KA (2007) A simple distribution-free test for nonnested model selection. *Political Analysis* 15(3):347-363.

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1):114.

— (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64(3):1155-1170.

Eastwood J, Snook B, Luther K (2012) What people want from their professionals: Attitudes toward decision-making strategies. *Journal of Behavioral Decision Making* 25(5):458-468.

Elmaghraby W, Jank W, Zhang S, Karaesmen IZ (2015) Sales force behavior, pricing information, and pricing decisions. *Manufacturing & Service Operations Management* 17(4):495-510.

Feiler DC, Tong JD, Larrick RP (2013) Biased judgment in censored environments. *Management Science* 59(3):573-591.

Fildes R, Goodwin P, Lawrence M, Nikolopoulos K (2009) Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting* 25(1):3-23.

Freeman M, Robinson S, Scholtes S (2021) Gatekeeping, fast and slow: An empirical study of referral errors in the emergency department. *Management Science* 67(7):4209-4232.

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

Gavetti G, Greve HR, Levinthal DA, Ocasio W (2012) The behavioral theory of the firm:

Assessment and prospects. *Academy of Management Annals* 6(1):1-40.

Gavirneni S, Xia Y (2009) Anchor selection and group dynamics in newsvendor decisions—A note. *Decision Analysis* 6(2):87-97.

Germann M, Merkle C (2020) Algorithm Aversion in Delegated Investing. Available at SSRN 3364850.

Greene W (2004) The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *The Econometrics Journal* 7(1):98-119.

Greene WH (2003) Econometric analysis (Pearson Education India).

Gruen TW, Corsten DS (2007) A comprehensive guide to retail out-of-stock reduction in the fast-moving consumer goods industry (Procter & Gamble).

Gruen TW, Corsten DS, Bharadwaj S (2002) Retail Out of Stocks: A Worldwide Examination of Extent, Causes, and Consumer Responses (Grocery Manufacturers of America, Washington, DC).

Hasebe T (2013) Copula-based maximum-likelihood estimation of sample-selection models. *The Stata Journal* 13(3):547-573.

Heckman JJ (1979) Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society* 153-161.

Hewage HC, Perera HN, De Baets S (2022) Forecast adjustments during post-promotional periods. *European Journal of Operational Research* 300(2):461-472.

Ibanez MR, Clark JR, Huckman RS, Staats BR (2018) Discretionary task ordering: Queue management in radiological services. *Management Science* 64(9):4389-4407.

JCMR (2022) Supply Chain Planning System of Record Industry Analysis, Market Size, Share, Trends, Growth and Forecast 2021 - 2029. Report.

Kesavan S, Kushwaha T (2020) Field experiment on the profit implications of merchants' discretionary power to override data-driven decision-making tools. *Management Science* 66(11):5182-5190.

Khosrowabadi N, Hoberg K, Imdahl C (2022) Evaluating Human Behaviour in Response

to AI Recommendations for Judgemental Forecasting. *European Journal of Operational Research* 303(3): 1151-1167.

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

Kirshner SN, Moritz BB (2021) Measuring demand chasing behavior. *Decision Sciences* 52(6):1264-1281.

Kremer M, Minner S, Van Wassenhove LN (2010) Do random errors explain newsvendor behavior? *Manufacturing & Service Operations Management* 12(4):673-681.

Kuncel NR, Klieger DM, Ones DS (2014) In hiring, algorithms beat instinct. *Harvard Business Review* 92(5):p32-32.3.

Lau N, Bearden JN (2013) Newsvendor demand chasing revisited. *Management Science* 59(5):1245-1249.

Lee YS, Siemsen E (2017) Task decomposition and newsvendor decision making. *Management Science* 63(10):3226-3245.

Liu A, Mazumdar T, Li B (2015) Counterfactual decomposition of movie star effects with star selection. *Management Science* 61(7):1704-1721.

Liu J, Xie J, Yang KK, Zheng Z (2019) Effects of rescheduling on patient no-show behavior in outpatient clinics. *Manufacturing & Service Operations Management* 21(4):780-797.

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151:90-103.

Lokshin M, Sajaia Z (2011) Impact of interventions on discrete outcomes: Maximum likelihood estimation of the binary choice models with binary endogenous regressors. *The Stata Journal* 11(3):368-385.

Maddala GS (1986) Limited-dependent and qualitative variables in econometrics (Cambridge university press).

Marra G, Radice R, Zimmer D (2022) A Unifying Switching Regime Regression Framework with Applications in Health Economics.

McKinsey (2020) Future of retail operations: Winning in a digital era. Report.

Meehl PE (1954) Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.

Moritz BB, Hill AV, Donohue KL (2013) Individual differences in the newsvendor problem: Behavior and cognitive reflection. *Journal of Operations Management* 31(1-2):72-85.

Önkal D, Goodwin P, Thomson M, Gönül S, Pollock A (2009) The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making* 22(4):390-409.

Park W-Y, Nickerson J, Bigelow L (2022) Stuck in the Middle (of Time): Strategic Repositioning and Survival in Response to an Innovation Shock in a Growing Market.

Phillips R, Şimşek AS, Van Ryzin G (2015) The effectiveness of field price discretion: Empirical evidence from auto lending. *Management Science* 61(8):1741-1759.

Prahl A, Van Swol L (2017) Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting* 36(6):691-702.

Rudi N, Drake D (2014) Observation bias: The impact of demand censoring on newsvendor level and adjustment behavior. *Management Science* 60(5):1334-1345.

Sachs AL, Becker-Peth M, Minner S, Thonemann UW (2022) Empirical newsvendor biases: Are target service levels achieved effectively and efficiently? *Production and Operations Management* 31(4):1839-1855.

Schweitzer ME, Cachon GP (2000) Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Science* 46(3):404-420.

Shi C, Chen W, Duenyas I (2016) Nonparametric data-driven algorithms for multiproduct inventory systems with censored demand. *Operations Research* 64(2):362-370.

Snyder C, Keppler S, Leider S (2022) Algorithm Reliance Under Pressure: The Effect of

Customer Load on Service Workers. Available at SSRN 4066823.

Sun J, Zhang DJ, Hu H, Van Mieghem JA (2021) Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Science*.

Tan TF, Staats BR (2020) Behavioral drivers of routing decisions: Evidence from restaurant table assignment. *Production and Operations Management* 29(4):1050-1070.

Tong J, Feiler D (2017) A behavioral model of forecasting: Naive statistics on mental samples. *Management Science* 63(11):3609-3627.

Tong J, Feiler D, Larrick R (2018) A behavioral remedy for the censorship bias. *Production and Operations Management* 27(4):624-643.

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185(4157):1124-1131.

Van de Ven WP, Van Praag BM (1981) The demand for deductibles in private health insurance: A probit model with sample selection. *Journal of Econometrics* 17(2):229-252.

van Donselaar KH, Gaur V, van Woensel T, Broekmeulen RA, Fransoo JC (2010) Ordering behavior in retail stores and implications for automated replenishment. *Management Science* 56(5):766-784.

Vuong QH (1989) Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica: *Journal of the Econometric Society*:307-333.

Wansink B, Kent RJ, Hoch SJ (1998) An anchoring and adjustment model of purchase quantity decisions. *Journal of Marketing Research* 35(1):71-81.

Wojtyś M, Marra G, Radice R (2016) Copula regression spline sample selection models: the r package semiparsamplesel. *Journal of Statistical Software* 71:1-66.

Yeomans M, Shah A, Mullainathan S, Kleinberg J (2019) Making sense of recommendations. *Journal of Behavioral Decision Making* 32(4):403-414.

Chapter 2

Newsvendor Double-Counting Bias in Automatic Grocery Retail Replenishment

2.1 Introduction

Day-to-day replenishment ordering in grocery retail is a highly automated operation. Modern supermarkets use automatic store replenishment (ASR) systems that run self-adaptive forecasting and data-driven inventory optimization models. Yet, typical implementations of these systems leave full discretion to users to adjust the replenishment order proposals before they are converted to actual purchase orders to the suppliers (e.g., van Donselaar et al. 2010). This discretion is critical to allow for the cognitive flexibility of the human mind and to heed demand signals from daily encounters with customers that the algorithms of an ASR system could not possibly capture. This is particularly important for dealing with demand peaks that are born in media or social media, such as celebrity-endorsed foods, new ingredients' health benefit claims, and other food trends (Kambhampaty and Creswell 2021, Partridge 2022), or driven by idiosyncratic events that time series-based forecasting

models cannot predict. The inception of the COVID-19 crisis makes a prominent example of such an event, with sudden demand surges for many items, from disinfectants to toilet paper (Campbell et al. 2020). Daily supermarket operations are filled with less drastic instances of the same phenomenon, such as a competitor's stores stocking out on some items, or some special event being organized in the neighborhood of a store location. To avoid stockouts in these situations, the users of ASR systems must be empowered to override the system and augment the proposed replenishment orders with the quantities that they believe the algorithm has missed.

Although allowing users to augment the order proposals of their ASR system (henceforth ASR orders) sounds intuitively important to avoid stockouts when facing emerging trends or idiosyncratic demand peaks, the effectiveness of such discretion is unclear. Clarifying the effects would be a step towards a better understanding of the operations managerial implications of retail automation, to which Caro et al. (2020) have called for more research attention. In the context of ASR systems, the pioneering study of van Donselaar et al. (2010) and a working paper by Li et al. (2015, as cited in Lee and Siemsen (2017)) have suggested that discretionary adjustments may have positive performance effects, but overall, research has been mixed regarding the performance implications of managerial discretion in the use of decision support systems. Caro and Saez de Tejada Cuenca (2022) report a negative revenue impact when managers override the recommendations of a pricing support system in fashion retail. Kesavan and Kushwaha (2020) report a negative profitability effect for managerial overriding of software-based inventory optimization at an automotive spare parts retailer. Outside the retail context, Ibanez et al. (2018) report a negative productivity effect for physician discretion over work sequences in a diagnostic task, and Käki et al. (2019) find mostly negative effects from production planners overriding their decision support system in power plant operations. Yet, positive performance effects from managerial discretion have been reported in the context of staffing financial service operations (Campbell and Frei 2011) and in car rental pricing

decisions (Phillips et al. 2015). The absence of universally positive performance effects is understandable, given the susceptibility of human discretion to various decision-making biases (Davis 2018).

In this study, we investigate the performance effects of ASR system users' decisions to increase the orders (i.e., upward deviations from ASR orders) created by their system, exploring the different stimuli for the decisions and how the stimuli associated with behavioral biases influences the effectiveness of upward deviations. Inspired by the work of van Donselaar et al. (2010), we suggest a novel behavioral bias that may lead to unnecessary upward deviations: newsvendor double counting. We suspect that ASR system users may misinterpret ASR orders as demand forecasts and apply the newsvendor optimization logic to them. This would be a double-counting bias (Bunn and Salo 1996), since the ASR system already applies the newsvendor optimization when creating the orders. We investigate this proposed new behavioral bias on data from an upmarket European supermarket chain where ASR system users have full discretion to increase ASR orders. Since stockout avoidance is the reason why users are allowed discretion to increase ASR orders, and since every unnecessarily increased order inflates the store inventory and the associated inventory costs, spoilage risk, and in-store logistics challenges, it is in the interest of the supermarket chains to reduce the ineffective decisions to increase ASR orders. We use a variety of econometric methods to account for the endogeneity of ASR system users' decisions. The results show that (i) under the stimulus for newsvendor double counting (i.e., high critical ratio of the newsvendor model), the likelihood of an upward deviation is greater, and (ii) when ASR order is increased under that condition, the order is less likely to deliver any benefit in stockout avoidance, resulting instead in extra inventory. We compare the prevalence and the effects of this newly proposed newsvendor double-counting bias to two well-known behavioral ordering biases that can also trigger upward deviations, namely supply line underweighting (Sterman 1989, Bloomfield and Kulp 2013) and anchoring (Tversky and Kahneman 1974, Schweitzer and Cachon 2000).

According to our results, newsvendor double counting explains upward deviations from ASR orders better than the anchoring bias and supply line underweighting do. Under strong versus weak stimulus for newsvendor double counting (i.e., plus/minus one standard deviation from the mean of the newsvendor ratio), the likelihood of an upward deviation increases from 2.7 percent to 4.6 percent. This estimated 1.9 percentage point increase is greater than the corresponding effects of anchoring bias and supply line underweighting, which are estimated at 0.9 and 1.0 percentage points, respectively. Newsvendor double counting also emerges as the strongest predictor of unnecessary upward deviations (i.e., increases that are not necessary to avoid a stockout and only result in extra inventory). When ASR orders are increased under strong stimulus for newsvendor double counting, the desired outcome of stockout avoidance is less than half as likely (9.5%) than when ASR orders are increased under weak stimulus for newsvendor double counting (25.9%). The result holds when controlling for the size of the ASR order and is thus not explained by the fact that the stockout risk is lower when the newsvendor ratio is high.

Our results contribute to the behavioral inventory management literature by introducing newsvendor double counting as a new decision-making bias. If replicated and corroborated in further research, this bias can complement our field's understanding of the downsides of human discretion in automated or algorithmically-supported inventory management decisions. Newsvendor double counting is curious behavior in the sense that many experimental studies have shown that decision makers, both student subjects and professional managers (Bolton et al. 2012), fail to sufficiently consider the newsvendor logic when deciding order quantities based on past demand data, a phenomenon that has been largely explained by anchoring bias (Tversky and Kahneman 1974, Schweitzer and Cachon 2000). Curiously, the observed insufficient consideration of the newsvendor logic seems to turn into over eagerness to consider it when moving from the laboratory setting into the day-to-day supermarket operations where ordering decisions are not based on the decision maker's own analysis of the past demand but instead supported by software-

generated orders.

2.2 Literature Review

Inventory management is one of the classic core areas of operations management and it has been studied from many different perspectives. Our study is informed by and aims to contribute to the behavioral aspects of inventory management, with a particular focus on the behavior related to the newsvendor model, which has received a lot of research attention since the study of Schweitzer and Cachon (2000) revealed how anchoring bias (Tversky and Kahneman 1974) drives suboptimal stocking decisions. According to the newsvendor model, the optimal stocking quantity that maximizes expected profit is

$$Q^* = F^{-1}(\frac{UnderstockingCost}{UnderstockingCost + OverstockingCost})$$

where $F^{-1}(.)$ is the inverse of the cumulative distribution function for demand. In the experiments of Schweitzer and Cachon (2000), like in their numerous subsequent replications (for a review, see Becker-Peth and Thonemann 2018), subjects systematically chose stocking quantities that lie between Q* and the mean demand. The anchoring on the mean demand, the so-called "pull-to-center effect" (Bostian et al. 2008), has been studied the most, but experimental research has also investigated other anchor points, exploring, for example, how subjects consider recent demand and previously placed replenishment orders (i.e., demand chasing) when making their stocking decisions (Bostian et al. 2008, Lau and Bearden 2013, Kirshner and Moritz 2021). Although all of these anchor points – mean demand, recent demand, and previous replenishment orders – are straightforward in a laboratory setting, they may not be so evident anchors in retail practice. In practice, demand is unlikely stable over time (a standard assumption in the past lab studies) and is usually not a single figure because of different demand cycles for different product types. In addition, demand may be totally unobservable due to the fact

that lost sales are usually not observable in practice (Feiler et al. 2013). The last placed order may not be such intuitive anchor in practice because of the weekday seasonality patterns in demand (van Donselaar et al. 2010). Hence, instead of those two frequently observed anchor points in laboratory experiments, other anchor points may be more logical for managers in practice, such as the typical order level for each stock-keeping unit (SKU). We hence complement the behavioral newsvendor literature by testing this idea empirically, which also allows us to control for anchoring bias and use it as a benchmark.

Another important behavioral bias that has been observed in the past research on replenishment ordering is supply line underweighting (Sterman 1989). In the influential study on the bullwhip effect, Sterman (1989) has proposed that human decision makers are unable to assess the delayed feedback properly when demand distribution is unknown and nonstationary. Their misperception of feedback makes them underweight the supply line, leading them to overorder. Croson and Donohue (2006) have reported that supply line underweighting is still observed even when the demand distribution is known and stationary, and even when inventory information is shared. In their lab experiment, Bloomfield and Kulp (2013) has moved the investigation of this bias from the bullwhip phenomenon to the newsvendor setting, and showed the prevalence of supply line underweighting in newsvendor order decisions. We also control for this bias and use it as another benchmark to evaluate the effect of newsvendor double-counting bias on augmentation decisions. By doing so, we test the prevalence of it in the retail practice.

2.3 Hypotheses

The evidence from the past research on decision makers' insufficient use of the newsvendor logic in ordering decisions, especially when the studied subjects have been practitioners (Bostian et al. 2008, Bolton et al. 2012, Moritz et al. 2013), may seem surprising given how intuitive it is to consider unit profit (an understocking cost) and perishability (an

overstocking cost) when deciding on order quantities. Importantly, however, the existing research has not indicated that practitioners are entirely ignorant of the newsvendor logic, just that, on average, they are not using it to a sufficient extent to make optimal decisions. In fact, contrary to ignorance, empirical research outside the laboratory experiments has shown that practitioners' inventory management decisions are generally well aligned with the newsvendor logic, particularly regarding the profit margins of the stocked products (Gaur et al. 2005, Corbett and Fransoo 2007, Rumyantsev and Netessine 2007). Meanwhile, even in the laboratory setting, the inventory ordering behavior has been observed to become more optimal when products are perishable (Bloomfield and Kulp 2013), that is, when the overstocking cost is more tangible than the mere opportunity cost of capital, to which inventory decision makers have been shown to be quite insensitive (Rumyantsev and Netessine 2007). These provide evidence that inventory management professionals actually follow the newsvendor logic in their decision making, albeit not necessarily reaching fully optimal decisions, as evidenced by the laboratory studies.

Algorithm aversion literature states that decision makers often do not understand or know how the algorithmic solutions of their decision support systems are produced (Burton et al. 2020). In the retail ordering context, managers have shown a tendency to confuse demand forecasts with ordering decisions (Fildes et al. 2009). In fact, practitioners have manifested this tendency in the particular context of automated ordering in supermarkets. Specifically, van Donselaar et al. (2010) hypothesized and found partial support for the idea that ASR system users are more prone to advance (i.e., augment) ASR orders for products that have higher profit margins. This inspired us to contemplate that perhaps the partial support for their hypothesis was due to the fact that profit margin only captures the understocking cost part of the newsvendor logic. Also, the inventory cost should be estimated to capture the overstocking cost of the newsvendor ratio. Guided by the previous findings stating that practitioners are particularly sensitive to perishability (Bloomfield and Kulp 2013) and not sensitive to capital costs (Rumyantsev and Netessine

2007), we proxy the overstocking cost with a product's spoilage propensity. We expect that the resulting newsvendor ratio will predict ASR system users' behavior better than the mere product profitability. Yet, mentally doing the newsvendor optimization on ASR orders would be a double-counting bias (Bunn and Salo 1996), since the ASR system already applies the newsvendor optimization when creating the orders. Therefore, an augmentation made under the stimulus for newsvendor double counting is likely to be an uninformed deviation; hence, it is less likely to result in any additional sales. Therefore, we hypothesize the following two hypotheses:

Hypothesis 1: The likelihood of augmenting an ASR order is positively related to the newsvendor ratio.

Hypothesis 2: The likelihood of avoiding a stockout is lower when an augmentation decision is susceptible to newsvendor double-counting bias.

2.4 Setting and Data

We collaborated with an industry-leading optimization software provider (JCMR 2022) to collect data consisting of ordering decisions spanning a four-month period, from one of its customers, an upmarket European supermarket chain. The dataset includes three store locations that are based in the three largest cities of the studied chain's primary market. This chain began using the ASR system more than three years before we collected the data. In all three stores, managers are given discretionary power to augment ASR orders since it is important to incorporate personnel's demand insights stemming from their daily encounters with customers. Such customer encounters, promotion events, celebrity endorsed products, or local events in the neighborhood may signal an increasing demand that could be missed by the ASR system. To avoid stockouts in these circumstances, ASR system users may choose to increase the proposed replenishment orders since stockouts pose a great financial risk for retailers (Anderson et al. 2006). Yet, those increases

may stem from behavioral biases instead of demand signals, meaning that operational efficiency is at risk because of excess inventory and consequential spoilage. Hence, it becomes essential to distinguish upward deviations from ASR orders that are triggered by behavioral biases from ones that are driven by demand signals.

The two product groups, packaged meats and ready-made meals, were chosen to be included in our data for a number of reasons. First, the SKUs in these product groups are handled in integer piece quantities instead of weight units, which makes the operationalization of variables more straightforward. Second, these two product groups allowed us to avoid promotional effects that could potentially taint the results. Third, the studied supermarket chain had a particular interest in these two product groups, especially because any order increase carries a higher risk of spoilage for perishable products. These product groups include 472 different SKUs, culminating in 1,416 different SKU-location combinations. The many different characteristics of these SKUs, such as margin, forecasting error, and demand volatility, help us control for the antecedents of deviations reported in the literature (van Donselaar et al. 2010), which is essential to disentangle the effect of newsvendor double-counting bias.

Our data include 21,192 ASR orders, and upward deviation decisions were made for 804 of them. Out of 21,192 cases, 11,434 include a nonzero ASR order. The performance outcome we focus on is whether stockouts are avoided during an SKU's order coverage period. In our data, order coverage periods range from one to seven days. Lead times range from two to four days, and they are known and reliable.

There is one dedicated ASR system user to each product group in each store location, meaning that ASR system users review the ASR orders for the SKUs of their own product group before finalizing replenishment orders. The algorithms of the ASR system are a black box to the user, as they were in other studies of discretionary behavior (e.g., Sun et al. 2021). These users are not restricted by any budget limitations while increasing the orders. The compensation they receive is a fixed salary plus a bonus based on an-

nual store profits, and it does not include an individual performance-based component. Hence, their incentives are in line with those of the supermarket chain, that being profit maximization. One limitation we face is that our data do not include ASR system users' individual characteristics. Yet, we mitigate this by adding the interactions of store location and product group fixed effects as there is only one dedicated ASR system user to each product group in each store location. Another limitation is that, because we are bound by the confidentiality of the proprietary algorithms of the ASR system, we are unable to disclose those proprietary algorithms, which is also the case in similar studies (e.g., Kesavan and Kushwaha 2020).

2.5 Measures

2.5.1 Dependent Variables

Upward Deviation is a binary variable taking the value of 1 if the purchase order quantity is greater than the ASR order. The upward deviation decisions also include the cases in which the ASR system did not recommend any order quantity for a particular SKU, yet its user decided to create an order anyway.

To assess the effectiveness of increased orders, our main dependent variable is *Stockout Avoided*, taking the value of 1 if there have been sales from the upward deviation, in other words, if the upward deviation decision has prevented a stockout. In Section 2.8, the robustness of the results will be assessed with alternative operationalizations of the performance.

2.5.2 Independent Variables

We use *Newsvendor Ratio* to study the implications of newsvendor double counting bias on increasing an ASR order.

$$UnitProfitMargin = (\frac{Price - UnitCost}{Price})$$

$$NewsvendorRatio = (\frac{UnderstockingCost}{UnderstockingCost + OverstockingCost})$$

$$= \frac{Price - UnitCost}{Price - UnitCost + UnitCost * SpoilagePropensity}$$

We include the well-known behavioral biases affecting newsvendor order decisions, namely, anchoring (Schweitzer and Cachon 2000) and supply line underweighting (Bloomfield and Kulp 2013). To operationalize the anchoring bias, we calculate *Relative Proposal Size* by taking into account the salience of past purchase orders. Thus, this variable is calculated as

$$\frac{ProposalSize_{kj}}{MeanPastPurchaseOrders_{kj}} - 1$$

where k and j represent the SKU and location, respectively. The mean in the denominator of this operationalization takes the mean of the past seven days (The robustness of the results will be tested with different time windows.), purchase orders. We expect that the smaller the ASR order relative to the mean of the past purchase orders, the higher the likelihood of an *Upward Deviation*. Next, to operationalize supply line underweighting, we create *Goods in Transit*, which equals the quantity of goods in the pipeline. We expect that the greater the quantity of inbound deliveries, the higher the likelihood of an *Upward Deviation*.

2.5.3 Controls

We include *Proposal Size* as a control variable to predict *Upward Deviation* and *Stockout Avoided*. The expected coefficient of *Proposal Size* is negative in predicting *Upward De-*

viation since ASR system users would be less inclined to increase a large ASR order. We expect a positive coefficient for this variable in predicting *Stockout Avoided*. *Proposal Size* is winsorized at the 1% tails. It is worth noting that *Proposal Size* is different from *Relative Proposal Size* (anchoring bias). The former reflects the reaction to and the salience of encountering a large or small ASR order, whereas the latter captures the stimuli for anchoring bias.

Upward Deviation Size is added as a control variable only in the second stage of our selection model, as it is only available for the cases where ASR orders have been increased. We argue that the magnitude of the increase itself should have an impact on *Stockout Avoided*. For example, ordering 10 more items yields a higher prospect of additional sales compared to ordering three more items. To control for this type of a difference in the performance, *Upward Deviation Size* should be added. We winsorize this variable at the 1% tails.

In addition to *Proposal Size* and *Upward Deviation Size*, we control for the fixed effects of locations, product groups, weeks, proposal weekdays, and the interactions of locations and product groups. These interactions are added to capture the ASR system users' individualistic qualities, such as risk aversion, as there is only one designated ASR system user for each product group in each store location. The fixed effects of weeks and proposal weekdays are added to capture time-related effects such as seasonality.

2.5.4 Exclusion Restrictions

ASR system users' decisions to increase ASR orders are endogenous. To tackle this problem, we employ the probit model with sample selection (Van de Ven and Van Praag 1981) that uses a set of exclusion restrictions, like instruments. We borrow these variables from the study of van Donselaar et al. (2010), which investigates deviations from algorithmic suggestions in ordering decisions. In particular, the variables adopted from this study are *Case Pack Coverage*, *On-hand Inventory*, *Item Size*, *Margin*, *Variety*, *Sea-*

sonality Error, and *Forecast Dispersion*.

We follow the approach of van Donselaar et al. (2010) to measure *Case Pack Coverage*, that is, an SKU's case pack size divided by its weekly mean sales. This variable is winsorized at the 1% tails. We expect a negative effect because an increase for a large case pack coverage would produce more significant implications compared to an increase for an SKU that has a small case pack coverage. *On-hand Inventory* is measured by dividing an SKU's end of the day inventory by its weekly mean end of the day inventory balances. We winsorize *On-hand Inventory* at the 1% tails. We expect a negative effect as ASR system users would be more encouraged by a lean inventory to increase the orders. We use *On-hand Inventory* to capture the opposite effect of the net shelf space variable in the study of van Donselaar et al. (2010) since our data do not have the SKUs' shelf space allocations. We use package width and height to define *Item Size* rather than the approach employed by van Donselaar et al. (2010), which captures the three-dimensional volume. This is because the studied supermarket chain has a greater interest in the area measure owing to the studied products' shapes. The expected effect of *Item Size* is positive since the results of van Donselaar et al. (2010) suggest that ASR system users want to bring in larger SKUs earlier. Our measure of *Margin*, winsorized at the 1% tails, is an SKU's absolute profit margin, the same measure used by van Donselaar et al. (2010). We expect a positive effect since ASR system users should have a high level of stockout avoidance for high margin products, leading them to increase the orders. Our measure of *Variety* for an SKU is the number of other SKUs in the same product subgroup, the identical approach employed by van Donselaar et al. (2010). We expect a positive effect for this variable since a high level of variety means a low degree of substitution, and variety stimulates demand (Baumol and Ide 1956), making ASR system users more inclined to increase the orders. van Donselaar et al. (2010) focus on *Seasonality Error* and *Forecast Dispersion* to investigate the implications of demand uncertainty on order advancements. We add the same variables, both winsorized at the 1% tails, by employing the same mea-

surements. Specifically, *Seasonality Error* is calculated as the root mean squared error of the differences between the seasonality index predictions based on past demand and the daily demands of the previous seven days. Our measure of *Forecast Dispersion* is computed by first using a trend-based forecasting model to obtain daily forecast errors. Then, we divide the weekly standard deviation of those forecast errors by average sales. The expected effect for both *Seasonality Error* and *Forecast Dispersion* is positive since a high level of uncertainty would lead ASR system users to hold a greater level of safety stocks.

2.6 Identification Strategy

We investigate whether a high stimulus for newsvendor double-counting bias leads ASR system users to increase orders, and the performance implications of such decisions. Upward deviations from ASR orders are not random; that is why we model the these decisions by employing the probit model with sample selection (Van de Ven and Van Praag 1981) that accounts for the inherent endogeneity of such decisions. Probit model with sample selection estimates a selection and an outcome equation via a maximum likelihood estimator. To improve identification, we include exclusion restrictions, borrowed from van Donselaar et al. (2010) and explained in detail in the previous section, in the selection equation. It is important to note that van Donselaar et al. (2010) only examine discretionary behavior in an inventory replenishment setting, leaving the performance implications out of their analysis.

The probit models to estimate the decision to increase an ASR order and the perfor-

Table 2.1: Descriptive Statistics and Correlation Table

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
mean	0.038	0.088	5.019	0.707	5.371	5.848	1.074	16.02	0.342	6.475	0.16	1.241	0.146	10.83
sd	0.191	1.402	10.384	0.256	7.874	6.198	0.847	15.309	0.3	5.342	0.15	1.553	0.353	12.189
min	0	-1	0	0.067	0	0.012	0	5	-0.146	0	0	0	0	1
max	1	5	76	0.996	40	36	5.529	60	1.425	21	0.847	11.757	1	72
1 Upward Deviation														
2 Relative Proposal Size	-0.023													
3 Goods In Transit	0.111	-0.088												
4 Newsvendor Ratio	0.075	-0.030	0.150											
5 Proposal Size	0.053	0.453	0.228	0.212										
6 Case Pack Coverage	-0.090	-0.104	-0.238	-0.122	-0.229									
7 Onhand Inventory	-0.027	-0.251	-0.075	-0.018	-0.169	0.024								
8 Item Size	0.017	-0.078	-0.038	0.281	0.104	0.216	-0.019							
9 Margin	0.055	-0.008	0.034	0.404	0.173	-0.050	-0.051	0.331						
10 Variety	-0.012	-0.011	-0.006	0.023	-0.014	0.006	-0.003	0.135	0.049					
11 Seasonality Error	-0.020	0.003	-0.038	-0.085	-0.121	0.038	0.053	-0.068	-0.064	0.083				
12 Forecast Dispersion	0.007	0.005	0.006	0.007	-0.044	-0.013	0.003	-0.063	-0.028	-0.037	-0.010			
13 Stockout Avoided	NA	-0.070	0.021	-0.232	-0.102	-0.007	-0.033	-0.124	-0.135	-0.003	0.067	0.029		
14 Upward Deviation Size	NA	-0.126	-0.006	0.150	0.127	0.065	0.107	0.190	0.332	0.039	0.015	0.065	0.034	

Note:
All $|\rho| > 0.013$ are statistically significant at p$<$0.05.

mance implications of such decisions take the following forms:

$$\begin{aligned} UpwardDeviation_i^* =& \alpha_0 + \alpha_1 ProposalSize_i + \alpha_2 CasePackCoverage_i + \\ & \alpha_3 OnhandInventory_i + \alpha_4 ItemSize_i + \alpha_5 Margin_i + \\ & \alpha_6 Variety_i + \alpha_7 ForecastDispersion_i + \alpha_8 SeasonalityError_i + \\ & \alpha_9 GoodsInTransit_i + \alpha_{10} RelativeProposalSize_i + \\ & \alpha_{11} NewsvendorRatio_i + \Theta \boldsymbol{X_i} + \varepsilon_i \\ UpwardDeviation_i =& \mathbb{I}[UpwardDeviation_i^* > 0] \end{aligned}$$

and

$$\begin{aligned} StockoutAvoided_i^* =& \beta_0 + \beta_1 ProposalSize_i + \beta_2 UpwardDeviationSize_i + \\ & \beta_3 GoodsInTransit_i + \beta_4 RelativeProposalSize_i + \\ & \beta_5 NewsvendorRatio_i + \boldsymbol{\gamma X_i} + \zeta_i \\ StockoutAvoided_i =& \mathbb{I}[StockoutAvoided_i^* > 0] \end{aligned}$$

where i represents each ASR order. $UpwardDeviation_i^*$ and $StockoutAvoided_i^*$ are the latent variables for ordering more than an ASR order and avoiding a stockout, respectively, and $\mathbb{I}[.]$ is the indicator function. As for the exclusion restrictions, the effects of variables from van Donselaar et al. (2010) are represented by α_2 to α_8. In both equations, to control for as many confounders as possible, we add the fixed effects for locations, product groups, weekdays, weeks, and the interactions of locations and product groups, represented by X_i. The interactions fixed effects are added as a proxy for the fixed effects of the ASR system users since in each store location, there is one dedicated user for each product group.

2.7 Results

2.7.1 Determinants of Upward Deviations

To test Hypothesis 1, we follow a hierarchical approach shown in Table 2.2. The first column of this table represents the probit model to estimate *Upward Deviation* with only controls and exclusion restrictions. In the probit model in the second column, we add newsvendor double counting and two benchmark biases, namely, anchoring and supply line underweighting. We observe that *Item Size* and *Margin* are not significant. Since the inclusion of valid and strong exclusion restrictions are critical for endogeneity correction in a probit model with sample selection, we delete these two variables from our final model, shown in the third column. It is worth noting that deleting these two exclusion restrictions does not change the significance of *Newsvendor Ratio*, as can be seen in column 2.

In column 3, we first observe that the estimated coefficient of *Proposal Size* is negative and significant ($p<0.01$); the larger the ASR order, the less likely that ASR system users deviate upward from it. From van Donselaar et al. (2010), *Case Pack Coverage*, *On-hand Inventory*, *Variety*, *Seasonality Error*, and *Forecast Dispersion* are significant at the expected direction in predicting *Upward Deviation*.

To test the hypothesis that an upward deviation decision is positively related to newsvendor double counting, we check the estimated coefficient of *Newsvendor Ratio*. The positive and significant ($p<0.01$) coefficient estimation supports Hypothesis 1, showing that ASR system users evaluate the products and ordering decisions holistically by mentally doing the newsvendor optimization, instead of putting their focus on only unit margin as hypothesized by van Donselaar et al. (2010). The magnitude of the average marginal effect (AME) of *Newsvendor Ratio* suggests that, ceteris paribus, a one-unit increase in the stimulus for newsvendor double counting increases the likelihood of *Upward Deviation* by 3.8 percentage points. To understand the magnitude of this effect better, we next examine our benchmark biases: anchoring (Schweitzer and Cachon 2000), and supply line under-

weighting (Bloomfield and Kulp 2013). First, the estimated coefficient of *Relative Proposal Size* is negative and significant ($p<0.01$), as predicted by anchoring bias. This finding suggests that ASR system users anchor on the mean of the past orders when evaluating a new ASR order. The magnitude of the AME of *Relative Proposal Size* implies that, ceteris paribus, a one-unit increase in this variable decreases the likelihood of *Upward Deviation* by 0.4 percentage points. Next, as predicted by supply line underweighting, the estimated coefficient of *Goods in Transit* is positive and significant ($p<0.01$): the higher the quantity of inbound deliveries, the higher the likelihood of *Upward Deviation*. The magnitude of the AME of *Goods in Transit* implies that, ceteris paribus, a one-unit increase in this variable increases the likelihood of *Upward Deviation* by 0.1 percentage points.

2.7.2 Performance Implications of Upward Deviation Decisions

Table 2.3 shows the estimation results of the performance equation, where we investigate whether upward deviation decisions made under the stimulus for newsvendor double counting lead to stockout prevention. First, we observe that *Newsvendor Ratio* is negative and significant ($p<0.01$), suggesting that when newsvendor double-counting bias is effective in triggering an upward deviation decision, this decision is less likely to result in additional sales. In particular, a one-unit increase in *Newsvendor Ratio* decreases the likelihood of preventing a stockout by 52 percentage points. This provides support for our Hypothesis 2. Although we observe that the stimulus for newsvendor double-counting bias decreases the potential benefits of upward deviation decisions, our benchmark biases are not significant in predicting *Stockout Avoided*. This implies that, though anchoring and supply line underweighting cause upward deviation, they do not have a statistically significant impact on the performance of such upward deviation decisions. Furthermore, we observe that *Upward Deviation Size* is significant ($p<0.05$) and positive, as expected. A one-unit increase in *Upward Deviation Size* increases the likelihood of preventing a stockout by 0.3 percentage points. While a large upward deviation is understandably a

Table 2.2: Effect of Newsvendor Double Counting on Upward Deviation Decisions

	Coefficients (1)	Coefficients (2)	Coefficients (3)	AME
Constant	-1.084^{***}	-1.664^{***}	-1.657^{***}	
	(0.151)	(0.169)	(0.169)	
Proposal Size	-0.007^{***}	-0.008^{***}	-0.008^{***}	-0.001
	(0.002)	(0.003)	(0.003)	
Case Pack Coverage	-0.064^{***}	-0.054^{***}	-0.054^{***}	-0.004
	(0.005)	(0.005)	(0.005)	
On-hand Inventory	-0.082^{***}	-0.092^{***}	-0.092^{***}	-0.007
	(0.024)	(0.025)	(0.025)	
Item Size	0.002	0.001		
	(0.001)	(0.001)		
Margin	0.083	0.015		
	(0.057)	(0.060)		
Variety	-0.009^{***}	-0.009^{***}	-0.009^{***}	-0.001
	(0.003)	(0.003)	(0.003)	
Seasonality Error	-0.531^{***}	-0.338^{*}	-0.341^{*}	-0.025
	(0.179)	(0.182)	(0.182)	
Forecast Dispersion	0.022^{**}	0.019^{*}	0.018^{*}	0.001
	(0.011)	(0.011)	(0.011)	
Newsvendor Ratio		0.505^{***}	0.519^{***}	0.038
		(0.091)	(0.086)	
Relative Proposal Size		-0.048^{***}	-0.048^{***}	-0.004
		(0.018)	(0.018)	
Goods In Transit		0.009^{***}	0.009^{***}	0.001
		(0.001)	(0.001)	
Pseudo R^2 (McFadden)	0.103	0.120	0.120	
N (ASR Orders)	21,192	21,121	21,121	

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$. Specification includes fixed effects to control for locations, product groups, proposal weekdays, weeks, and the interactions of locations and product groups.

Table 2.3: Effect of Newsvendor Double Counting on Stockout Avoided

	Coefficients	AME
Constant	1.253^{*}	
	(0.709)	
Proposal Size	-0.011	-0.004
	(0.010)	
Upward Deviation Size	0.008^{**}	0.003
	(0.004)	
Newsvendor Ratio	-1.456^{***}	-0.520
	(0.228)	
Relative Proposal Size	-0.032	-0.011
	(0.069)	
Goods In Transit	-0.005	-0.002
	(0.004)	
N (Upward Deviation Decisions)	803	
ρ	-0.574^{***}	

*p$<$0.1; **p$<$0.05; ***p$<$0.01. Specification includes fixed effects to control for locations, product groups, proposal weekdays, weeks, and the interactions of locations and product groups.

significant predictor of ASR system users having some private knowledge, it is considerably worse at predicting good decisions in comparison to how strongly *Newsvendor Ratio* predicts uninformed upward deviation decisions.

2.8 Robustness Checks

Our first robustness check relates to the operationalization of anchoring bias. In the main analysis, we used the mean of the past seven days' orders to calculate *Relative Proposal Size*. To ensure that our results are robust to the selection of this time window, we replicate our results by using the mean of the past 14 days and 21 days. As we increase the time window, the effect of anchoring bias on upward deviation decisions becomes weaker but stays in the expected direction. More importantly, *Newsvendor Ratio* remains significant (p$<$0.01) in the expected directions in both stages, providing further support for both

hypotheses.

The next robustness check also relates to the operationalization of anchoring bias. In Section 2.2, we argue that, in practice, past orders are easier to capture as an anchoring point than the past demand, which is the typical anchor in laboratory experiments (e.g., Bolton and Katok 2008). To assess the robustness of the effect of newsvendor double-counting bias, if anchoring on the mean demand is also observed in the grocery replenishment practice, we create *Relative Proposal Size* based on the mean of the past seven days' sales figures. The probit model with sample selection with this new operationalization shows that the estimated coefficient of *Relative Proposal Size* is not significant in the first stage, suggesting that the demand is indeed not salient enough in the retail practice for ASR system users to anchor on. This is likely because demand is difficult to capture in practice as it typically constitutes different cycles for different product types, and because lost sales are usually not observable (Feiler et al. 2013). Nevertheless, the estimated coefficient of *Newsvendor Ratio* in both stages is in the expected direction and significant ($p<0.01$), providing further support for both hypotheses.

To assess the robustness of our results to the operationalization of performance, we create a continuous dependent variable, namely, *Sales from Upward Deviation*. This variable is calculated as follows:

$$(\frac{Sales - (On - handInventory + ProposalSize)}{UpwardDeviationSize})$$

and is only available when there is an upward deviation; thus, we employ a Heckman selection model (Heckman 1976) which includes two steps: The selection equation to predict *Upward Deviation*, and the performance equation to predict *Sales from Upward Deviation* when *Upward Deviation* equals 1. The selection equation of the Heckman selection model is the same as that of the probit model with sample selection. There are two differences in the performance equation: i) the dependent variable is *Sales from Upward Deviation* instead of *Stockout Avoided*, and ii) the two-stage Heckman selection with a

Table 2.4: Effect of Newsvendor Double Counting on Sales from Upward Deviation

	Coefficients
Constant	0.317***
	(0.108)
Proposal Size	−0.001
	(0.001)
Upward Deviation Size	0.002***
	(0.001)
Newsvendor Ratio	−0.216***
	(0.037)
Relative Proposal Size	0.008
	(0.008)
Goods In Transit	−0.001
	(0.001)
N (Upward Deviation Decisions)	803
Adjusted R^2	0.075
Inverse Mills Ratio	−0.082**

*p<0.1; **p<0.05; ***p<0.01. Specification includes fixed effects to control for locations, product groups, proposal weekdays, weeks, and the interactions of locations and product groups.

continuous dependent variable estimates an inverse mills ratio and puts it in the performance equation, accounting for the potential sample selection bias (Certo et al. 2016). We observe that *Newsvendor Ratio* is positive and significant ($p<0.01$) in the first stage of this two-stage model, providing support for Hypothesis 1. Table 2.4 shows that *Newsvendor Ratio* is negative and significant ($p<0.01$) in predicting *Sales from Upward Deviation*, supporting Hypothesis 2.

We next create another operationalization for the performance: *Sales from Order*, that is the quantity of the sales from the purchase order. We winsorized *Sales from Order* at the 1% tails. Since *Sales from Order* is available when there has been no upward deviation as well as when there has been one, selection models are not suitable. Hence, we

employ an endogenous treatment regression model, estimated by a full maximum likelihood estimator, with instrumental variables (Heckman 1978, Maddala 1986). This model simultaneously estimates two equations: one for the binary endogenous treatment (i.e., *Upward Deviation*) and one for the performance (i.e., *Sales from Order*). The performance equation is only identified if the rank condition is satisfied (Ibanez et al. 2018). This condition means that at least one exogenous regressor with a nonzero coefficient from the treatment equation should be excluded from the performance equation. This requires us to include variables akin to instruments in the treatment equation. Such variables are valid if they are correlated with *Upward Deviation*, and uncorrelated with the error term of the performance equation. Our exclusion restrictions in our selection model serve as instrumental variables in this model. Thus, the treatment stage of this model is the same as the selection stage of our probit model with sample selection. In the performance equation of this model, we follow what we call the marker approach to determine the upward deviation decisions that are marked with the behavioral biases of interest. Particularly, we created a binary marker variable for each of the biases that assess whether the purchase order equals the quantity predicted by those biases. *Newsvendor Double Counting Marker* takes the value of 1 if the purchase order falls within the interval determined by *Newsvendor Ratio*. Let us assume that, for Product A, the newsvendor optimal order is 57.5. Product A is ordered in case packs that contain six items each. The closest possible order to this newsvendor optimal would be 60. If the purchase order is also 60, then *Newsvendor Double Counting Marker* equals 1. *Anchoring Marker* takes the value of 1 if the purchase order is withing the range predicted by anchoring bias. *Supply Line Underweighting Marker* takes the value of 1 if the upward deviation size equals the quantity of goods in the pipeline. We interact these markers with *Upward Deviation* in the performance equation below as we are interested in the performance implications of upward deviation decisions that are marked by newsvendor double counting, anchoring, and the

supply line underweighting biases.

$$
\begin{aligned}
SalesFromOrder_i = {} & \beta_0 + \beta_1 UpwardDeviation_i + \beta_2 ProposalSize_i \\
& + \beta_3 AnchoringMarker_i + \beta_4 SupplyLineUnderweightingMarker_i \\
& + \beta_5 NewsvendorDoubleCountingMarker \\
& + \beta_6 UpwardDeviation_i * AnchoringMarker_i \\
& + \beta_7 UpwardDeviation_i * SupplyLineUnderweightingMarker_i \\
& + \beta_8 UpwardDeviation_i * NewsvendorDoubleCountingMarker_i + \boldsymbol{\Gamma X_i} + \zeta_i
\end{aligned}
$$

The endogenous treatment regression model assumes that the error terms of the equations, ε_i and ζ_i, follow a bivariate normal distribution with mean zero and covariance matrix $\begin{bmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{bmatrix}$. The endogenous treatment regression model has been previously used by Ibanez et al. (2018) to examine the productivity effect of radiologists' decisions to deviate from their prescribed work sequence, and by Moreno and Terwiesch (2015) to examine the impact of manufacturing mix flexibility on discounts in automobile industry. The results of the treatment equation show that *Newsvendor Ratio* is positive and significant ($p<0.01$) in predicting *Upward Deviation*, providing support for Hypothesis 1. Next, we observe that ρ, the correlation estimation between ε_i and ζ_i, is positive and significant ($p<0.05$), implying that unobservables that increase *Sales from Order* tend to occur with unobservables that increase the decisions of *Upward Deviation* (positive bias). As for the results of the second stage, Table 2.5 shows that all interaction terms between *Upward Deviation* and the markers are negative and significant, showing that upward deviation decisions marked by anchoring, supply line underweighting, and newsvendor double counting negatively affect *Sales from Order*. This further supports Hypothesis 2.

Lastly, we analyze whether phantom inventory leads the store managers to increase the ASR orders. Phantom inventory, coined by Gruen and Corsten (2007), refers to an inventory record inaccuracy, that is, the inventory system shows a greater availability of

Table 2.5: Effect of Newsvendor Double Counting on Sales from Order

	Coefficients
Constant	-0.033
	(0.049)
Proposal Size	0.038^{***}
	(0.001)
Upward Deviation	0.311^{***}
	(0.082)
Newsvendor Double Counting Marker	0.353^{***}
	(0.018)
Anchoring Marker	-0.097^{***}
	(0.014)
Supply Line Underweighting Marker	0.262^{***}
	(0.020)
Upward Deviation x Newsvendor Double Counting Marker	-0.600^{***}
	(0.082)
Upward Deviation x Anchoring Marker	-0.207^{**}
	(0.081)
Upward Deviation x Supply Line Underweighting Marker	-0.452^{***}
	(0.098)
N (ASR Orders)	21,121
ρ	0.086^{**}

*p<0.1; **p<0.05; ***p<0.01. Specification includes fixed effects to control for locations, product groups, proposal weekdays, weeks, and the interactions of locations and product groups.

an item. In our setting, if the ASR system has this inaccuracy yet the store managers are aware of the actual availability of an SKU, *Upward Deviation* would be more likely. To test this idea, we operationalize the phantom inventory as a binary variable that is equal to 1 when the end of day inventory count for the focal SKU is 1, 0 otherwise. We choose 1 as the threshold since phantom inventory is typically at low levels. When we add this binary variable in the first stage of our main model, its estimated coefficient is not significant ($p>$0.05). When we take the end of day inventory count threshold as 2 to operationalize this binary variable, the results are the same. Importantly, the drivers for the newsvendor double counting, the anchoring bias, and the supply line underweighting are still significant ($p<$0.05) and at the expected directions. Hence, we conclude that, phantom inventory is not a concern in our examination.

2.9 Structural Estimation

In this section, using our field data, we investigate store managers' psychological costs of overage and of underage when they replenish the inventory. Ho et al. (2010) suggest that aside from the actual underage and overage costs, newsvendor has an additional psychological underage and overage costs. Through their structural estimation model, they find that the psychological aversion to overage is greater than the psychological aversion to underage. If asymmetrical, these additional psychological costs would also effect the deviation decisions of the decision makers in our data. Hence, by utilizing a structural estimation model, we estimate the weights given by the store managers to overage and underage when they decide on the order quantities.

We utilize a maximum likelihood estimator that estimates α, β, γ, and δ:

$$CriticalRatio = (\frac{(\alpha+1) * UnderageCost}{(\alpha+1) * UnderageCost + (\beta+1) * OverageCost})$$

$$OrderEstimated = ASRProposal + \sigma_{ASRProposal} * F^{-1}CriticalRatio+$$
$$\gamma * (MeanPastPurchaseOrders - ASRProposal) + \delta * GoodsInTransit$$

$$\min_{\alpha,\beta,\gamma,\delta}(-\sum log(f\left(\frac{Order - OrderEstimated}{\sigma_{Order}}\right)))$$

where α, β, γ, and δ represent the parameters for underage cost, overage cost, anchoring bias, and supply line underweighting, respectively. On top of the parameters for underage and overage costs, we add parameters for the anchoring and supply line underweighting biases. Given that our results as well as previous research suggest that these biases are effective in decision makers' ordering behaviors, we need to include them in our structural estimation. $\sigma_{ASRProposal}$ and σ_{Order} are the standard deviations of ASR proposals and order realizations, respectively.

In this estimation, the objective is to minimize the difference between the realizations of the orders and the estimated orders. We standardize these differences by the standard deviation of the order realizations before taking the logarithm of the probability density function. The assumptions of this process are i) ASR proposals follow a normal distribution, and ii) the standardized error follows a standard normal distribution. We select the initial values for the parameters to be optimized over as 0, representing the null model in which decision makers do not have additional psychological costs of over- and under-ordering, do not anchor on the mean of previous orders, and do not underweight the supply line. We constructed the weights for underage and overage costs as $(\alpha + 1)$ and $(\beta + 1)$ to evaluate their statistical difference from 0. Table 2.6 shows the parameter estimations and their standard errors.

We observe that, all parameter estimates are significantly different from 0 ($p<0.001$), suggesting that because of the psychological costs they carry, decision makers weights the underage and overage costs. Also, we see that anchoring and supply line under-

Table 2.6: Structural Estimation Results

Parameter		Estimate Average
α	Underage cost	-0.806***
		(0.005)
β	Overage cost	5.523***
		(0.0001)
γ	Anchoring on past orders	0.064***
		(0.007)
δ	Supply line underweighting	0.114***
		(0.007)

*p<0.5; **p<0.01; ***p<0.001.

weighting biases are in function when decision makers see an ASR order and decides on a final purchase order. Specifically, in our setting, decision makers' psychological cost of overage is larger than the psychological cost of underage. This is in line with the findings of Ho et al. (2010) and Zhao et al. (2021).

We then investigate whether newsvendor double counting still affects ordering decisions, after accounting for the psychological costs of overage and underage. To do so, we create a new critical ratio variable by incorporating the parameter estimates reported in Table 2.6. In other words, this new critical ratio is created by multiplying the underage cost and overage cost by 0.194 and 6.523, respectively. Then, we utilize another maximum likelihood estimator:

$$OrderEstimated = ASRProposal + \sigma_{ASRProposal} * F^{-1}NewCriticalRatio * \omega + 0.064 * (MeanPastPurchaseOrders - ASRProposal) + 0.114 * GoodsInTransit$$

$$\min_{\omega}(-\sum log(f\left(\frac{Order - OrderEstimated}{\sigma_{Order}}\right)))$$

where ω represent whether newsvendor double counting is in function. If this parameter estimate is significantly different from 0, it means that decision makers are applying the newsvendor logic and deviating from $ASRProposal$ accordingly, even when their psycho-

logical costs of overage and underage are accounted for. We observe that μ = 0.098 ($p<0.001$). All combined, through our structural estimation analysis, we see that the newsvendor double-counting bias is robust even after controlling for the psychological costs of overage and underage.

2.10 Discussion

Although retailers widely use ASR systems in ordering decisions, human decision makers have the final say. In the grocery inventory replenishment context, store managers are empowered to increase ASR orders to ensure that demand is met and stockouts are prevented as these managers may have private information that is missed by the ASR system's algorithms. Yet, this study shows that store managers suffer from a bias we coined as newsvendor double counting in their upward deviation decisions from ASR orders. Specifically, store managers mentally apply the newsvendor logic on ASR orders by considering overordering and underordering costs. This would be a double-counting bias since this newsvendor optimization logic is applied to forecasts by the ASR system before those forecasts are converted into order proposals. By analyzing ordering decisions of store managers of a supermarket by various econometrics techniques, we show that upward deviation decisions triggered by newsvendor double-counting bias do not lead to additional sales. This shows that the discretionary power of the users to increase orders may lead to overstocking, increasing the inventory holding costs and waste.

This study makes several contributions to the growing literature on behavioral newsvendor decisions. First, we propose a novel behavioral bias that explains a good portion of upward deviation decisions. If this bias, newsvendor double counting, is corroborated by future research, it can increase our understanding of human-machine interaction in the usage of ASR systems for the inventory replenishment task. Second, this study complements the study of Sachs et al. (2022) by providing an empirical examination of be-

havioral biases with secondary data. As Sachs et al. (2022) state, studies analyzing ordering behavior in actual practice are needed to improve our field's understanding of biases affecting the newsvendor behavior. Our study does so by showing that anchoring and supply line underweighting trigger ASR system users to deviate upward from ASR orders. To the best of our knowledge, this study is the first empirical examination of the prevalence of supply line underweighting in retail practice. In addition, our study contributes the understanding of anchoring bias by suggesting a new anchor that is more prevalent in retail practice than the mean demand, that is the purchase orders. We think that future research may be inspired by our alternative operationalization of anchoring bias to consider that different anchor points are salient to decision makers depending on the decision-making context.

This study is not free from limitations. First, the data we use from the supermarket chain do not include ASR system users' individual characteristics; that is why we are unable to control for the user-level characteristics such as risk aversion. We tackle this issue by adding the interactions of store and product group fixed effects. If individual-level data on decision makers are collected, future research can examine the association between individual qualities and the degree to which one has the impetus for newsvendor double-counting bias. Second, we only have data from two product groups that are perishable. Since critical ratio of the newsvendor optimization is dependent upon the spoilage propensity of SKUs, the effect sizes we report may not be empirically generalizable to other product groups.

Chapter 2 References

Anderson ET, Fitzsimons GJ, Simester D (2006) Measuring and mitigating the costs of stockouts. *Management Science* 52(11):1751-1763.

Baumol WJ, Ide EA (1956) Variety in retailing. *Management Science* 3(1):93-101.

Becker-Peth M, Thonemann UW (2018) Behavioral inventory decisions. *The handbook of behavioral operations* 11393.

Bloomfield RJ, Kulp SL (2013) Durability, transit lags, and optimality of inventory management decisions. *Production and Operations Management* 22(4):826-842.

Bolton GE, Katok E (2008) Learning by doing in the newsvendor problem: A laboratory investigation of the role of experience and feedback. *Manufacturing & Service Operations Management* 10(3):519-538.

Bolton GE, Ockenfels A, Thonemann UW (2012) Managers and students as newsvendors. *Management Science* 58(12):2225-2233.

Bostian AA, Holt CA, Smith AM (2008) Newsvendor "pull-to-center" effect: Adaptive learning in a laboratory experiment. *Manufacturing & Service Operations Management* 10(4):590-608.

Bunn DW, Salo AA (1996) Adjustment of forecasts with model consistent expectations. *International Journal of Forecasting* 12(1):163-170.

Burton JW, Stein MK, Jensen TB (2020) A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 33(2):220-239.

Campbell D, Frei F (2011) Market heterogeneity and local capacity decisions in services. *Manufacturing & Service Operations Management* 13(1):2-19.

Campbell MC, Inman JJ, Kirmani A, Price LL (2020) In times of trouble: A framework for understanding consumers' responses to threats. *Journal of Consumer Research* 47(3):311-326.

Caro F, Saez de Tejada Cuenca A (2022) Believing in analytics: Managers' adherence to price recommendations from a DSS. *Manufacturing & Service Operations Management*, 25(2):524-542.

Caro F, Kök AG, Martínez-de-Albéniz V (2020) The future of retail operations. *Manufacturing & Service Operations Management* 22(1):47-58.

Certo ST, Busenbark JR, Woo Hs, Semadeni M (2016) Sample selection bias and Heck-

man models in strategic management research. *Strategic Management Journal* 37(13):2639-2657.

Corbett CJ, Fransoo JC (2007) Entrepreneurs and newsvendors: do small businesses follow the newsvendor logic when making inventory decisions? Available at SSRN 1009330.

Croson R, Donohue K (2006) Behavioral causes of the bullwhip effect and the observed value of inventory information. *Management Science* 52(3):323-336.

Davis AM (2018) Biases in Individual Decision-Making. *The Handbook of Behavioral Operations*:149-198.

Feiler DC, Tong JD, Larrick RP (2013) Biased judgment in censored environments. *Management Science* 59(3):573-591.

Fildes R, Goodwin P, Lawrence M, Nikolopoulos K (2009) Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting* 25(1):3-23.

Gaur V, Fisher ML, Raman A (2005) An econometric analysis of inventory turnover performance in retail services. *Management Science* 51(2):181-194.

Gruen TW, Corsten DS (2007) A comprehensive guide to retail out-of-stock reduction in the fast-moving consumer goods industry (Procter & Gamble).

Heckman J (1978) Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica* 46(4):931-59.

Heckman JJ (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of economic and social measurement*, volume 5, number 4 (NBER), 475-492.

Ho T-H, Lim N, Cui TH (2010) Reference dependence in multilocation newsvendor models: A structural analysis. *Management Science* 56(11):1891–1910.

Ibanez MR, Clark JR, Huckman RS, Staats BR (2018) Discretionary task ordering: Queue management in radiological services. *Management Science* 64(9):4389-4407.

JCMR (2022) Supply Chain Planning System of Record Industry Analysis, Market Size,

Share, Trends, Growth and Forecast 2021 - 2029. Report.

Käki A, Kemppainen K, Liesiö J (2019) What to do when decision-makers deviate from model recommendations? Empirical evidence from hydropower industry. *European Journal of Operational Research* 278(3):869-882.

Kambhampaty AP, Creswell J (2021) The Era of the Celebrity Meal. The New York Times https://www.nytimes.com/2021/12/08/style/celebrity-fast-food-partnerships.html.

Kesavan S, Kushwaha T (2020) Field experiment on the profit implications of merchants' discretionary power to override data-driven decision-making tools. *Management Science* 66(11):5182-5190.

Kirshner SN, Moritz BB (2021) Measuring demand chasing behavior. *Decision Sciences* 52(6):1264-1281.

Lau N, Bearden JN (2013) Newsvendor demand chasing revisited. *Management Science* 59(5):1245-1249.

Lee YS, Siemsen E (2017) Task decomposition and newsvendor decision making. *Management Science* 63(10):3226-3245.

Li B, Oliva R, Watson N (2015) Do retail managers rock or paddle the boat? Empirical findings from restocking decisions. Report.

Maddala GS (1986) Limited-dependent and qualitative variables in econometrics (Cambridge university press).

Moreno A, Terwiesch C (2015) Pricing and production flexibility: An empirical analysis of the US automotive industry. *Manufacturing & Service Operations Management* 17(4):428-444.

Moritz BB, Hill AV, Donohue KL (2013) Individual differences in the newsvendor problem: Behavior and cognitive reflection. *Journal of Operations Management* 31(1-2):72-85.

Partridge J (2022) Customers queue at Aldi at dawn for YouTubers' Prime Hydration drink. The Guardian https://www.theguardian.com/media/2022/dec/29/aldi-youtubers-prime-hydration-drink-ksi-logan-paul-queue.

Phillips R, Şimşek AS, Van Ryzin G (2015) The effectiveness of field price discretion: Empirical evidence from auto lending. *Management Science* 61(8):1741-1759.

Rumyantsev S, Netessine S (2007) What can be learned from classical inventory models? A cross-industry exploratory investigation. *Manufacturing & Service Operations Management* 9(4):409-429.

Sachs AL, Becker-Peth M, Minner S, Thonemann UW (2022) Empirical newsvendor biases: Are target service levels achieved effectively and efficiently? *Production and Operations Management* 31(4):1839-1855.

Schweitzer ME, Cachon GP (2000) Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Science* 46(3):404-420.

Sterman JD (1989) Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Science* 35(3):321-339.

Sun J, Zhang DJ, Hu H, Van Mieghem JA (2021) Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Science*.

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185(4157):1124-1131.

Van de Ven WP, Van Praag BM (1981) The demand for deductibles in private health insurance: A probit model with sample selection. *Journal of Econometrics* 17(2):229-252.

van Donselaar KH, Gaur V, van Woensel T, Broekmeulen RA, Fransoo JC (2010) Ordering behavior in retail stores and implications for automated replenishment. *Management Science* 56(5):766-784.

Zhao, H., Xu, L., & Siemsen, E. (2021). Inventory sharing and demand-side underweighting. *Manufacturing & Service Operations Management*, 23(5), 1217-1236.

Chapter 3

Diversifying the Retail Experience: An Empirical Study on Spillover Effects of Experiential Services

3.1 Introduction

The rise of e-commerce and accessibility to online shopping channels have made attracting customers into brick-and-mortar stores one of the biggest challenges facing today's retailers. To give customers more incentives to visit physical stores, retailers have started to rethink the roles of their brick-and-mortars (Gao and Su 2019) by adopting innovative business models. These include Buy-Online-Pickup-in-Store services (Gao and Su 2017; Gallino and Moreno 2014), using stores as a showroom channel (Lim et al. 2023; Bell et al. 2018), and offering coupons and directions to the nearest store via online adds (WeatherAds 2023).

In order to thrive in this fastly changing environment, retailers may need to enrich the customer experience even further since "boring, undifferentiated, irrelevant and unremarkable stores are most definitely dying" (Dennis 2018). To do so, some retailers have

diversified the whole store experience by expanding the services provided. For example, Target features a bar in its Streeterville store in Chicago, which required an investment of approximately $9.6 million (Selvam 2015), enabling shoppers to walk around the store with their drinks. Also, some grocery stores feature restaurants inside; this is called the "grocerant" business model and is deemed the future of food shopping (McGrath 2016). Whole Foods, for instance, generates approximately 19% of its revenue from its prepared foods and bakery section (Statista 2017). Diversifying the retail experience is not limited to grocery retailers. Activewear shops like Athleta and Lululemon host free workout classes in selected stores s (Gurfein 2015; Mutlu et al. 2023), while an Urban Outfitters in Austin introduced an in-store diner concept called Space 24 Twenty (Pendrill 2016). In this paper, we are interested in these experience-enhancing ancillary services, which we call "experiential services."

Research on the value of experiential services is limited and mostly analytical. The first study in operations management literature is the recent work of Mutlu et al. (2023). In that paper, the authors analytically examine the impact of experiential offerings on the price and demand of main products. They complement their models by investigating the profit levels of retailers under different pricing strategies (e.g., conditionally free on purchase). In the retail literature, Ham et al. (2021) conduct a survey to understand the behavioral intentions to consume in grocerant stores. Building on these, our paper is the first empirical examination of experiential services, focusing on their spillover effect. Specifically, our research questions are: 1) How does an additional experiential service affect the measures of main products of the retailer? 2) What are the mechanisms behind this effect, if any? It is important to examine these effects because retailers invest a great deal of resources in experiential services believing that they will generate demand for the main products and attract more customers. But, is it really the case that experiential services drive sales of main products? The impact is not obvious beforehand.

On the one hand, increased time spent in a store could lead customers to purchase

more as extended stay means a greater exposure to products. Also, offering an experiential service may attract new customers, increasing footfall, which may convert into more sales. In addition, experiential services may draw customers towards related products within the main retail section through the complementary sales mechanism (Mutlu et al. 2023). For example, consumers having a drink in the bar inside the Target store in Streeterville could feel compelled to buy crisps or pretzels, which are complementary to the bar offerings. On the other hand, cannibalization could occur such that after spending a certain amount of their budget on the experiential service, customers may be less willing to spend in the main retail section. Some customers may come to the store only for the experiential service, creating store traffic that does not convert into any sales for the main products. For example, free classes offered by active wear shops are subject to such danger (Kadet 2018). Furthermore, the increased crowd could disturb the operations of the retail space, reduce the service level, and deter potential customers who intend to buy. Also, the extended stay, although argued by practitioners as a desired outcome (Buildd 2023), may not be so desirable from an operational efficiency standpoint. Therefore, an empirical examination of the potential spillover effects of experiential services is needed in order to comprehensively evaluate the economics of introducing these services in physical stores.

To investigate this, we collaborate with a leading supermarket chain in the Southeast of United States, which introduced a taproom service in some of its stores. Specifically, we analyze 371,662,651 detailed stock-keeping unit (SKU) transactions from this supermarket chain between June 2018 and April 2020. We combine these transaction-level data with store-level foot traffic data from SafeGraph. To compare the stores with a taproom with the ones without a taproom, we employ the synthetic difference-in-differences (S-DiD) estimator (Arkhangelsky et al. 2021) to tackle the inherent endogeneity of the supermarket chain's decision to introduce the experiential service in certain stores.

Our findings suggest that the introduction of a taproom service increase store-level

performance measures. Specifically, we find an average of 5.97% increase in sales volume, 6.82% in sales revenue, 5.76% in the number of transactions, and 6.18% in assortment size. We then investigate the mechanisms driving these results through additional analyses that (1) estimate heterogeneous effects across departments within a store, (2) evaluate basket-level changes, (3) identify product characteristics, and (4) check for changes in several customer-reach measures. Our results suggest that the store-level increases are mainly driven by the sales increase in a few departments. The argument by the practitioners behind featuring a taproom service is that customers would spend more time in stores, which increase their exposure to more products. Although we find an average increase of 15.49% in time-spent in stores, this increase exposure only translate into higher sales for certain departments in our data. These departments are all food and beverage related, and contain either complementary products to the taproom offerings, or perishable products. The former suggest a complementarity sales mechanism of the experiential service (Mutlu et al. 2023), whereas the latter can be explained by impulse purchases (Sowder 2023) or anticipated regret (Zeelenberg et al. 1996). Furthermore, the analysis indicates that although customers visit the stores more frequently after the introduction of the experiential service by the focal retailer, the average basket size decreases by 1.21%. Despite going against expectations (Maynard 2017; Moore 2021), this behavior aligns with the budget constraints customers face. Because they visit the stores more often, they tend to purchase fewer items during each visit on average. Furthermore, our results point out that a taproom service increases the number of new consumers by 7.14%, suggesting that an experiential service could enhance both intensive and extensive margins.

3.2 Related Literature

There are three streams of literature that are relevant to our work: (1) experiential retailing, (2) bundling, complementarity, and ancillary services, and (3) retail agglomeration, which are reviewed in this section.

The first stream of literature we build on is experiential retailing (Pine and Gilmore 1998). In their seminal piece, Pine and Gilmore (1998) state that experiences are distinct economic offerings that increase the perceived value of an ordinary good or service according to consumers. Although they introduced the concept of experience economy in the 1990s, Mutlu et al. (2023) state that analytical and empirical studies on the issue have been limited. In a conceptual piece, Schmitt (1999) argues that consumers are not only concerned about functional features of goods and services; they are emotional decision makers who care about achieving pleasurable experiences. Experiential retailing is expected to increase overall consumption because of the convenience of having two services at the same location, or because of the enticement of an attractive shopping environment (Kim 2001). Testing whether consumers are indeed more enticed by the experiential retailing concept, previous research utilized interviews and surveys. For example, Ham et al. (2021) investigate the factors increasing the engagement in grocerants. By conducting semi-structured in-depth interviews, Ballantine et al. (2010) suggest that experiential stores attract both utilitarian and hedonic consumers, the latter being dissatisfied with non-experiential retailers. Our paper contributes to this stream of literature by being the first empirical study that investigates the spillover impact of experiential services, providing a more complete picture of the economics of this type of service introduction.

Our paper also relates to the literature on product and service bundles. Bundling products and services can be the main profit generator for firms (Singh et al. 2021) since it allows the exploitation of complementarities (Milgrom and Roberts 1990). For

example, Liu et al. (2010) find strong complementarities between the consumption of core broadband service and of related categories, such as cable television. Dong et al. (2018) report that banks enjoy a profit increase from bundling mobile money with credit services. United Airlines generated more than 15% of its total revenue from ancillary services in 2013 (Wang et al. 2019). Yet, Dong et al. (2018) also report that this positive effect is not prevalent in the whole value chain; in particular, the bundling of complementary services does not create added value for mobile network operators. The bundling and ancillary services literature tend to focus on different pricing strategies (e.g., subscription (Wang et al. 2019)). For example, Cui et al. (2018) show analytically that unbundling the ancillary service is optimal for a uniform-pricing firm if a high portion of consumers who highly value the main service also value the ancillary service. This analytical model assumes that ancillary services cannot be sold independently. This is typically the case in the literature on bundling and ancillary services (Mutlu et al. 2023). Our study's focus, experiential service offerings, is also an additional service, yet in contrast to this stream of literature, these additional service offerings can be enjoyed even without a purchase from retailers' main functions. As Mutlu et al. (2023) point out, the main focus of research on ancillary services is either bundling or pricing separately; yet, with the experiential service offerings, the main question is whether to offer them in the first place (Mutlu et al. 2023). This decision to offer an additional service should also depend on its spillover effects. Therefore, we contribute to this stream by examining the spillover effects of an additional service by relaxing the assumption that this service cannot be sold separately.

The literature on retail agglomeration provides a theoretical basis to understand the impact of offering experiential services. This stream of literature asks why retailers co-locate despite the increased competition risk, and suggests that limited information of customers is the reason (Stahl 1982; Wolinsky 1983). Specifically, the retail agglomeration effect suggests that customers try to minimize their search costs; hence, they are more likely to shop from a retailer located close to other retailers, compared to an isolated

one (Olivares and Cachon 2009). Since customers want to minimize their search cost as well as the traveling cost (Dellaert et al. 1998), they mostly make multipurpose visits instead of single purpose ones (Marianov et al. 2018); hence, they prefer retail clusters. This stream of literature focuses solely on different retailers. Yet, we argue that different offerings of a retailer can be thought of as members of a cluster. Shoppers' desire to minimize their search and traveling costs may drive a demand for experiential services. For example, customers of Target in Streeterville may prefer to have their drinks in the bar inside the shop after their grocery shop, instead of relocating to another bar. Hence, our study contributes to this stream of literature by rethinking and redefining retail clusters.

3.3 Data and Measures

3.3.1 Empirical Setting and Data

To analyze the spillover effects of introducing an experiential service, we collaborated with a supermarket chain that operates more than 500 stores in several states in the Southeast of the United States. This supermarket chain has built taprooms inside some of its stores, where customers can enjoy alcoholic refreshment, coffee, wings, or pizza slices before, after, or during their grocery shop. The first taproom was introduced in January, 2019. As of 2023, there are a total of eight stores featuring the taproom service. The supermarket chain adopted a staggered approach to introduce the taproom service in these eight stores; in other words, stores are treated at different time periods, presented in Table 3.1.

This setting provides an ideal context to examine the spillover effects of introducing an experiential service for several reasons. First, taprooms are considered an experiential service for grocery retailers by previous studies (Mutlu et al. 2023), supermarket practitioners and consultants (Ruback 2023). The reason is that taprooms may transform a

mundane household errand into a social, fun, and interesting experience. Dixon (2017) argues that supermarkets with a taproom or a restaurant are experience destinations, not mere shopping markets. Second, the availability of granular data in the grocery retail context allows us to analyze the effect of the experiential service with respect to various shopping behaviors and product characteristics, such as impulse purchases or perishability. Third, grocery retail chains invest in taprooms with the objective to sell more of their main products. This serves as an ideal context for our study since we are interested in evaluating the spillover effects of experiential services.

Table 3.1: Taproom Introduction Dates

Store	Taproom Opening Day
1	January 18, 2019
2	November 6, 2019
3	August 28, 2019
4	March 27, 2020
5	November 11, 2020
6	December 9, 2020
7	December 9, 2020
8	September 8, 2021

We obtained detailed SKU-transaction level data, which aggregated into a total of 371,662,651 transactions. For each transaction, we have the day, the store, the quantity and the price of the SKUs in the basket, and to which product category the SKU belongs. We do not have the customer ID corresponding to each transaction. Our data encompass 456 product categories. Because of the possible confounding effect of the COVID-19 pandemic, we restrict our main analysis to include data before April 2020, when state-level measures against the pandemic were put in effect, so that we can alleviate the potential confounding impact of pandemic measurements as well as pandemic-induced changes in store demands.

As will be explained in Section 3.5, we adopt the S-DiD estimator to conduct our analysis. This estimator requires a balanced panel. Since some control stores are not observed

on every day in our data, we drop them to have a balanced panel. This model requirement, combined with the time restriction we put because of the COVID-19 pandemic, leads us to have 219,128 store-day observations of four treated and 340 control stores between June 30, 2018, and March 31, 2020. As robustness, we expand the analysis to include seven treated stores and show the results in Section 3.7.1. As observed in Table 3.1, the supermarket chain introduced its first taproom service on January 18, 2019. To ensure that we have enough data on this store before its treatment, we started our data collection period on June 30, 2018.

We combine these supermarket data with SafeGraph's Weekly Patterns data, which provide the estimates of weekly foot traffic around each brick-and-mortar store. Specifically, SafeGraph uses data shared by many mobile app partners to create geospatial data that include the number of unique visitors, the number of visits, median dwell time in minutes, and median distance from home traveled by visitors in meters. Note that these point-of-interest data were released monthly for periods before December 2018, and weekly for the periods after December 2018. This leads us to have 21,710 week-store observations between February 31, 2018 and March 30, 2020. This mobile patterns database of SafeGraph has been recently used by Babar et al. (2023) and Shin et al. (2023).

3.3.2 Analysis Roadmap

As we aim to analyze the spillover effects of introducing the taproom service onto the supermarket's performance related to its main products, our main analyses pertain the store-level measures. Specifically, at the store level, we analyze how the number of transactions, sales volume, sales revenue, and assortment size change with the taproom introduction. Then, we aim to understand the mechanisms behind the results we observe at the store level. To do so, we produce a road map that is informed by the literature and industry perspectives. First, Mutlu et al. (2023) point out that experiential offerings may be a complementary sales mechanism for a retailer's main products; in other words, the

main products that are complementary to the additional service offerings may experience a greater effect. For example, if the spillover effects really exist, consumers having a drink in the bar inside Target Streeterville should be more urged to buy crisps or pretzels, which are complementary to the bar offerings, than to buy detergent. Or in the case of free workout classes of Lululemon, the spillover effect should be stronger for yoga pants or t-shirts compared to swimsuits. Inspired by this, we analyze how the spillover effects are different for different product categories by focusing on heterogeneous effects on departmental outcomes. Second, as Dixon (2017) argues, alcohol is an effective way to keep consumers longer in a store. Since alcohol is one of the main products offered by a taproom service, this inspired us to investigate whether consumers indeed spend more time in treated stores after the taproom service being introduced. Furthermore, he continues his argument by stating that as consumers take more time in a store, they see more products they would have missed if they were quickly in and out. This motivated us to examine how an average basket changes with the introduction of a taproom. Does an average basket include more items now? Or does it include a higher degree of variety in items? Besides, Coolidge (2015) state that featuring a taproom in supermarkets surge demand and foot traffic. If this actually happens, it can explain the store-level results. Hence, we analyze the foot traffic argument by checking the number of visits as well as number of new customers. Furthermore, the literature suggests that alcohol consumption is associated with high levels of impulse purchases (Harnish et al. 2023). This motivated us to examine whether there is an increase in the sales of items that are typically characterised as impulse items.

3.3.3 Unit of Analysis and Variable Operationalizations

To investigate spillover effects of the taproom introduction, we chose store-day as our unit of analysis. The independent variable of interest, $Treated_{it}$, is a binary variable taking the value of 1 if store i introduced a taproom before day t.

Our main dependent variables of interest reflect the store-level performance measures. First, the dependent variable to capture the daily sales quantity is $LogSalesVolume_{it}$, calculated as the logarithm of the sum of all transaction quantities in store i on day t. On top of the sales quantity, we measure $LogSalesRevenue_{it}$ capturing the daily sales revenue, that is, the logarithm of the sum of all transaction revenues in store i on day t. To capture the effect on the store visits frequency, we create $LogNumberofTransactions_{it}$, calculated as the logarithm of the number of transactions that happened in store i on day t. Another dependent variable we analyze is $LogAssortmentSize_{it}$, calculated as the logarithm of the number of SKUs purchased in store i on day t.

To understand how an average basket in treated stores was changed by the taproom service, we create basket-level performance measures. To start with, $LogBasketSize_{it}$ captures the logarithm of the mean of transaction quantities in store i on day t. Next, we take the logarithm of the mean of transaction values in dollars in store i on day t to create $LogBasketSalesRevenue_{it}$. Besides, to understand how variety of products purchased is affected, we create $LogBasketAssortmentSize_{it}$, which is the logarithm of the mean of the number of SKUs purchased in each transaction that took place in store i on day t. Last, we create $LogItemPrice_{it}$ by taking the logarithm of the mean of the average item price in each transaction in store i on day t.

Table 3.2 presents the descriptive statistics of the variables.

3.4 Model-Free Evidence

Prior to specifying formal econometric models, we present model-free evidence on the possible effects of the introduction of a taproom service in grocery stores. To do so, we first aggregate our transactional data into weekly store-level data to obtain a neater picture of trends.We then use the synthetic control method (Xu 2017; Abadie, Diamond, et al. 2010) to obtain the counterfactual control stores that match with treated stores.

Table 3.2: Descriptive Statistics

Variable Name	Description	Mean	Standard Deviation	Minimum	Maximum
$Treated_{it}$	1 if store i has a taproom on day t, 0 otherwise	0.004	0.06	0	1
$LogSalesVolume_{it}$	Logarithm of the number of items sold in store i on day t	8.496	1.42	2.996	11.093
$LogSalesRevenue_{it}$	Logarithm of the dollar amount earned in store i on day t	10.311	1.101	4.588	12.551
$LogNumberofTransactions_{it}$	Logarithm of the number of transactions in store i on day t	6.717	0.84	0.693	8.589
$LogAssortmentSize_{it}$	Logarithm of the number of SKUs sold in store i on day t	8.194	1.441	2.639	10.77
$LogBasketSize_{it}$	Logarithm of the average basket size sold in store i on day t	1.779	0.611	0.381	3.46
$LogBasketSalesRevenue_{it}$	Logarithm of the dollar amount earned from an average basket in store i on day t	3.594	0.315	2.429	4.848
$LogBasketAssortmentSize_{it}$	Logarithm of the number of SKUs in an average basket sold in store i on day t	1.477	0.631	0.216	2.743
$LogItemPrice_{it}$	Logarithm of the average item price (in dollars) in an average basket sold in store i on day t	2.501	0.352	1.401	4.183
$LogMedianDwellTime_{ij}$	Logarithm of the median of time spent in store i at week j, in minutes	2.532	0.348	0	6.201
$LogDistanceFromHome_{ij}$	Logarithm of the median of the distances from store i to visitors' houses at week j, in meters	8.546	0.76	6.653	14.291
$LogNumberofUniqueVisitors_{ij}$	Logarithm of the number of unique visitors in store i at week j	5.7	0.584	1.609	7.526
$LogNumberofWeeklyVisits_{ij}$	Logarithm of the number of visits in store i at week j	5.899	0.605	1.609	7.716

Next, we obtain the averages of several performance indicators of the treated group, and of the control group, shown in Figure 3.1. Specifically, Figure 3.1 includes four plots corresponding to four outcomes. On the x-axis of each plot, we have the time relative to the treatment timing, represented by week 0. As for the y-axis, the top left plot shows the average weekly sales volume of the treated stores and that of the counterfactual control stores. On the top right, we have the average weekly revenue of the treated stores and that of the counterfactual control stores. On the bottom left, we include the average weekly number of transactions of the treated stores and that of the counterfactual control stores. Finally, the bottom right plot demonstrates the average weekly assortment size of the treated stores and that of the counterfactual control stores.

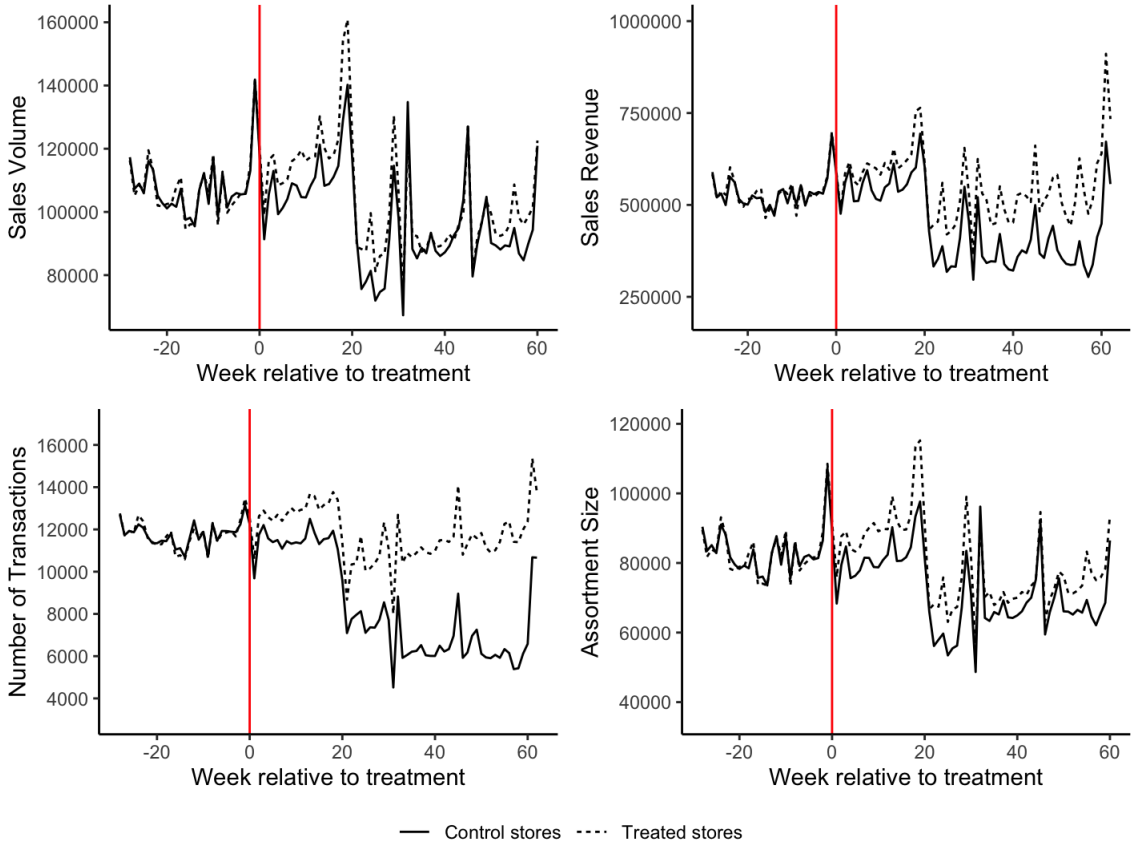

Figure 3.1: Model-free Evidence

On each of these plots, we observe that the lines are practically on top of each other until the treatment timing. After the treatment, we observe that the lines diverge. On each

plot, we see that the average of the treatment group is greater than that of the control group, showing that the taproom service increased all of the outcomes examined.

Although the plots demonstrate a gap between the lines that represent the average outcomes for control and treatment groups, we cannot know whether the difference is statistically significant. To check this, we next run eight paired sample t-tests, two for each outcome: one for until the treatment averages, and one for after the treatment averages. Table 3.3 shows the results. We observe that, for each outcome variable, the mean difference between the averages of the control group and the treated group before the treatment is not statistically significantly different from 0. Yet, all of the mean differences are significant for the averages after the treatment. These results present initial evidence of the impact of introducing a taproom service on store-level performance measures.

Table 3.3: Model-free Evidence: Paired Sample t-test Results

Panel A: Sales Volume		
	Before Treatment	After Treatment
Mean Difference	154.749	6659.9***
Confidence Interval (95%)	[-698.931, 1008.429]	[5069.242, 8250.557]
Panel B: Sales Revenue		
	Before Treatment	After Treatment
Mean Difference	3051.691	111,172.4***
Confidence Interval (95%)	[-2136.234, 8239.617]	[96,238.54, 126,106.2]
Panel C: Number of Transactions		
	Before Treatment	After Treatment
Mean Difference	4.749	3494.159***
Confidence Interval (95%)	[-71.253, 80.751]	[3053.927, 3934.39]
Panel D: Assortment Size		
	Before Treatment	After Treatment
Mean Difference	36.258	7756.277***
Confidence Interval (95%)	[-663.875, 736.391]	[6610.623, 8901.931]

*p<0.1; **p<0.05; ***p<0.01.

3.5 Empirical Modeling and Identification

The previous section shows the initial evidence of the impact of featuring a taproom on store performance. Yet, to assess the causal effect of this, there are several estimation issues we need to address. First, the studied supermarket chain's decision to introduce a taproom service in certain stores is not random. For instance, this chain might have added the taproom service to specific stores that tend to record higher sales or higher customer traffic. This endogeneity may lead to bias in our estimates if not accounted for. Second, the studied supermarket chain did not introduce the taproom service at the same time in treated stores, meaning that our data have a staggered treatment timing design. We cannot employ the classic approach of the dynamic two-way fixed effects DiD model since it would produce biased estimations when there is heterogeneity in treatment effect and when there is a staggered treatment timing design (Baker et al. 2022; Callaway and Sant'Anna 2021; Wooldridge 2021). Third, there are only a few treated stores in our data, especially since we restrict our examination period to avoid the confounding impact of the COVID-19 pandemic. The synthetic control model addresses having only one (Abadie and Gardeazabal 2003) or few (Xu 2017) treated units by creating counterfactual controls that have perfectly matches the pre-treatment trends of treated units, hence relaxing the parallel trends assumption. Hence, we employ the S-DiD estimator with a staggered treatment design (Arkhangelsky et al. 2021).

The S-DiD estimator, developed by Arkhangelsky et al. (2021), merges the strengths of the DiD model and those of the synthetic control model developed by Abadie and Gardeazabal (2003) and Abadie, Diamond, et al. (2010). Just like the synthetic control method, it creates an artificial counterfactual for each treated store by putting weights on control stores to match the pre-treatment outcome trends between treated and control groups (Farronato et al. 2023; Arkhangelsky et al. 2021), and it can be employed when there are very few, even single, treated units (Arkhangelsky et al. 2021). On top of

that, this method alleviates the concerns related to researcher discretion in the choice of pre-treatment period length by estimating weights assigned to pre-treatment time periods (Farronato et al. 2023). By weighting the time periods, the S-DiD estimator balances pre-treatment periods with post-treatment periods; if a certain pre-treatment time period has a higher power in predicting the post-treatment outcomes, it receives a higher weight (Arkhangelsky et al. 2021; Berman and Israeli 2022). This weighting of time periods and units captures a great degree of the variance in the outcome; hence, the S-DiD estimator has a better precision than the synthetic control estimation does. Another advantage of this method is that it does not strictly impose a parallel-trends assumption (Berman and Israeli 2022). Given that the parallel trend assumption is often violated (Xu 2017), this method relaxes it by using a a weighted average of outcomes from control units to predict the outcomes of the treated stores "as if" they did not introduce the taproom service. The weights are chosen to optimally match the pre-exposure outcomes of the treated stores, and thus, they capture any possible trends that might affect identification without requiring a parallel-trends assumption (Berman and Israeli 2022). Lastly, as Berman and Israeli (2022) note, the S-DiD estimator is consistent even under an unobserved correlation between the treatment exposure and the store-level time trends, alleviating the concern of endogeneity.

Originally, Arkhangelsky et al. (2021) develop this method for settings with block assignment, in other words, for settings in which treatment is received by all treated units at the same point in time. Following the procedure developed by Ben-Michael et al. (2022), the S-DiD can be estimated for each period of treatment adoption separately. Then, the estimated average treatment effects on treated (ATTs) are combined with a weighted average. This approach is recently employed by Berman and Israeli (2022) in their examination of how retail analytics dashboard adoption affects performance of online retailers.

For a panel of N units and T time periods, our S-DiD model takes the following form:

$$(\hat{\tau}, \hat{\mu}, \hat{\alpha}, \hat{\beta}) = \underset{\tau,\mu,\alpha,\beta}{\arg\min}\{\sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{it} - \mu - \alpha_i - \beta_t - Treated_{it}\tau)^2\hat{\omega}_i\hat{\lambda}_t\} \tag{3.1}$$

where μ, α_i, and β_t represent an intercept, store fixed effects, and day fixed effects, respectively. The inclusion of these fixed effects allow us to control for the time invariant unobservables. $\hat{\omega_i}$, unit weights, are found to align pre-exposure trends in the outcome of control units with those for the exposed units (Arkhangelsky et al. 2021). $\hat{\lambda_t}$, time weights, are estimated to balance pre-treatment periods with post-treatment ones (Arkhangelsky et al. 2021). These weights are used to create the synthetic controls, with which a two-way fixed effects regression is used to estimate τ, which is the average causal effect of the treatment.

We employ the placebo variance estimation algorithm of Arkhangelsky et al. (2021) to obtain the standard errors for $\hat{\tau_i}$. We chose the placebo algorithm over bootstrap- and jacknife-based methods in our main analysis since it is more reliable when the number of treated units is low (Arkhangelsky et al. 2021). This method uses placebo predictions by replacing the treated units with control units to estimate the standard errors and confidence intervals. As robustness, we use alternative variance estimation algorithms and discuss the approaches in Section 3.7. In the estimation of standard errors, we set the number of repetitions to 50. According to Mooney et al. (1993), 50 to 200 replications are generally suitable for estimates of standard errors and thus are acceptable for normal-approximation confidence intervals.

3.6 Results

In this section, we presents the estimation results to assess spillover impact of introducing a taproom service. In Section 3.6.1, results regarding the store-level performance measures are presented. In Sections 3.6.2, 3.6.3, 3.6.4, and 3.6.5, the mechanisms behind the store-level results are investigated. Specifically, in Section 3.6.2, we discuss the het-

erogeneous effects across departments within a store. Section 3.6.3 discusses how an average basket changes post-exposure to treatment. In Section 3.6.4, basket changes related to specific product characteristics are reported. Finally, Section 3.6.5 presents the changes in several customer-reach measures.

3.6.1 Effects on Store Performance

In Table 3.4, we present the store-level results. Overall, our findings show a significant and positive effect of the introduction of a taproom service in grocery stores. Specifically, the estimated ATT of $Treated_{it}$ on $LogSalesVolume_{it}$ is 0.058 ($p<0.01$), showing that the total quantity sold is increased by an average of 5.97% after the taproom introduction in treated stores. The estimated ATT on $LogSalesRevenue_{it}$, 0.066 ($p<0.01$), shows that sales revenue in treatment stores increases by 6.82% relative to control stores. We next turn to the estimated ATT on $LogNumberofTransactions_{it}$, which is 0.056 ($p<0.01$), corresponding to an increase of 5.76%. Additionally, the estimated ATT for $LogAssortmentSize_{it}$, 0.06 ($p<0.01$), suggests that the variety of SKUs purchased on an average day increases by 6.18% after the introduction of a taproom service.

These positive effects on $LogSalesVolume_{it}$ and $LogSalesRevenue_{it}$ can be explained by several mechanisms. First, offering an experiential service may attract new customers, which leads to increased footfall. The newly-attracted customer base will be presented with the retailer's main products, which increases the probability of making a purchase. Second, on top of attracting new customers, the existing customers may increase their visit frequencies because of the taproom, which again would increase the footfall. The estimated ATT on $LogNumberofTransactions_{it}$ supports this explanation that experiential services increase the foot traffic to stores.

After obtaining these store-level results, we ask why? What are the mechanisms that drive the positive impact at the store-level? The following sections investigate various possible mechanisms to answer this question.

Table 3.4: Store-level Estimation Results

	Log Sales Volume	Log Sales Revenue	Log Number of Transactions	Log Assortment Size
ATT	0.058***	0.066***	0.056***	0.06***
	(0.02)	(0.023)	(0.019)	(0.019)
Confidence Interval (95%)	[0.018, 0.098]	[0.02, 0.111]	[0.018, 0.095]	[0.023, 0.098]
Store Fixed Effects	YES	YES	YES	YES
Day Fixed Effects	YES	YES	YES	YES
Sample Size	219,128	219,128	219,128	219,128

*p<0.1; **p<0.05; ***p<0.01. Numbers in parentheses are placebo standard errors bootstrapped for 50 times.

3.6.2 Heterogeneous Effects

To better understand what drives the estimated store-level effects, we next conduct a separate analysis for individual departments. The 14 departments and example product categories within each department are provided in 3.5.

By these separate analyses, we recognise the potential variability in how the introduction of a taproom service might affect different store sections. For example, the effect of the taproom on consumer behavior and purchasing patterns may be greater or different for snacks and beverages than it is for cleaning supplies. Shopping habits and customer choices, such as the types of products purchased, might be uniquely influenced by the taproom introduction. By examining each department separately, we aim to reveal these variations, providing insights into the positive effect observed at the store-level.

Table 3.6 reports the estimated ATT for $LogSalesVolume_{it}$, $LogSalesRevenue_{it}$, and $LogAssortmentSize_{it}$ in each of these departments. Our analyses reveal that the positive impact we observe at the store-level is driven by the impact of the taproom service on certain departments. Particularly, Table 3.6 shows that the significant ($p<0.05$) and positive impact is observed for the following departments: Alcohol & Beverages, Bakery, Meat & Seafood, Produce, and Snacks. It is worth noting that all these departments are related to food and beverages. The effect of the taproom service is never significant in

Table 3.5: Departments and Exemplary Products

Department	Products
Alcohol & Beverages	Beer, wine, tea, coffee, juice, soft drinks, spirits etc.
Bakery	Bagels, pies, pancake mix, carrot cake etc.
Deli	Stuffed olives, pre-made sandwiches, sliced ham, Bologna sausage, salami etc.
Dairy	Yogurt, butter, cream cheese, milk, sour cream etc.
Meat & Seafood	Poultry, smoked meats, sausages, bacon, raw shrimp, shellfish, salmon etc.
Produce	Tomato, eggplant, onion, melon, kiwi, herbs, potatoes etc.
Home & Decor	Pool supplies, fabric dye, kitchen gadgets, candles, air filters, cookware etc.
Frozen Food	Frozen pizza, frozen vegetables, frozen fish etc.
Personal Care	Sunscreen, toothpaste, tampons, deodorants, shampoos etc.
Household Cleaning	Napkins, detergent, sponges etc.
Pharmacy	Pain relief, digestive help etc.
Food Cupboard	Olive oil, sugar, pasta, tomato paste, sugar, mustard, cereal etc.
Snacks	Candy, crisps, cookies etc.
Miscellaneous	Greeting cards, pet supplies, tobacco, baby formula, bait etc.

Table 3.6: Separate Department Analysis

	Log Sales Volume	Log Sales Revenue	Log Assortment Size
Alcohol & Beverages	0.076** [0.015, 0.137]	0.087*** [0.036, 0.138]	0.088*** [0.021, 0.155]
	(0.031)	(0.026)	(0.034)
Bakery	0.041** [0.001, 0.081]	0.048** [0.001, 0.094]	0.039 [-0.01, 0.088]
	(0.02)	(0.024)	(0.025)
Deli	0.064 [-0.016, 0.144]	0.072 [-0.02, 0.164]	0.067 [-0.033, 0.168]
	(0.041)	(0.047)	(0.051)
Dairy	0.049 [-0.005, 0.102]	0.045 [-0.021, 0.112]	0.052 [-0.02, 0.124]
	(0.027)	(0.034)	(0.037)
Meat & Seafood	0.049** [0.003, 0.096]	0.055** [0.003, 0.107]	0.044* [-0.002, 0.09]
	(0.024)	(0.026)	(0.023)
Produce	0.096*** [0.028, 0.163]	0.121*** [0.037, 0.206]	0.112*** [0.043, 0.18]
	(0.034)	(0.043)	(0.035)
Home & Decor	0.018 [-0.058, 0.093]	0.004 [-0.068, 0.076]	0.02 [-0.028, 0.067]
	(0.039)	(0.037)	(0.024)
Frozen Food	0.033 [-0.041, 0.106]	0.03 [-0.044, 0.105]	0.036 [-0.035, 0.107]
	(0.038)	(0.038)	(0.036)
Personal Care	0.06 [-0.007, 0.128]	0.045 [-0.014, 0.105]	0.059 [-0.004, 0.121]
	(0.034)	(0.031)	(0.032)
Household Cleaning	0.039 [-0.012, 0.09]	0.046 [-0.005, 0.098]	0.033 [-0.037, 0.102]
	(0.026)	(0.026)	(0.035)
Pharmacy	0.036 [-0.044, 0.115]	0.031 [-0.055, 0.116]	0.036 [-0.037, 0.109]
	(0.041)	(0.044)	(0.037)
Food Cupboard	0.052 [-0.003, 0.107]	0.052 [-0.002, 0.113]	0.055 [-0.009, 0.118]
	(0.028)	(0.031)	(0.032)
Snacks	0.026** [0.0002, 0.053]	0.039** [0.002, 0.077]	0.022 [-0.105, 0.149]
	(0.013)	(0.019)	(0.065)
Miscellaneous	0.014 [-0.101, 0.129]	0.139 [-0.013, 0.29]	0.043 [-0.083, 0.169]
	(0.059)	(0.077)	(0.064)
Store Fixed Effects	YES	YES	YES
Day Fixed Effects	YES	YES	YES

*p<0.1; **p<0.05; ***p<0.01. Specification includes store fixed effects and week fixed effects. Numbers in parentheses are placebo standard errors bootstrapped for 50 times.

our data for departments that are unrelated to food and beverages, such as Pharmacy, or Household Cleaning. The positive impact observed at the departments of Alcohol & Beverages, Bakery, and Snacks can be explained by the complementary sales mechanism of offering an experiential service (Mutlu et al. 2023). Snacks and bakery products are complementary to alcohol and coffee, the main products offered at the taproom service. Yet, we observe that it is not only the departments that are complementary to the taproom offerings (like Snacks) that enjoy a positive effect of the taproom introduction. Interestingly, the departments of Meat & Seafood and Produce also experience a positive and significant ($p<$0.05) impact. One potential explanation for the effect at the Meat & Seafood department is again the complementary sales mechanism. Chicken wings, ribs, or subs are some of the main products offered by the taproom in addition to alcohol and beverages. Consuming these, or even seeing those options offered, could increase the likelihood of the purchase of similar items. Yet, the complementary sales mechanism fails to explain why we observe a positive and significant effect for the department of Produce. One potential mechanism behind it is impulse purchasing. Sowder (2023) suggests that around 40% of produce sales at the grocery stores are from impulse buys, given that produce is typically arranged in the shelves with a visual appeal, which could lead to impulse purchases. Another potential explanation for the estimated ATTs for Produce is anticipated regret. Anticipated regret refers to a situation in which customers anticipate a post-purchase regret; hence, they ex ante make choices to minimize it (Zeelenberg et al. 1996). The impact of anticipated regret on consumers' shopping behaviors have been reported in various contexts (Zeelenberg et al. 1996; Jiang et al. 2017; Zeelenberg 1999; Simonson 1992). In our context, the taproom offerings are products typically deemed as unhealthy, such as beer, wine, chicken wings, or subs. Customers may try to minimize their potential regret after consuming these products by purchasing the main products of the retailer that are healthier. This would explain why we observe an increase in the sales of the SKUs within the department of Produce.

3.6.3 Changes in Basket Composition

In this section, we analyze how an average basket changes after the introduction of a taproom service to further understand the mechanisms leading to the positive estimated results for the store-level performance metrics. We run the S-DiD estimator for $LogBasketSize_{it}$, $LogBasketSalesRevenue_{it}$, $LogBasketAssortmentSize_{it}$, and $LogItem\ Price_{it}$. Table 3.7 shows that the impact of the taproom service did not significantly change the $LogBasketSalesRevenue_{it}$ and $LogBasketAssortmentSize_{it}$. Interestingly, we observe a negative and significant ($p<0.05$) effect on $LogBasketSize_{it}$, demonstrating that an average basket is lighter by 1.21%. Yet, we see a positive and significant ($p<0.05$) effect on $LogItemPrice_{it}$, showing that the average item price in the average basket is higher roughly by 1.71%.

Table 3.7: Basket-level Estimation Results

	Log Basket Size	Log Basket Sales Revenue	Log Basket Assortment Size	Log Item Price
ATT	-0.012**	-0.009	-0.007	0.017**
	(0.005)	(0.012)	(0.009)	(0.008)
Confidence Interval (95%)	[-0.022, -0.001]	[-0.032, 0.014]	[-0.026, 0.011]	[0.001, 0.033]
Store Fixed Effects	YES	YES	YES	YES
Day Fixed Effects	YES	YES	YES	YES
Sample Size	219,128	219,128	219,128	219,128

*p<0.1; **p<0.05; ***p<0.01. Numbers in parentheses are placebo standard errors bootstrapped for 50 times.

The negative ATT estimate for $LogBasketSize_{it}$, may first seem counterintuitive given the positive ATT estimates we report at the store-level. Yet, this result should be interpreted in tandem with the increase in $LogNumberofTransactions_{it}$. So, while consumers are making more frequent purchases and spending more overall, they are buying fewer items per transaction on average. This is also in line with the fact that consumers have budget constraints. Since their allocated budget to supermarket spending is limited, and since they visit the store more frequently, they now buy less per transaction. Related to

the limits of the budget, consumers might also allocate a portion of their budget towards spending at the taproom, thereby affecting their overall expenditure for the main products of the retailer. Overall, with this analysis, we conclude that the introduction of the taproom service leads to more frequent visits, and fewer but more expensive purchases per visit. Consequently, there is no statistically significant change in the average basket value.

3.6.4 Product Characteristics

As we report in Section 3.6.2, the impact of the taproom service varies for different departments. Also in Section 3.6.3, we report that the average item price in the average basket increases by 1.7%. These results suggest that the taproom service may switch the consumer behavior such that consumers' propensity to buy certain products increases. Inspired by this, in this section, we analyze whether how the spillover effects of the taproom service change for different products.

The first product characteristics we analyze is related to impulse purchasing. Previous research has associated alcohol consumption with poor impulse control and unplanned purchases (Harnish et al. 2023). Given that the main product category of the taproom service is alcohol, there could be an increase in impulse items sales in treated stores in our setting, too. To test this idea, we first conduct a survey following Narasimhan et al. (1996) and Trivedi et al. (2017). The survey asks the respondents to select the closest rating to them on 5-point Likert scale, where 5 means "very often" and 1 means "very rarely," to the following two statements: "I often buy the products within this category on a whim when I pass by it in the store." and "I typically like to buy the products within this category when the urge strikes me." The categories referred in these statements are 14 departments that we create and analyze in Section 3.6.2. In the beginning of the survey, exemplary products within each category are given to the respondents to ensure that the respondents understand what product categories represent. The survey is conducted at a business school, utilizing students as respondents, leading us to have a total of 86

responses. For each product category, we take the average score of the responses to two questions. The categories and the average impulse ratings are as follows: Snacks = 3.645, Bakery = 3.365, Alcohol & Beverage = 3.00, Dairy = 2.86, Personal Care = 2.76, Meat & Seafood = 2.65, Produce = 2.595, Deli = 2.565, Household Cleaning = 2.435, Home & Decor = 2.425, Miscellaneous = 2.405, Food Cupboard = 2.395, Frozen Food = 2.22, and Pharmacy = 2. We create a binary variable that takes the value of 1 if the impulse rating is greater than or equal to 2.5, 0 otherwise. Then, for each basket, we calculate an $ImpulseBasketRatio$ by dividing the basket size by the number of impulse items in the basket. To obtain what an average basket looks like in store i on day t, we take the daily-store average of $ImpulseBasketRatio$s. In Table 3.8, we observe that the estimated effect in this variable is positive and significant ($p<0.05$), indicating a 0.7% increase in impulse purchasing in the average basket of treated stores after the treatment. To investigate whether the observed change at the basket level is reflected at the store level, we create $ImpulseRatio_{it}$ for each store i on day t by dividing the daily store sales by the number of impulse items sold in that store on that day. The estimated effect in this variable is positive and significant ($p<0.05$) as well, demonstrating a 0.5% increase in impulse purchasing in treated stores after the treatment.

Table 3.8: Impulse items

	Impulse Basket Ratio	Impulse Ratio
ATT	0.007***	0.005***
	(0.001)	(0.002)
Confidence Interval (95%)	[0.004, 0.009]	[0.001, 0.008]
Store Fixed Effects	YES	YES
Day Fixed Effects	YES	YES
Sample Size	219,128	219,128

*p<0.1; **p<0.05; ***p<0.01. Numbers in parentheses are placebo standard errors bootstrapped for 50 times.

Another product characteristic we consider is perishability. According to our results in Table 3.6, the treatment effect on the sales volume, revenue, and assortment size for the

products within the departments of Meat & Seafood and Produce is positive and significant. Since both of these product categories contain highly perishable SKUs, we further investigate whether there is a particular increase in the perishable products' sales with the treatment. To do so, we hand code each SKU, based on the category they belong to, as perishable or not (i.e., 1 if perishable, 0 otherwise). Then, we create each basket's perishability proportion by dividing the number of perishable items in a basket to the basket size. We next take the mean of these to get the average basket perishability proportion in store i on day t to obtain $BasketPerishabilityRatio_{it}$. The estimated effect in this variable is positive and significant ($p<0.05$), indicating a 0.4% increase in the quantity of perishable items in the average basket of treated stores after the treatment. To investigate whether the observed change at the basket level is reflected at the store level, we create $StorePerishabilityRatio_{it}$ for each store i on day t by dividing the daily store sales by the number of perishable items sold in that store on that day. The estimated effect in this variable is positive and significant ($p<0.05$) as well, demonstrating a 0.4% increase in perishable items' sales in treated stores after the treatment.

Table 3.9: Perishable items

	Basket Perishability Ratio	Store Perishability Ratio
ATT	0.004**	0.004**
	(0.001)	(0.001)
Confidence Interval (95%)	[0.002, 0.005]	[0.001, 0.007]
Store Fixed Effects	YES	YES
Day Fixed Effects	YES	YES
Sample Size	219,128	219,128

*p<0.1; **p<0.05; ***p<0.01. Numbers in parentheses are placebo standard errors bootstrapped for 50 times.

3.6.5 Customer Reach Analysis

In this section, we utilize SafeGraph's Weekly Patterns data to continue to understand what is behind the results we observe at the store level. These geospatial data at the

Table 3.10: Customer Reach Analysis Results

	Log Median Dwell Time	Log Distance From Home	Log Number of Unique Visitors	Log Number of Weekly Visits
ATT	0.144**	0.108	0.069***	0.071***
	(0.073)	(0.076)	(0.023)	(0.022)
Confidence Interval (95%)	[0.019, 0.269]	[-0.042, 0.257]	[0.043, 0.094]	[0.029, 0.113]
Store Fixed Effects	YES	YES	YES	YES
Week Fixed Effects	YES	YES	YES	YES
Sample Size	21,710	21,710	21,710	21,710

*p<0.1; **p<0.05; ***p<0.01. Numbers in parentheses are placebo standard errors bootstrapped for 50 times.

week-store level provide us with the number of unique visitors, number of visits, median dwell time in minutes, and median distance from home traveled by visitors in meters. We take the logarithm of all these dependent variables before running the S-DiD models. Since these data are weekly, as opposed to daily, like our supermarket data, the sample size is lower in this analysis.

The results are shown in Table 3.10. We observe that the estimated ATT for $LogMedian\ DwellTime_{ij}$ is 0.144 (p<0.05), suggesting an increase in average weekly time spent of roughly 15.49%. This could explain the increased store-level sales and assortment size. By spending more time in the store, consumers are now seeing more products, increasing their purchase likelihood.

We also observe that the estimated ATT for $LogNumberofWeeklyVisits_{ij}$ is 0.071 (p<0.05), corresponding to an increase by 7.36%. This result from this measure from the SafeGraph's Weekly Patterns data is consistent with what we report regarding the impact on $LogNumberofTransactions_{it}$ in Section 3.6.1. Another measure to assess the store traffic is $LogNumberofUniqueVisitors_{ij}$. Since we do not have the customer ID corresponding to each transaction in the supermarket data, this measure allows us to examine whether the taproom service attracts new consumers to treated stores. The estimated ATT for this variable is 0.069 (p<0.01), meaning that the introduction of a taproom service enhances the total influx of distinct consumers to the store. This change is roughly a

7.14% increase. Lastly, we observe that the estimated ATT for $LogDistanceFromHome_{ij}$ is not significant, indicating that the taproom service did not change the geographical reach and effective trading radius. Combining these, we suggest that the increased time spent, the increased frequency in visits, and the increased number of new customers explain the store-level results. Actually, we see that store-level results are driven by both intensive and extensive margins. The number of new consumers can be viewed as extensive margins whereas the frequency with which stores are visited can be seen as intensive margins. Given that both are significant in our analysis, we conclude that the store-level results are a product of both.

3.7 Robustness Checks

In this section, we present our additional analyses to test the robustness of our main results. In Section 3.7.1, we utilize our data in differing ways by changing the number of treated stores included in the analysis. In Section 3.7.2, we incorporate demographic and economic covariates in the matching procedure. In Section 3.7.3, we employ two other estimators to run our analysis. Section 3.7.4 presents the falsification tests while Section 3.7.5 shows the results when we employ an alternative standard error estimation algorithm.

3.7.1 Data Utilization

In this section, we present the robustness checks pertaining to how we utilize our data. In our main analysis, we use data observed before April 2020 to ensure that the measures against the COVID-19 pandemic do not interfere with our estimation. One store of the studied supermarket chain introduced the taproom on the 27th of March, 2020. This means that, in our main analysis, this store only had four days of observations as treated. Hence, in this analysis, we drop this store from our data to assess whether the limited

post-treatment data drives the estimated treatment effects. The results are shown in Table 3.11. We observe that all of the estimated ATTs remain significant ($p<0.05$) and positive, and the magnitudes of the effects are very similar to what we obtain in the main analysis.

Table 3.11: Store-level Results: Three Treated Stores

	Log Sales Volume	Log Sales Revenue	Log Number of Transactions	Log Assortment Size
ATT	0.059**	0.066***	0.057**	0.061***
	(0.027)	(0.021)	(0.023)	(0.023)
Confidence Interval (95%)	[0.007, 0.111]	[0.025, 0.107]	[0.012, 0.103]	[0.016, 0.106]
Store Fixed Effects	YES	YES	YES	YES
Day Fixed Effects	YES	YES	YES	YES
Sample Size	218,491	218,491	218,491	218,491

*p<0.1; **p<0.05; ***p<0.01. Numbers in parentheses are placebo standard errors bootstrapped for 50 times.

Next, we choose to not restrict our data period by the timing of the state level measures against the COVID-19 pandemic, which allows us to leverage all of the data we have and to include a larger portion of the treated stores. The last day we have in our data is 31st of March, 2021. As Table 3.1 shows, the last taproom introduction was on 8th of September, 2021. Since we do not have data until that point, we are not able to add this last store in our analysis. Yet, we are still able to incorporate the other three treated stores, leading us to have a total of seven treated stores. Table 3.12 shows the results. We observe that all of the estimated ATTs are significant ($p<0.05$) and positive. Interestingly, for all four dependent variables, the magnitudes of the ATTs are greater in this analysis compared to our main results (Table 3.4), suggesting the latter is conservative. This further supports the robustness of our results.

Table 3.12: Store-level Results: Seven Treated Stores

	Log Sales Volume	Log Sales Revenue	Log Number of Transactions	Log Assortment Size
ATT	1.101***	1.206***	0.822***	1.089***
	(0.433)	(0.333)	(0.16)	(0.288)
Confidence Interval (95%)	[0.253, 1.949]	[0.553, 1.859]	[0.508, 1.136]	[0.525, 1.654]
Store Fixed Effects	YES	YES	YES	YES
Day Fixed Effects	YES	YES	YES	YES
Sample Size	404,000	404,000	404,000	404,000

*p<0.1; **p<0.05; ***p<0.01. Numbers in parentheses are placebo standard errors bootstrapped for 50 times.

3.7.2 Incorporating Covariates

Our main estimator, S-DiD, creates counterfactual control stores. These stores' pre-treatment trends in outcome variables match the pre-treatment trends of the treated stores' outcome variables (Arkhangelsky et al. 2021). That is why, studies utilizing this estimator, such as Berman and Israeli (2022), do not use exogenous time-varying covariates to create the control units. Yet, as noted by Arkhangelsky et al. (2021) and Clarke et al. (2023), relevant covariates can be included in the estimation procedure to ensure that their impact is accounted for during matching. First, the impact of the change in those covariates on the outcome is removed, then the S-DiD algorithm is applied on the residuals (Clarke et al. 2023). In this section, we utilize this ability of the model to incorporate covariates, and add two additional variables: The first is the median household income in inflation-adjusted dollars of the ZIP code area within which each store is located. The second one is the population of the ZIP code area within which each store is located. Since economic and demographic factors may affect the store outcomes and the decision to choose which stores get exposed to the treatment, we choose to use these variables from the 2018, 2019, and 2020 American Community Survey by Census Block Group (CBG) data. In this dataset, some ZIP codes are not available for 2020. Hence, we drop those stores that are located in those ZIP code areas, leading us to have four treated and

216 control stores.

Table 3.13 shows the estimation results. We observe that the estimated ATTs for $LogSalesVolume_{it}$, $LogSalesRevenue_{it}$, $LogNumberofTransactions_{it}$, and $LogAssortmentSize_{it}$ remain significant ($p<0.05$) and positive when we utilize some demographic and economic factors in the matching procedure.

Table 3.13: Store-level Results: Incorporating Census Data

	Log Sales Volume	Log Sales Revenue	Log Number of Transactions	Log Assortment Size
ATT	0.057**	0.055**	0.058**	0.062**
	(0.029)	(0.025)	(0.026)	(0.031)
Confidence Interval (95%)	[0.001, 0.113]	[0.006, 0.104]	[0.006, 0.109]	[0.001, 0.123]
Store Fixed Effects	YES	YES	YES	YES
Day Fixed Effects	YES	YES	YES	YES
Sample Size	140,140,491	140,140	140,140	140,140

*p<0.1; **p<0.05; ***p<0.01. Numbers in parentheses are placebo standard errors bootstrapped for 50 times.

3.7.3 Alternative Estimation Strategies

Porreca (2022) compares the S-DiD estimator with other matching and DiD techniques for a staggered treatment timing design. His simulation results support the argument of Arkhangelsky et al. (2021) on the precision and the reliability of the S-DiD estimator. Nevertheless, in this section, to test whether our results are robust to different estimators, we utilize the "group time average treatment effect estimator" developed by Callaway and Sant'Anna (2021) and the generalized synthetic control method (Xu 2017).

Callaway and Sant'Anna DiD Framework

Callaway and Sant'Anna (2021) extends the classic two-way fixed effects DiD estimator for cases in which units receive the treatment at different periods, and after they are exposed to the treatment, they remain as treated. This estimator decomposes a staggered

design into smaller DiDs and estimate an ATT for each DiD (Xu 2023) by picking different control groups for each treatment period. In other words, ATTs are estimated separately by group and by time period. These separate ATTs are aggregated into group-time average treatment effects. If differing lengths of exposure to the treatment needs to be taken account, Callaway and Sant'Anna (2021) proposes a dynamic approach to do so. This method has been recently used by Bekkerman et al. (2023) to examine the impact of short term rental platforms on residential real estate investment, and by Eftekhari et al. (2023) to investigate how Health Information Exchanges affect referral decisions.

Table 3.14: Callaway and Sant'Anna Estimator Results

		Log Sales Volume	Log Sales Revenue	Log Number of Transactions	Log Assortment Size
Never-treated as Controls	Dynamic-aggregated ATT	0.161**	0.156**	0.079**	0.147**
		(0.029)	(0.034)	(0.025)	(0.025)
		[0.104, 0.219]	[0.09, 0.223]	[0.03, 0.127]	[0.098, 0.197]
	Group-aggregated ATT	0.099**	0.083**	0.01**	0.08**
		(0.003)	(0.004)	(0.003)	(0.003)
		[0.092, 0.105]	[0.076, 0.091]	[0.005, 0.015]	[0.074, 0.086]
Later-treated as Controls	Dynamic-aggregated ATT	0.161**	0.156**	0.079**	0.147**
		(0.031)	(0.032)	(0.022)	(0.031)
		[0.1, 0.222]	[0.094, 0.218]	[0.036, 0.121]	[0.086, 0.208]
	Group-aggregated ATT	0.099**	0.083**	0.01**	0.08**
		(0.003)	(0.004)	(0.003)	(0.003)
		[0.092, 0.105]	[0.076, 0.091]	[0.004, 0.015]	[0.074, 0.086]
Store Fixed Effects		YES	YES	YES	YES
Day Fixed Effects		YES	YES	YES	YES
Sample Size		219,128	219,128	219,128	219,128

*p<0.1; **p<0.05; ***p<0.01. Numbers in parentheses are bootstrapped standard errors.

To run this estimator, we use the *did library* in R, with both options of group aggregation and dynamic aggregation. Also, the *did library* allows using either never-treated groups or later-treated ones as control groups. We run our models by utilizing both options. The results are presented in Table 3.14. There are four estimated ATTs for each ourcome variable. We observe that all ATTs are significant and positive, corroborating our main findings.

Table 3.15: Generalized Synthetic Control Method Estimation Results

	Log Sales Volume	Log Sales Revenue	Log Number of Transactions	Log Assortment Size
ATT	0.062**	0.062**	0.117***	0.087*
	(0.029)	(0.032)	(0.046)	(0.048)
Confidence Interval (95%)	[0.005, 0.12]	[0.001, 0.124]	[0.026, 0.208]	[-0.008, 0.181]
Store Fixed Effects	YES	YES	YES	YES
Day Fixed Effects	YES	YES	YES	YES
Sample Size	219,128	219,128	219,128	219,128

*p<0.1; **p<0.05; ***p<0.01. Numbers in parentheses are bootstrapped standard errors for 50 times.

Generalized Synthetic Control Method

Our main model, the S-DiD (Arkhangelsky et al. 2021), in an extension of the synthetic control method. Another extension to this method is using factor-augmented models (Xu 2023) that rewrites the fixed effects as interactive. We use one such model in this section, the generalized synthetic control method (Xu 2017). This method relaxes the parallel trend assumption, arguing that in most cases this assumption fails. It is built on the synthetic control method (Abadie and Gardeazabal 2003), that is, it reweights observations from pre-exposure periods of control units to create counterfactual controls. Yet, it generalizes this method by i) allowing staggered treatment timing and multiple treated units, ii) producing standard errors, and iii) combining it with linear fixed effects models. These generalizations improve the efficiency of the estimator (Xu 2017). We run this model to estimate our four main dependent variables. The results are shown in 3.15. We observe that the effect on $LogSalesVolume_{it}$ (p<0.05), $LogSalesRevenue_{it}$ (p<0.05), $LogNumberofTransactions_{it}$ (p<0.05), and $LogAssortmentSize_{it}$ (p<0.1) remain significant and positive.

3.7.4 Placebo Analysis

In this section, we report the results from our placebo analyses (i.e., a falsification tests) that aim to assess whether the estimated results are an artifact of the research design we employed.

Table 3.16: Randomly Assigned Treated Stores

	Log Sales Volume	Log Sales Revenue	Log Number of Transactions	Log Assortment Size	
ATT	-0.038	0.004	0.059	0.014	Randomization
	(0.02)	(0.007)	(0.034)	(0.024)	1-4
ATT	-0.009	0.024	0.012	0.013	Randomization
	(0.013)	(0.031)	(0.007)	(0.007)	5-8
ATT	0.004	-0.007	-0.002	-0.001	Randomization
	(0.002)	(0.03)	(0.063)	(0.021)	9-12
ATT	-0.134	0.006	-0.014	0.045	Randomization
	(0.088)	(0.028)	(0.008)	(0.068)	13-16
ATT	0.019	0.001	0.007	-0.012	Randomization
	(0.021)	(0.005)	(0.005)	(0.008)	17-20
ATT	0.008	-0.009	0.011	-0.013	Randomization
	(0.016)	(0.005)	(0.019)	(0.021)	21-24
ATT	0.005	0.003	-0.005	0.003	Randomization
	(0.003)	(0.006)	(0.005)	(0.004)	25-28
ATT	0.013	0.012	0.022	-0.036	Randomization
	(0.035)	(0.009)	(0.028)	(0.023)	29-32
ATT	-0.024	-0.009	0.041	-0.011	Randomization
	(0.018)	(0.007)	(0.022)	(0.021)	33-36
ATT	0.024	-0.031	0.019	0.001	Randomization
	(0.014)	(0.017)	(0.011)	(0.004)	37-40
Store Fixed Effects	YES	YES	YES	YES	
Day Fixed Effects	YES	YES	YES	YES	
Sample Size	216,580	216,580	216,580	216,580	

*p<0.1; **p<0.05; ***p<0.01. Numbers in parentheses are placebo standard errors bootstrapped for 50 times.

By following Schmidt and Raman (2022), we drop the treated stores from the data. We then randomly select four stores to be "treated" and assign them the same treatment days of the original treated stores. We run the same specification to estimate the logarithms of $LogSalesVolume_{it}$, $LogSalesRevenue_{it}$, $LogNumberofTransactions_{it}$, and $LogAssortmentSize_{it}$. We repeat this procedure for 10 times, which leads us to have 10 different sets of placebo treated stores, to ensure that the results are not an artifact of

the stores selected by the randomization. We observe that, the estimated effect is not significant in any of the 40 models we run. We provide the results in Table 3.16.

3.7.5 Standard Error Calculation

In the main analysis, we employ the placebo variance estimation algorithm of Arkhangelsky et al. (2021) to obtain the standard errors and to assess the significance of the estimated ATTs. Arkhangelsky et al. (2021) note that the placebo variance estimation algorithm relies on the two following assumptions: the error term i) follows a Gaussian distribution, and ii) is homoskedastic across units. In case of heteroskedasticity across units, results are suspected to be different; hence, in this section, we employ the bootstrap variance estimation algorithm (Arkhangelsky et al. 2021) to test whether results hold. We again set the number of replications to 50 as suggested by Mooney et al. (1993). The results are shown in Table 3.17. The estimated effects stay significant and positive for all of the main outcome variables.

Table 3.17: Store-level Results: Alternative Standard Error Calculation

	Log Sales Volume	Log Sales Revenue	Log Number of Transactions	Log Assortment Size
ATT	0.058***	0.066***	0.056***	0.06**
	(0.014)	(0.012)	(0.008)	(0.028)
Confidence Interval (95%)	[0.031, 0.086]	[0.043, 0.088]	[0.042, 0.072]	[0.004, 0.116]
Store Fixed Effects	YES	YES	YES	YES
Day Fixed Effects	YES	YES	YES	YES
Sample Size	219,128	219,128	219,128	219,128

*p<0.1; **p<0.05; ***p<0.01. Numbers in parentheses are placebo standard errors bootstrapped for 50 times.

3.8 Conclusion

In this study, we examine the potential spillover effects of offering an experiential service in the retail context. Although retailers have been transforming their brick-and-mortars to attract more customers, the research on potential effects of such services is limited. By analyzing data from a supermarket chain that introduced a taproom service in some stores, this study fills this gap. To our knowledge, we provide the first empirical evidence of spillover effects of an experiential service. Since this service format is becoming increasingly prevalent, the findings of this study have significant managerial implications for retailers.

Our results suggest that, relative to control stores, treatment stores experience an increase in the total quantity of items sold, sales revenue, number of transactions, and the variety of SKUs sold by 5.97%, 6.82%, 5.76%, and 6.18%, respectively. These estimations show a significant spillover effect of an experiential service. Our department-level investigation suggests that the store-level increases are mainly driven by the impact in a few departments. We also find that, after the treatment, the average item price increases by 1.71% yet the average basket size decreases by 1.21% in treatment stores compared to control stores. This suggests that customers visit the stores more frequently, yet they buy fewer yet more expensive items in each transaction. When we analyze the changes based on specific product characteristics, we find that impulse purchasing increases by 0.5% in treatment stores. We also find that purchases of perishable items increase by 0.4%. Furthermore, our customer-reach analysis suggests that both extensive (i.e., reaching new consumers) and intensive (i.e., overall increase in store visits) margins drive the store-level results.

These findings generate valuable insights for grocery retailers. For a retailer that wishes to increase store sales revenue and store sales volume, our results suggest that introducing a taproom service would help to achieve this goal. This outcome is achieved

via an increase in store visits, unique visitors, and dwell time. Note that these mechanisms may influence operational efficiency. Mitigations should be made by store mangers to avoid aisle and checkout congestion which could lead to a disrupted customer flow through the store, resulting in queue abandonment, customer dissatisfaction, and reputational damage. Therefore, managers may wish to consider that checkouts are staffed and self-checkouts are fully functioning.

Furthermore, our results have inventory management implications. We find an increase in sales of:

- perishable products;
- products that are susceptible to impulse buying; and
- products from the following departments: Alcohol & Beverage, Snacks, Bakery, Meat & Seafood, and Produce.

Hence, store managers may want to reconsider the inventory replenishment policies of these products. By changing the order quantities or order frequencies, they can ensure the availability of these items. On top of the availability, for perishable products, they may want to consider the freshness as well so that they can fully capitalize on the observed spillover effect.

Our paper is not free from limitations. First, we analyze a single type of an experiential service, limiting the generalizability of our results. Although there are other experiential services that retailers can offer, a taproom service is a favored choice among retailers. This is the case not only for supermarkets but also other types of shops, such as bookstores, nail salons or farmers markets (Karlamangla 2016). Because of this popularity of taproom services, it is important to deepen our understanding of their impact. Yet, we acknowledge that the empirical generalizability to the impact of other types of services might be limited. Second, in our data, we do not observe whether a customer used the taproom service during their store visit. Hence, we are not able to investigate whether baskets of

taproom users are different from baskets of non-users. By such analysis, future research could examine if the mere presence of a taproom has an impact on shopping habits.

Chapter 3 References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association* 105(490), 493–505.

Abadie, Alberto and Javier Gardeazabal (2003). The economic costs of conflict: A case study of the Basque Country. *American Economic Review* 93(1), 113–132.

Arkhangelsky, Dmitry et al. (2021). Synthetic difference-in-differences. *American Economic Review* 111(12), 4088–4118.

Babar, Yash, Ali Mahdavi Adeli, and Gordon Burtch (2023). The Effects of Online Social Identity Signals on Retailer Demand. *Management Science* 69(12), 7335-7346.

Baker, Andrew C, David F Larcker, and Charles CY Wang (2022). How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics* 144(2), 370–395.

Ballantine, Paul W, Richard Jack, and Andrew G Parsons (2010). Atmospheric cues and their effect on the hedonic retail experience. *International Journal of Retail & Distribution Management* 38(8), 641–653.

Bekkerman, Ron et al. (2023). The effect of short-term rentals on residential investment. *Marketing Science* 42(4), 819–834.

Bell, David R, Santiago Gallino, and Antonio Moreno (2018). Offline showrooms in omnichannel retail: Demand and operational benefits. *Management Science* 64(4), 1629–1651.

Ben-Michael, Eli, Avi Feller, and Jesse Rothstein (2022). Synthetic controls with staggered adoption. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84(2), 351–381.

Berman, Ron and Ayelet Israeli (2022). The value of descriptive analytics: Evidence from online retailers. *Marketing Science* 41(6), 1074–1096.

Buildd (2023). The IKEA Food Effect: How a Swedish restaurant made IKEA even more successful. https://buildd.co/marketing/ikea-food-effect.

Callaway, Brantly and Pedro HC Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics* 225(2), 200–230.

Clarke, D., Pailanir, D., Athey, S., and Imbens, G. (2023). Synthetic difference in differences estimation. *arXiv preprint arXiv:2301.11859.*

Cui, Yao, Izak Duenyas, and Ozge Sahin (2018). Unbundling of ancillary service: How does price discrimination of main service matter? *Manufacturing & Service Operations Management* 20(3), 455–466.

Dellaert, Benedict GC et al. (1998). Investigating consumers' tendency to combine multiple shopping purposes and destinations. *Journal of Marketing Research* 35(2), 177–188.

Dennis, Steve (2018). Physical Retail Isn't Dead. Boring Retail Is. https://www.forbes.com/sites/stevendennis/2018/03/19/physical-retail-is-not-dead-boring-retail-is-understanding-retails-great-bifurcation

Dixon, Vince (2017). The Rise of the Grocerant. https://www.eater.com/2017/2/27/14706474/whole-foods-restaurant-grocery-store

Dong, Yan et al. (2018). Banking on "mobile money": The implications of mobile money services on the value chain. *Manufacturing & Service Operations Management*.

Eftekhari, Saeede et al. (2023). Impact of health information exchange adoption on referral patterns. *Management Science* 69(3), 1615–1638.

Farronato, Chiara, Jessica Fong, and Andrey Fradkin (2023). Dog eat dog: Balancing network effects and differentiation in a digital platform merger. *Management Science* 70(1), 464-483.

Gallino, Santiago and Antonio Moreno (2014). Integration of online and offline channels in retail: The impact of sharing reliable inventory availability information. *Management*

Science 60(6), 1434–1451.

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

— (2019). New functions of physical stores in the age of omnichannel retailing. *Operations in an Omnichannel World*, pp. 35–50.

Gurfein, Laura (2015). Athleta Now Has Free Fitness Classes Under Its Flatiron Store. https://ny.racked.com/2015/11/19/9760574/athleta-fitness-studio-free-flatiron-nyc

Ham, Sunny et al. (2021). The rise of the grocerant: Patrons' in-store dining experiences and consumption behaviors at grocery retail stores. *Journal of Retailing and Consumer Services* 62, p. 102614.

Harnish, Richard J, Nicole C Ryerson, and Piotr Tarka (2023). Purchasing under the influence of alcohol: The impact of hazardous and harmful patterns of alcohol consumption, impulsivity, and compulsive buying. *Psychological Reports*, p. 00332941231164348.

Jiang, Baojun, Chakravarthi Narasimhan, and Ozge Turut (2017). Anticipated regret and product innovation. *Management Science* 63(12), 4308–4323

Kadet, Anne (2018). Psst...Here's a Secret to Keeping That New Year's Workout Resolution. https://www.wsj.com/articles/psst-heres-a-secret-to-keeping-that-new-years-workout-resolution-1514905200

Karlamangla, Soumya (2016). Beer at your bookstore or nail salon? Alcohol at unexpected businesses could draw customers; but also health concerns. Newspaper Article. https://www.latimes.com/local/california/la-me-ln-alcohol-outlets-20161214-story.html

Kim, Youn Kyung (2001). Experiential retailing: an interdisciplinary approach to success in domestic and international retailing. *Journal of Retailing and Consumer Services* 8(5), 287–289.

Lim, Stanley Frederick WT, Fei Gao, and Tom Fangyun Tan (2023). Channel Changes

Choice: An Empirical Study About Omnichannel Demand Sensitivity to Fulfillment Lead Time. *Management Science*.

Liu, Hongju, Pradeep K Chintagunta, and Ting Zhu (2010). Complementarities and the demand for home broadband internet services. *Management Science* 29(4), 701–720.

Marianov, Vladimir, Horst A Eiselt, and Armin L uer-Villagra (2018). Effects of multipurpose shopping trips on retail store location in a duopoly. *European Journal of Operational Research* 269(2), 782–792.

Maynard, Micheline (2017). From Sofas To Meatballs: Ikea May Be The Next Player In The Restaurant Business. https://www.forbes.com/sites/michelinemaynard/2017/04/17/from-sofas-to-meatballs-ikea-may-be-the-next-player-in-the-restaurant-business/?sh=f5892e31287f

McGrath, Maggie (2016). Why 'Grocerants' Are The Future Of Food Shopping. https://www.forbes.com/sites/maggiemcgrath/2016/06/16/why-grocerants-are-the-future-of-food-shopping

Milgrom, Paul and John Roberts (1990). The economics of modern manufacturing: Technology, strategy, and organization. *The American Economic Review*, 511–528.

Mooney, Christopher Z, Robert D Duval, and Robert Duvall (1993). Bootstrapping: A nonparametric approach to statistical inference. 95. Sage.

Moore, Logan (2021). Winn-Dixie's 'sip and shop' concept taking off in Jacksonville. https://www.bizjournals.com/jacksonville/news/2021/09/08/winn-dixie-was-pivoting-to-the-sip-and-shop-con.html

Mutlu, Nevin, Hadi El-Amine, and Ozge Sahin (2023). Offering Memories to Sell Goods? Pricing and Welfare Implications of Experiential Retail. *Manufacturing & Service Operations Management*.

Narasimhan, Chakravarthi, Scott A Neslin, and Subrata K Sen (1996). Promotional elasticities and category characteristics. *Journal of Marketing* 60(2), 17–30.

Olivares, Marcelo and Gerard P Cachon (2009). Competing retailers and inventory: An

empirical investigation of General Motors' dealerships in isolated US markets. *Management Science* 55(9), 1586–1604.

Pendrill, Katherine (2016). 22 Examples of Diversified Retail. https://www.trendhunter.com/slideshow/diversified-retail

Pine, B Joseph and James H Gilmore (1998). Welcome to the experience economy. Vol. 76. Harvard Business Review Press Cambridge, MA, USA.

Pollard, Amelia, Jeremy Hill, and Giulia Morpurgo (2023). Retail Apocalypse Returns. https://www.bloomberg.com/news/newsletters/2023-02-07/retail-distress-hits-companies-like-bed-bath-beyond-bbby-party-city-prty

Porreca, Zachary (2022). Synthetic difference-in-differences estimation with staggered treatment timing. *Economics Letters* 220, 110874.

Reuter, Dominick (2023). Nearly 2,400 stores are closing across the US in 2023. Here's the full list. https://www.businessinsider.com/stores-closing-in-2023-list

Ruback, Brianna (2023). 5 Grocery Chains With In-Store Bars—So You Can Drink While Shopping. https://www.eatthis.com/grocery-chains-with-in-store-bar

Schmidt, William and Ananth Raman (2022). Operational disruptions, firm risk, and control systems. *Manufacturing & Service Operations Management* 24(1), 411–429.

Schmitt, Bernd (1999). Experiential marketing. *Journal of Marketing Management* 15(1-3), 53–67

Selvam, Ashok (2015). Confirmed: America's First Boozy Target Is Bringing Wine Bar, Small Plates to Chicago. https://chicago.eater.com/2015/9/4/9261323/target-wine-bar-chicago-americas-first

Shin, Minkyu et al. (2023). The impact of the gig economy on product quality through the labor market: Evidence from ridesharing and restaurant quality. *Management Science* 69(5), 2620–2638.

Simonson, Itamar (1992). The influence of anticipating regret and responsibility on purchase decisions. *Journal of Consumer Research* 19(1), 105–118.

Singh, Shivendu Pratap, Chris F Kemerer, and Narayan Ramasubbu (2021). Innovation in complex assembled electronic products: An analysis of the evolution of television components. *Journal of Operations Management* 67(6), 680–703.

Sowder, Amy (2023). How to snag digital impulse buys for produce. Newspaper Article. https://www.thepacker.com/news/retail/how-snag-digital-impulse-buys-produce

Stahl, Konrad (1982). Differentiated products, consumer search, and locational oligopoly. *The Journal of Industrial Economics*, 97–113.

Statista (2017). Sales share of Whole Foods Market worldwide from 2010 to 2017, by product category. https://www.statista.com/statistics/258680/sales-share-of-whole-foods-market-worldwide-by-product-category/

Trivedi, Minakshi, Dinesh K Gauri, and Yu Ma (2017). Measuring the efficiency of category-level sales response to promotions. *Management Science* 63(10), 3473–3488

VanderWeele, Tyler J (2011). Sensitivity analysis for contagion effects in social networks. *Sociological Methods & Research* 40(2), 240–255.

Wang, Ruxian, Maqbool Dada, and Ozge Sahin (2019). Pricing ancillary service subscriptions. *Management Science* 65(10), 4712–4732.

WeatherAds (2023). 5 Scarily Effective Weather-Triggered Ad Campaigns. https://www.weatherads.io/blog/5-scarily-effective-weather-triggered-ad-campaigns

Wolinsky, Asher (1983). Retail trade concentration due to consumers' imperfect information. *The Bell Journal of Economics*, 275–282.

Wooldridge, Jeffrey M (2021). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. In: Available at SSRN 3906345.

Xu, Yiqing (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis* 25(1), 57–76.

— (2023). "Causal inference with time-series cross-sectional data: a reflection". Available at SSRN 3979613.

Zeelenberg, Marcel (1999). Anticipated regret, expected feedback and behavioral deci-

sion making. *Journal of Behavioral Decision Making* 12(2), 93–106.

Zeelenberg, M., Beattie, J., Van der Pligt, J., & De Vries, N. K. (1996). Consequences of regret aversion: Effects of expected feedback on risky decision making. *Organizational behavior and human decision processes*, 65(2), 148-158.

Conclusion

In conclusion, this dissertation investigates two very timely and important topics in the grocery retail sector: digitalization and retail store transformation.

Regarding digitalization, this dissertation suggests that behavioral biases play a key role in deviations from algorithmic suggestions. Although it is important to incorporate private human knowledge into inventory decisions, we see that censorship bias, anchoring bias, supply line underweighting, and newsvendor double counting trigger discretionary behavior in terms of ordering more or less than the algorithmic suggestions. The findings contribute to the behavioral operations literature by unearthing a novel bias, testing the prevalence of previously suggested biases in the field, testing the performance implications of decisions triggered by these biases, and moving the conversation on discretionary behavior from logistical reasons to behavioral reasons. The findings are relevant to retailers since they can be used in the development of algorithmic inventory decision-making systems or in the training of employees.

In terms of retail transformation, this dissertation suggests that offering an experiential service positively and significantly affects store-level performance measures. Yet, for a grocery retailer who aims to increase basket size, we find that introducing a taproom may not be beneficial. In addition, the findings suggest that the impact of this taproom service is not significant in all store departments, explained by the complementarity sales mechanism of experiential services. Taken together, the findings of this dissertation generate practical insights to retailers who may introduce an experiential service.

Conclusión

En conclusión, esta disertación investiga dos temas muy actuales e importantes en el sector minorista de comestibles: la digitalización y la transformación de las tiendas minoristas.

En cuanto a la digitalización, esta tesis sugiere que los sesgos de comportamiento desempeñan un papel clave en las desviaciones de las sugerencias algorítmicas. Si bien es importante incorporar el conocimiento humano privado en las decisiones de inventario, vemos que el sesgo de censura, el sesgo de anclaje, la infraponderación de la línea de suministro y el doble conteo de los vendedores de periódicos desencadenan un comportamiento discrecional en términos de realizar pedidos más o menos que las sugerencias algorítmicas. Los hallazgos contribuyen a la literatura sobre operaciones conductuales al descubrir un sesgo novedoso, probar la prevalencia de sesgos previamente sugeridos en el campo, probar las implicaciones de desempeño de las decisiones desencadenadas por estos sesgos y trasladar la conversación sobre el comportamiento discrecional de razones logísticas a razones conductuales. Los hallazgos son relevantes para los minoristas ya que pueden usarse en el desarrollo de sistemas algorítmicos de toma de decisiones de inventario o en la capacitación de empleados.

En términos de transformación minorista, esta tesis sugiere que ofrecer un servicio experiencial afecta positiva y significativamente las medidas de desempeño a nivel de tienda. Sin embargo, para un minorista de comestibles que pretende aumentar el tamaño de su cesta, encontramos que introducir una taberna puede no ser beneficioso. Además,

los hallazgos sugieren que el impacto de este servicio de taberna no es significativo en todos los departamentos de la tienda, explicado por el mecanismo de venta complementaria de los servicios experienciales. En conjunto, los hallazgos de esta tesis generan conocimientos prácticos para los minoristas que pueden introducir un servicio experiencial.