ESSAYS ON RETAIL OPERATIONS MANAGEMENT

Hyun Seok Lee

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Kenan-Flagler Business School (Operations).

Chapel Hill
2017

Approved by:

Vinayak Deshpande

Seyed Emadi

Saravanan Kesavan

Camelia Kuhnen

Bradley Staats

ABSTRACT

Hyun Seok Lee: Essays on Retail Operations Management
(Under the direction of Saravanan Kesavan)

Under the competitive nature, retailers need to consider numerous aspects to make better operational decisions. Retailers should understand customers' needs to provide better service; have the right amount of labor to match supply with demand; and anticipate investors' responses on the announcement of revealing private information such as building excess inventory to maximize retailers' objective. This dissertation empirically examines these three aspects – customer behavior, labor scheduling, and excess inventory announcement where there appears to be limited empirical research in the literature. I utilize individual retailer's micro-level proprietary data across multiple stores as well as publicly available firm-level financial data for multiple retailers.

In the first essay, I identify a new phenomenon called thwarting behavior, defined as a systematic change in customers' behavior when they experience congestion that imposes negative externalities on other customers. I provide empirical evidence for the thwarting behavior by analyzing archival data obtained from a retailer and by conducting a field study where I observe customer directly. I then quantify its impact on sales drop by running a field experiment.

In the second essay, I examine the impact of incentives for store managers on their labor scheduling decisions. I find empirical evidence that incentive improves outcomes of store managers’ labor scheduling decisions and this finding is mainly driven by effort effect rather than selection mechanism. I also find that the financial incentives have differential impact on underlying decisions that led to the labor scheduling outcomes such as forecasting, labor planning, and execution.

In the third essay, I analyze the determinants of excess inventory announcement and the stock market reaction to the announcement in the U.S. retail sector. I find that operationally competent retailers, measured by total factor productivity, have a lower probability of announcing excess inventory in the following year. In addition, the stock market penalizes excess inventory announcements made by operationally competent retailers more severely than those made by their less competent peers. Finally, providing action information, which the firm has taken or plans to take to deal with excess inventory, moderates the negative association between firm’s operational competence and abnormal returns due to the announcement.

To my wife, daughter, son, parents, sister and friends, I couldn’t have done this without you.
Thank you for all of your support along the way.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the constant inspiration, support and encouragement from the faculty members at University of North Carolina at Chapel Hill, fellow colleague, close friends and members of my family.

I would like to begin by thanking my advisor, Prof. Saravanan Kesavan for the continuous support of my PhD study, for his patience, motivation, and immense knowledge. Prof. Kesavan has been an enduring source of inspiration all through my doctoral program. His guidance has been instrumental in helping me develop the ability to identify and critically analyze the different facets to each research problem. As an advisor, a teacher, and a mentor, he has shown me, by example, the perseverance that is required to be a good researcher. Through various interactions and project engagements, I have learnt to appreciate that there is as much value in the journey itself, as in the ultimate goal that is to be reached at the end of each project. Under the guidance of Prof. Kesavan, I have had the advantage of not only learning the intricate and advanced tools and techniques required in the field of empirical research, but also to conduct research that would bring theoretical insights to practical applications. He has instilled in me a strong sense of the rigor and discipline that is required from academicians. I could not have imagined having a better advisor and mentor for my PhD study.

I would also like to express my sincere gratitude to our department chair, my co-author, and member of my dissertation committee, Prof. Vinayak Deshpande, for providing me extensive professional guidance throughout the dissertation process. His encouragement for the

PhD program and support for various conferences have given me the advantage to constantly keep in touch with recent research advancements, and interact with other academicians and practitioners in the field of retail operations. I would also like to thank my dissertation committee members, Prof. Camelia Kuhnen, Prof. Bradley Staats, and Prof. Seyed Emadi, for their advice, suggestions, and support throughout my dissertation. I have had the opportunity of working on a research project with Prof. Kuhnen and she has offered kind encouragements throughout my study. I have worked with Prof. Staats and Prof. Emadi as a teaching assistant and they have showed me the best practices in the classroom through their enthusiastic lectures and engagement with students, helping me teach well in my own class.

Besides my committee members, I would like to thank the rest of faculty members: Prof. Wendell Gilland, Prof. Bin Hu, Prof. Lauren Lu, Prof. Ann Marucheck, Prof. Adam Mersereau, Prof. Ali Parlakturk, Prof. Sandeep Rath, Prof. Nur Sunar, Prof. Jay Swaminathan, and Prof. Harvey Wagner, for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

I would also like to extend my gratitude to my senior colleagues – Aaron Ratcliffe, Karthik Natarajan, Adem Ordemir, Gang Wang, and Zhe Wang who were always available to help me navigate through the different stages of the program. I would also like to wish the upcoming PhD students – Ying Zhang, Pradeep Pendem, Nagarajan Sethuraman, Brett Hathaway, Oguz Cetin, Baiyang Huang, Anand Bhatia and Neha Jha – the very best in successfully completing the PhD program and very much look forward to hearing about their research topics. I would like to extend my thanks to my close Korean PhD friends Sunjin Park, Edward Sul, Hyeong-Tak Lee, Sunwoo Hwang, Prof. Chanil Boo, Prof. Changhyun Kim who were unwavering in their support and were always present to bolster my confidence during tough times. I am also indebted to the firm from whom we obtained the data for the empirical analysis in my dissertation. Last but not the least, I would like to say a word of appreciation to the efforts put in by Sharon Parks, Kim Yates, Erin Rimmer, and Holly Guthrie in making sure that there was always adequate administrative support during my time here.

Meeting the unique challenges of a doctoral program would not have been possible without the steadfast love and guidance of my parents, Changgab Lee and Hyungsook Kim. Their unequivocal belief in my potential to succeed in the doctoral program, and confidence in my capabilities has helped me remain resolute in the pursuit of my dreams. Through their own hard work and desire to learn continuously they have been my role models throughout my professional career. My appreciation also goes to my only sister, Saerom Lee, for her unconditional support in me. I should also thank to my parents-in-laws, Young Jai Chai and Jeonghee Jo, for their endless support and love throughout my life in completely different territory in the opposite side of the world. Finally, I would like to extend my deepest gratitude to Prof. Kyung Sam Park, for encouraging me to directly start academic career instead of industry career after college graduation. Thanks to him, I was able to learn more about academic life as a professor while preparing PhD in the master’s program and finally will start it soon.

Most importantly, everything has been possible with the love and endless support from my wife, Aeri Chae. She has been listened my dissertation more than any other people in the world and provided feedback from lay-back audience’s perspective. When I depressed, she was the one who made me smile and encouraged to continue my journey. She deserves more than half of my doctoral degree. My lovely daughter and son, Lindsay Jia Lee and Jay Hyun Lee, have been providing the true happiness to me and they have been the best motivation for me to endure the tough journey of PhD.

TABLE OF CONTENTS

LIST OF TABLES xiv

LIST OF FIGURES xvi

1 Introduction 1

1.1 Dissertation Overview 2

1.1.1 Understanding and managing customer-induced negative externalities in congested self-service environments 2

1.1.2 Can incentives improve labor scheduling decisions? Evidence from a quasi-experiment 3

1.1.3 Determinants of excess inventory announcement and stock market reaction in the retail sector 3

2 Understanding and Managing Customer-Induced Negative Externalities in Congested Self-Service Environments 5

2.1 Introduction 5

2.2 Prior Literature 10

2.3 Theory Development and Hypotheses 14

2.4 Data and Methodology for Archival Data Analysis 18

2.4.1 Data sources 18

2.4.2 Variables 20

2.4.2.1 Dependent variables 20

2.4.2.2 Key variables of interest 21

2.4.2.3 Control variables 21

2.4.3 Test of inverted-U relationship....................23

2.5 Results: The Negative Impact of Congestion....................25

2.5.1 Robustness check: Alternative ways to test for inverted U-shaped relationship....................28

2.5.2 Alternative explanations....................32

2.6 Field Study: Evidence of Thwarting Behavior....................33

2.6.1 Theory: Ideal experiment....................33

2.6.1.1 Handling selection biases....................34

2.6.2 Field study at retailer B....................36

2.6.2.1 Data collection....................37

2.6.3 Results....................38

2.6.3.1 Thwarting behavior 1: Bringing more clothes into fitting rooms increases waiting time for other customers....................38

2.6.3.2 Thwarting behavior 2: Leaving behind more clothes in fitting rooms increases chance of lost sales due to phantom stockouts....................40

2.7 Field Experiment: Mitigating the Negative Impact of Phantom Stockouts....................43

2.7.1 Experimental design....................44

2.7.2 Implementation and methodology....................45

2.7.3 Results....................46

2.8 Managerial Implications and Conclusion....................47

3 Can Incentives Improve Labor Scheduling Decisions?

Evidence from a Quasi-Experiment....................51

3.1 Introduction....................51

3.2 Literature & Hypotheses....................56

3.2.1 Related literature....................56

3.2.2 Hypotheses 58

3.3 Data & Methodology 64

3.3.1 Data sources....................64

3.3.2 Labor scheduling process 65

3.3.3 Incentive scheme 67

3.3.4 Variables....................70

3.3.5 Empirical methodology 72

3.4 Results 74

3.4.1 Can incentive improve outcomes of overall labor scheduling decisions?....................74

3.4.1.1 Model free evidence....................74

3.4.1.2 Main results....................77

3.4.1.3 Effort vs. selection 79

3.4.1.4 Robustness checks 80

3.4.2 Mechanisms: Underlying decisions for scheduling labor 84

3.4.2.1 Forecasting....................84

3.4.2.2 Labor planning....................88

3.4.2.3 Execution 89

3.4.2.4 Robustness check....................93

3.5 Conclusion 93

4. Determinants of Excess Inventory Announcement and Stock Market Reaction in the Retail Sector 97

4.1 Introduction....................97

4.2 Literature Review....................103

4.3 Hypothesis Development....................106

4.3.1 Determinants of excess inventory announcement 106

4.3.2 Market response on excess inventory announcement 107

4.4 Determinants of Excess Inventory Announcement..109

4.4.1 Estimation of TFP...109

4.4.2 Data and variables 114

4.4.2.1 Firm level TFP 114

4.4.2.2 Excess inventory announcement...115

4.4.2.3 Daily stock price 118

4.4.2.4 Variables 119

4.4.3 Model specification and analysis..121

4.4.4 Results 122

4.4.4.1 Results: Determinants of excess inventory announcement................................122

4.4.4.2 Robustness checks 124

4.5 Market Reaction...127

4.5.1 Event study methodology..127

4.5.2 Model specification 130

4.5.3 Results 130

4.5.3.1 Results: Main effect of excess inventory announcement on stock price 130

4.5.3.2 Results: Market reaction across the announcing firm's operational competence..132

4.5.3.3 Robustness checks 134

4.5.4 Value of information in excess inventory announcements..135

4.6 Summary and Conclusion...140

Appendix I: Validating the inverted U-shaped relationship at retailer B 143

Appendix II: Alternative view of the incentive change 148

Appendix III: Heterogeneity of firm level and industry level TFP in the U.S. retail sector 150

BIBLIOGRAPHY 154

LIST OF TABLES

2.1 Summary of Statistics of the Variables (Retailer A) 22

2.2 Pearson Correlation Coefficients (Retailer A) 22

2.3 Inverted-U Relationship (Retailer A) 28

2.4 Robustness Checks for Inverted-U Relationship (Retailer A) 30

2.5 Robustness Checks without Quadratic Functional Form Assumption (Retailer A) 31

2.6 Impact of Congestion on the Number of Items and
Time Spent in the Fitting Room (Retailer B) 39

2.7 Phantom Stockouts (Retailer B) 41

2.8 Experiment Design (Retailer B) 45

2.9 Impact on Sales of Recovery Operations by Fitting Room Associate (Retailer B) 47

3.1 Summary Statistics............71

3.2 Incentive Affects Overall Labor Scheduling Decisions............78

3.3 Selective Robustness Checks 81

3.4 Mechanism: (1) Forecasting (No impact on accuracy)............85

3.5 Mechanism: (1) Forecasting (More effort) 86

3.6 Mechanism: (1) Forecasting (Optimistic bias) 87

3.7 Mechanism: (2) Labor planning 88

3.8 Mechanism: (3) Execution (Last mile adjustment effort) 92

4.1-a Distibution of Excess Inventory Announcements by Industry 118

4.1-b Reasons of Excess Inventory Announcement 118

4.1-c Follow-up Actions to Excess Inventory Announcement............118

4.2-a Summary Statistics 120

4.2-b Pearson Correlation 121

4.3 Regression Result for Productivity and Excess
Inventory Announcement (Probit Model)....123

4.4 (Robustness Check) Regression Result for Productivity
and Excess Inventory Announcement 125

4.5 Cumulative Average Abnormal Returns across Different Event Windows 131

4.6 Regression Result for Productivity and Cumulative Abnormal Return....133

4.7 (Robustness Check) Regression Result for Productivity
and Cumulative Abnormal Return....135

4.8 Regression Result for Moderator of Productivity and Cumulative Abnormal Return 138

5.1.1 Inverted-U Relationship with Hourly Dummies (Retailer A)....143

5.1.2-a Inverted-U Relationship (Retailer B)....145

5.1.2-b Robustness Checks Using Spline (Retailer B) 145

5.1.3 Phantom Stockouts (Details on day 1, Sunday)....146

5.1.4 Phantom Stockouts with Price, Size, and Category (Details on day 2, Saturday)....147

5.2.1 Overall Store Performance is Smaller in the Post-Incentive Change Period....149

5.3.1 90-10 TFP Ratio across Industries (SIC2)....152

5.3.2-a Autoregressive Coefficients....153

5.3.2-b Autoregressive Coefficients across Industries (LP) 153

5.3.2-c Autoregressive Coefficients across Industries (OP) 153

LIST OF FIGURES

2.1 Conceptual Model......14

2.2 Fitting Room Areas......19

2.3 Distinguishing Incoming from Outgoing Traffic......20

3.1 Labor Scheduling Process......67

3.2 Histogram of Labor Scheduling Decisions......75

3.3 Accuracy in Labor Scheduling Decisions over Time......76

3.4 Accuracy in Labor Scheduling Decisions over Time (Alternative Method)......76

3.5 Average of Accuracy in Labor Scheduling Decisions......77

4.1 Distribution of Excess Inventory Announcements by Year......117

5.3.1 Mean Retail Productivity across years (1963-2011)......150

5.3.2 Heterogeneity of Retail Productivity across Industries (SIC2)......151

5.3.3 Heterogeneity of Retail Productivity across Years (1963-2011)......151

CHAPTER 1
Introduction

Facing enormous competitive pressures from online retailers, brick-and-mortar retailers are experiencing tough times. According to the recent report[1], nearly 80 percent of Americans do at least some shopping on the internet, with 43 percent shopping online on a regular basis such as "weekly" or "a few times a month." It is a significant increase as just 22 percent of Americans had made a purchase online in 2000. In addition, customers are capable of comparing price and products within the reach of their pocket in the new digital era stemming from big data, mobile commerce, and the explosion of omni-channel retailing. This makes retailers' margin thinner. Such intense competition and declining margins have forced many retailers to critically examine and redesign their operations in an effort to improve their performance.

To this end, retailers need to consider numerous aspects into account when they make operational decisions. They should understand customers' need to provide better service so that retain their existing customers and earning a bigger share of the customer's wallet. They want to have the right amount of labor to match supply with demand as overstaffing can increase store expenses and understaffing can result in lost sales and poor service. In the higher strategic level, retailers need to anticipate investors' responses on their decisions of revealing private information such as building excess inventory to make better decisions. This dissertation empirically examines these three aspects – customer behavior, labor scheduling, and excess

[1] See "Online Shopping and E-Commerce" by Aaron Smith and Monica Anderson, Pew Research Center, December 19,2016

inventory announcement – aiming to provide valuable insights to retailers helping them in better decision making.

In this empirical study, we utilize individual retailer's micro-level proprietary data across multiple stores as well as publicly available firm-level financial data for multiple retailers. This dissertation research consists of three chapters. In what follows, I briefly describe each chapter in sequence.

1.1 Dissertation Overview

1.1.1 Understanding and managing customer-induced negative externalities in congested self-service environments

In this essay I identify a new phenomenon called thwarting behavior, defined as a systematic change in customers' behavior when they experience congestion that imposes negative externalities on other customers. Using point-of-sale (POS), traffic, and labor data obtained from a retail technology platform firm, RetailNext, and one of its clients, I demonstrate an inverted U-shaped relationship between fitting room traffic and sales. This is consistent with thwarting behavior and shows that managing congestion in fitting rooms is critical for store performance. To provide the direct evidence for thwarting behavior, I conduct a field study observing customer behavior at another retailer. This field study provides evidence that customers indeed change their behavior during congestion that lead to increased waiting time for others and phantom stockouts. Finally, I run a field experiment to show that these phantom stockouts cause significant lost sales in retail stores. I quantify that mitigating phantom stockouts in the fitting room area using an associate increases checkout-counter-level hourly sales by 22.6%.

1.1.2 Can incentives improve labor scheduling decisions? Evidence from a quasi-experiment

There are conflicting views on the role of incentives in improving performance. Theoretical literature on agency theory, expectancy theory and goal-setting theory argue that incentives make agents exert more effort that leads to better performance whereas psychologists and behaviorists contend that extrinsic rewards have a negative effect on intrinsic motivation. However, there is limited research in operations management literature that has examined the role of pay-for-performance incentives on improving operational outcomes with field data. Here, I use detailed weekly scheduling data from a quasi-experimental setting at the retail chain to examine the impact of financial incentives on operational decisions, specifically labor scheduling decisions. I find that store managers make better labor scheduling decisions under the strong incentive scheme, indicating that incentives for store managers help them improve labor scheduling decisions. I further find that improvement in labor scheduling decisions due to incentive is mainly driven by change in effort rather than selection mechanism. Finally, I find that the financial incentive has differential impact on three underlying decisions (i.e., forecasting, labor planning, and execution) that lead outcomes of overall labor scheduling decisions.

1.1.3 Determinants of excess inventory announcement and stock market reaction in the retail sector

In this essay, I empirically analyze the determinants of excess inventory announcement and the stock market reaction to the announcement in the U.S. retail sector. I examine if the firm’s operational competence, as measured by total factor productivity (TFP), can explain the retailer’s excess inventory announcement. I also investigate if the stock market reaction to such announcements is conditional on the operational competence of the announcing firm. I use a combined dataset on excess inventory announcements, annual financial statements, and daily stock prices of publicly traded retailers in the U.S. between 1990 and 2011. I find that operationally competent retailers have a lower probability of announcing excess inventory in the following year. In addition, the stock market penalizes excess inventory announcements made by operationally competent retailers more severely than those made by their less competent peers. Finally, providing action information, which the firm has taken or plans to take to deal with the excess inventory, moderates the negative association between firm's operational competence and abnormal returns due to the announcement whereas I do not find such moderating effect with reason information.

CHAPTER 2
Understanding and Managing Customer-Induced Negative Externalities in Congested Self-Service Environments

2.1 Introduction

Congestion management is an important endeavor in service settings. Since congestion occurs when there is variability in inter-arrival times or service times, researchers have identified several ways to reduce or accommodate this variability to relieve congestion (Hassin and Haviv 2003; Lu et al. 2013). In these settings, customers are typically assumed to be either passive – simply waiting for service and receiving it – or strategic, but only to the extent of deciding to balk or renege from queues (Kulkarni 2009; Aksin et al. 2013; Dong et al. 2015; Batt and Terwiesch 2015; 2016). However, customers could show other types of strategic behavior during congestion different from balking and reneging. More importantly, some of those strategic behaviors may exacerbate congestion by imposing negative externalities on other customers, unlike balking and reneging that have a positive externality on other customers. We call such strategic behavior as *thwarting* behavior and define it as a systematic change in customers' behavior when they experience congestion in a way that imposes negative externalities on other customers.

Anecdotal evidence of such thwarting behavior can be found in popular press and literature. Transportation researchers find that drivers tend to exhibit aggressive behaviors such as honking at other cars and tailgating during congestion, which leads to accidents and exacerbates congestion (Hennessy and Wiesenthal 1997). In call centers, irritated customers due to long wait have been known to demand longer service (Dong et al. 2015) that could increase waiting time for others. It is unclear whether retail customers exhibit any thwarting behavior when stores become congested.

In this paper, we examine the phenomenon of thwarting behavior in fitting rooms of apparel retailers. We choose fitting rooms as they are usually self-service environments in most retail stores with little or no monitoring where we may be able to observe this behavior. During congestion, two thwarting behaviors may manifest in fitting rooms. Customers might take more clothes to try on in a fitting room when they experience congestion because they do not want to make multiple visits and are afraid of losing items. This requires them to occupy fitting rooms longer (service slowdown), increasing waiting time for other customers who want to use fitting rooms. In addition, they could leave behind clothes in fitting rooms if they do not purchase them or replace those items back in the shelf or a recovery rack. Such misplaced items could result in phantom stockouts when they are the last units of inventory in the store. Lost sales could increase due to these two types of thwarting behavior as it would lead to increased balking/reneging due to increased waiting time for other customers and more phantom stockouts (DeHoratius and Ton 2015) due to misplaced inventory.

We first analyze archival data on fitting room traffic and sales to detect a decline in sales expected with thwarting behavior. Although we expect sales to decline due to thwarting behavior in fitting rooms, it is possible that the sales drop might be driven by other factors such as crowded checkout counters, phantom stockouts in the rest of the store, or overall poor service level due to understaffing. Since it is impossible to rule out all alternative explanations without detailed data on every part of the store, we further conduct a field study and a field experiment to provide direct evidence for the thwarting behavior and to quantify its impact on store sales.

In the field study we directly observe the behaviors of customers entering the fitting room area at a major retailer. We measure the extent to which customers change their behavior during congestion by taking more clothes into the fitting room and leaving behind more unwanted clothes in the fitting room. We also measure the increase in waiting time to other customers and the phantom stockouts that occur due to thwarting behavior. While our field study can detect whether customers change their behavior when they face congestion, it does not directly reveal the impact of thwarting behavior on sales decline. So, we subsequently perform a field experiment to show its impact on sales.

Our data were obtained through collaborations with three companies. RetailNext is a leading in-store analytics provider to retailers. It collects traffic information from video cameras in retail stores to codify customer-arrival patterns as well as customer pathways in retail stores. In addition, they collate customer traffic information and point-of-sale (POS) data. These data were obtained from one of its clients, a large U.S.-based retailer (retailer A). This retailer's stores are about 50,000 sq. ft. in size and primarily carry men's, women's, and children's apparel, along with some home furnishing goods. The data obtained from RetailNext and retailer A helped us provide initial evidence for the presence of thwarting behavior and its impact on store sales. We further observe customers directly through a field study and conduct a field experiment at another department store retail chain (retailer B).

Our primary findings are as follows. First, we observe an inverted U-shaped relationship between fitting room traffic and sales. Sales initially increase with fitting room traffic as more customers intend to purchase. Beyond a certain point, however, we observe a decline in sales when fitting rooms become congested. Contrary to the conventional wisdom that more traffic drives more store sales, we identify that too much traffic in fitting rooms can hurt store sales.

This observation shows that managing congestion in fitting rooms is critical for brick-and-mortar apparel retailers.

Second, we identify two mechanisms that drive the inverted-U relationship observed in our data. By observing 209 customers who used fitting rooms for 24 non-continuous hours spanning 3 weekends, we find that customers, on average, take additional 1.38 (23.96%) clothes into the fitting room when they experience congestion, which requires them to occupy fitting rooms an average of 1.87 (18.61%) minutes longer. This behavior imposes a negative externality on other customers as overall waiting time increases. Our investigation of misplaced inventory shows that customers, on average, leave behind 2.69 (54%) items more in the fitting rooms during congested periods. So, customers take a few more extra items into the fitting rooms and leave behind many more unwanted items when they experience congestion, suggesting that the sales from fitting room users will be low during congested periods. In addition, further decline in sales occurs as the inventory left behind in the fitting rooms could result in phantom stockouts. We observe a 38.6% of phantom stockout rate among items left behind in the fitting room area. Further analysis shows that those misplaced items were high-price merchandise with significant opportunity costs. These two mechanisms are consistent with thwarting behavior that we define in the paper.

Finally, our field experiment shows that mitigating phantom stockouts in the fitting room area alone can significantly improve store sales. By comparing the treatment to control groups, we find that an addition of a backend recovery operation for items left behind in the fitting rooms using an associate increases checkout-counter-level hourly sales by $186.93 (22.6%). The large increase in sales from reducing phantom stockouts shows that thwarting behavior can significantly hurt store sales during congested periods.

Our research makes several contributions to the operations management literature. Our first contribution is the identification and demonstration of customers' thwarting behavior. Researchers have identified that customers exhibit strategic behaviors, e.g., balking or reneging, in the queue based on their service experience, which affects their attitude such as anxiety and fairness (e.g., Rafaeli et al. 2002; Bitran et al. 2008). While balking and reneging lead to positive externalities for other customers, none of the papers have empirically identified this specific strategic behavior, thwarting, which imposes a negative externality on others. We provide the first direct evidence for the presence of thwarting behavior in retail stores. Identification of thwarting behavior enables us to show a new dimension of customer impact in service operations as it reveals that *customer*-induced service slowdown, in contrast to server-driven slowdowns (Kc and Terwiesch 2009; Anand et al. 2011; Tan and Netessine 2014; Batt and Terwiesch 2016), can be an important reason for service slowdowns in retail settings.

Second, we measure instore-based phantom stockouts in retail stores, which is different from backroom-based phantom stockouts identified in prior literature (DeHoratius and Raman 2008; Ton and Raman 2010). While phantom stockouts are expected to result in lost sales, the magnitude of the lost sales has not been estimated before. By using a field experiment we show that mitigating phantom stockouts in the fitting room area during congested periods can lead to significant increase in store sales.

Third, we contribute to the expanding literature on retail labor (Fisher et al. 2007; Ton 2009; Netessine et al. 2010; Kesavan et al. 2014; Mani et al. 2015) by conducting the first field experiment in this line of research. While prior literature has argued for increasing store labor to drive sales, we use a field experiment to show that specifically using labor to perform a backend recovery operation in fitting rooms can increase sales significantly. In addition, field studies present important advantages over lab experiments to examine thwarting behavior of individual customers as the treatment effects are expected to interact with participant characteristics (Al-Ubaydli and List 2015). Since thwarting behavior causes negative externality on other customers, it is likely that this behavior is harder to study in a lab setting. Participants in the lab may either change behavior when they are aware of being observed or may result in systematically opting out, causing biases in our inferences. Due to the covert nature of the field, we can observe customers in their natural environment and understand the degree to which customers engage in thwarting behavior.

2.2 Prior Literature

Researchers have identified several ways to relieve congestion. In these settings, customers are typically assumed to be either passive or strategic, but only to the extent of deciding to balk or renege from queues, which have a positive externality for other customers. Examples of analytical works with passive arrivals are Deo and Gurvich (2011) and Lee et al. (2012) while those of empirical works include Chan et al. (2014) and Batt and Terwiesch (2016). Models with balking and reneging can be found in many books, e.g., Kulkarni (2009). Recent empirical works on this topic include Aksin et al. (2013) and Batt and Terwiesch (2015). However, customers show other types of strategic behavior in retail settings different from balking and reneging, which impose negative externalities on other customers. For example, Allon and Hanany (2012) analytically studied customer's behavior of cutting the line that can increase waiting time for other customers. To the best of our knowledge, our paper is the first paper that empirically identifies and demonstrates customers' behavior change, when they experience congestion, in a way that imposes a negative externality on other customers; we call it thwarting. Thwarting is distinctive from a negative externality caused by joining the queue that has been studied in the prior literature. Thwarting deals with additional negative externalities on other customers, beyond joining the queue, due to change in customers' behavior when they experience congestion.

Sasser (1976) identified customers as a "mixed blessing" in operating environments. On the one hand, customers could be a source of productive capacity by providing labor for self-service; on the other hand, they could introduce variability and uncertainty into operations. Frei (2006) further classified such customer-introduced variability into five categories (i.e., arrival, request, capability, effort, and subjective preference). We identify yet another type of variability induced by customers who face congestion. More importantly, this type of variability is driven by willful behavioral changes on part of strategic customers who are responding to the congestion they face.

Staffing problems have been studied in other service systems such as call center and healthcare settings (Gans et al. 2003; Green 2004). Recently Batt and Terwiesch (2016) empirically found that service rate is dependent on workload in a hospital's emergency department. Earlier papers have argued that high congestion levels may require servicepersons (hereafter, "servers") to multitask in parallel, which involves a cognitive switching cost (Kc 2013; Batt and Terwiesch 2016) and fatigue (Kc and Terwiesch 2009) that lead to server slowdown. Because customers themselves perform most activities in a fitting room, a self-service environment in general, servers do not cause an increase in service time in our research setting. Our paper therefore emphasizes *customer*-induced service slowdowns as opposed to *server*-driven slowdowns.

The importance of labor as a key execution issue in stores has been highlighted in Raman et al. (2001). Other examples of store execution issues are inventory record inaccuracy

(DeHoratius and Raman 2008) and phantom stockouts (Ton and Raman 2010), which were found to significantly impact retail store performance negatively. Using survey data collected from a small-appliance and furnishings retail store, Fisher et al. (2007) showed that labor issues, as an example of store execution issues, considerably affect both customer satisfaction and sales. In an apparel setting, overcrowded fitting rooms often result in temporary phantom stockouts within the store. Such customer-induced misplaced inventory *within the store*, different from phantom inventory *in the backroom* (Ton and Raman 2010), is expected to be one driver of the inverted-U relationship that we observe in our setting.

Ton and Huckman (2008) provided further evidence of the importance of store labor by demonstrating a link between an increase in employee turnover and a decrease in profit margin and customer service. In addition, Ton (2009) showed that an increase in store labor is associated with higher profits. Our paper is consistent with this work in demonstrating that fitting room associates can help improve store performance. More importantly, we conduct a field experiment involving labor intervention that has not been conducted in this stream of research thus far. While prior literature has argued for increasing store labor to drive sales, our field experiment shows that specifically using labor to perform a backend recovery operation of misplaced items in fitting rooms can increase sales significantly.

Recently, several authors have used retail traffic data in their empirical analyses. Perdikaki et al. (2012) studied relationships between sales, traffic, and labor for apparel retail stores. By decomposing sales into conversion rate and basket value, they found that, at an aggregate level, store sales have an increasing concave relationship with traffic; conversion rate decreases nonlinearly with increasing traffic; and labor moderates the impact of traffic on sales.

Our research results suggest that, beyond plateauing, sales *decline* when a self-service environment, a fitting room area in particular, is highly congested.

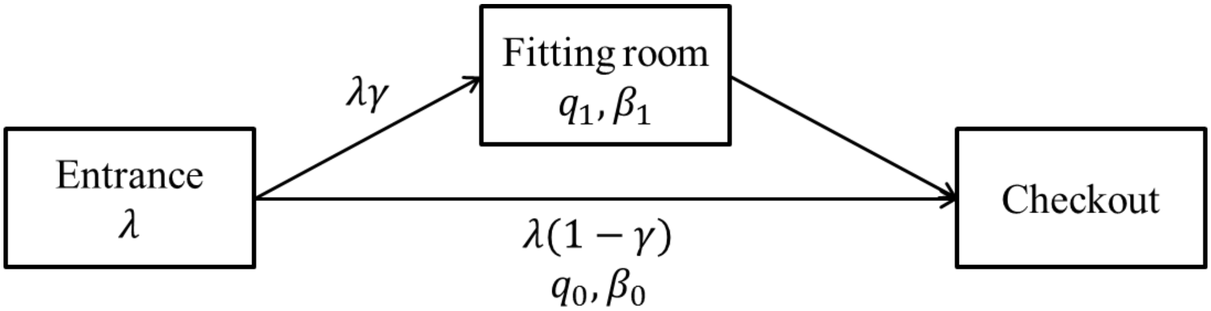
Tan and Netessine (2014) studied the impact of workload on servers' performance in a restaurant chain. They found that servers exert more effort on sales by sacrificing service speed when the overall workload is small, whereas servers start to reduce sales effort and increase service speed as workload increases. Our paper differs from previous papers in two ways. First, we focus more on how *customer behavior* affects service rates; previous work has emphasized servers. Second, while previous studies mostly used overall store traffic data, we demonstrate the value of *in-store* traffic data by showing the negative impact of congestion in fitting rooms on sales. Furthermore, we provide evidence for two thwarting behaviors through a field study observing customer behavior, which can explain the inverted U-shaped relationship observed in our data.

Finally, a stream of literature has focused on the self-service setting. Given traditional service (e.g., a teller in banking), most prior papers have studied the impact of introducing self-service technology (e.g., online banking) on customer satisfaction and retention in a number of settings (Buell et al. 2010). Buell and Norton (2011) revealed that engaging in operational transparency, which they termed labor illusion, is sufficient to increase perceived value. While those papers focus on the impact of operational change on consumer behavior, our work emphasizes the impact of customers' behavior change on operations. Also, as some prior literature (Moon and Frei 2000; Frei 2006; 2008) has consistently pointed out the importance of customers' impacts on operations, our paper adds to this stream by identifying another type of customer's impacts on operations during congested periods that we call thwarting behavior.

2.3 Theory Development and Hypotheses

We formalize our definition of thwarting behavior using a simple conceptual model in Figure 2.1 and then explain the hypotheses we plan to test. Consider a store where store traffic is λ per hour. Assume that a fraction (γ, where $0 \leq \gamma \leq 1$) of those customers enter the fitting room area. We call this $\lambda\gamma$ as fitting room traffic. The remaining customers, $\lambda(1-\gamma)$, leave the store without entering a fitting room. Assume that the conversion rate of customers who use fitting rooms to be q_1 and those who do not to be q_0. Let β_1 and β_0 be the basket values of the purchases made by customers who use fitting rooms and those who do not, respectively. We assume $q_1 > q_0$ as recent paper (Musalem et al. 2016) finds higher conversion rate of fitting room users than non-users. We also assume $\beta_1 > \beta_0$ as anecdotal evidence suggests that customers might need to use fitting rooms when they purchase high price clothes (e.g., formal suit) whereas they might not need to try on when they purchase low price clothes (e.g., basic white t-shirt). Finally, we assume that γ depends upon the waiting time in the fitting room area for impatient customers while conversion rate and basket value parameters are affected by product availability. The store sales in any given hour are the sum of sales from customers who used fitting rooms and those who did not, i.e., $Sales = \lambda\gamma q_1\beta_1 + \lambda(1-\gamma)q_0\beta_0$.

Figure 2.1: Conceptual Model



Now we construct hypotheses to investigate the relationship between fitting room traffic and store sales using this simple setup. We consider three possible scenarios.

Scenario 1: Traditional operations management literature assumes that customers are passive (i.e., patient) or have limited strategic behavior where they can only balk or renege (i.e., impatient). We consider the implications of this assumption on the relationship between fitting room traffic and sales in this scenario.

Fitting rooms play a critical role for apparel retailers as customers experience the product and evaluate alternatives there to complete their purchase. These steps have been identified as part of the core pathway in the consumer purchase decision process. Hence, as fitting room traffic increases, sales are expected to increase initially, since more shoppers reveal their intention to purchase. In the conceptual model, we see that sales increases with γ. If customers are assumed to be passive, sales are expected to increase linearly with fitting room traffic till fitting room capacity is reached. Once it's reached, customers may passively queue outside the fitting rooms so sales will flatten. In the case of limited strategic behaviors such as balking and reneging, sales would have an increasing concave relationship with fitting room traffic. For example, our model can capture balking by setting γ to decrease with the level of congestion. Those customers who balk or renege will end up not using fitting rooms. As the expected sales are higher for fitting room users than non-users, we anticipate sales to be increasing at a diminishing rate with higher fitting room traffic.

Since we expect some customers to balk and renege, we propose the following relationship based on operations management literature:

Hypothesis 1a: *There is an increasing concave relationship between fitting room traffic and store sales.*

Scenario 2: In this scenario, we consider the case where customers exhibit thwarting, beyond reneging and balking. Specifically, we assume that customers change their behavior when they experience congestion in ways that impose negative externalities on other customers that causes sales to decline.

Customers could become strategic by taking more clothes into the fitting room when they experience congestion. When the store is congested, customers might anticipate longer waiting time to secure a fitting room. So they may take more alternatives with them to reduce the chance of returning back to the selling floor. They may also take more alternatives for fear of stocking out when many customers are around. Taking more clothes into the fitting room would increase waiting time for the rest of the customers as the service time would increase. In this case where service time is state-dependent, the waiting time for other customers could grow much faster than usual, which will increase abandonments significantly (much smaller γ). This could lead to decline in sales at higher levels of fitting room traffic.

Furthermore, there is another type of strategic behavior that could affect product availability in the store leading to lower sales at higher levels of congestion. Customers could leave behind more unwanted clothes in the fitting room when they face congestion, instead of purchasing more or replacing those items back in the shelf or a recovery rack. Even those customers, who secure a fitting room but are unhappy with the fit, color, or design of their initial selections, may be reluctant to go back into the store to find alternatives due to the crowd. Literature in psychology has shown that customers in crowded stores feel disoriented (Dion 1999), less satisfied (Eroglu and Machleit 1990), more stressed, and tenser (Langer and Saegert 1977). So, they may just leave those unwanted clothes in the fitting room. Though store associates may eventually return these clothes to the proper locations, temporary phantom stockouts (Ton and Raman 2010) are likely to occur and affect sales, especially during congested periods when associates may be preoccupied with other customer-facing tasks. This phantom stockout could increase lost sales as customers cannot find the items they are looking for, resulting in lower conversion rates and basket sizes for both fitting room users as well as others.

Leaving behind more clothes in the fitting room could also make rooms dirty. It might cause customers to avoid some fitting rooms and increase waiting time for other rooms. This could lead to an increase in abandonments (smaller γ). Accordingly, we propose the following relationship based on thwarting:

Hypothesis 1b: *There is an inverted U-shaped relationship between fitting room traffic and store sales.*

Although Hypothesis 1b is based on thwarting behavior, there are other reasons that could drive inverted-U shaped relationship between fitting room traffic and store sales. For example, the rest of the store is likely to be congested when the fitting room is congested so the store sales could drop due to decrease in q_0 and β_0 as a result of poor customer service in the rest of the store, phantom stockouts in the rest of the store, or long waiting times in the checkout counter. So, the presence of an inverted-U relationship is not by itself an evidence of thwarting behavior.

Scenario 3: Next we argue for an alternate relationship between fitting room traffic and sales in which sales has an increasing convex relationship with fitting room traffic based on a different underlying consumer behavior. This could happen if customers who visit the store during congested hours are unable to find an unoccupied fitting room to try on the clothes at the store so they decide to purchase multiple items with an intention of returning some later (q_0 and β_0 increase as traffic increases). Further, even customers who find a fitting room may decide to purchase more as they take more items into the fitting room when they face congestion (q_1 and

β_1 increase as traffic increases). Both explanations would increase sales non-linearly when the fitting rooms are busy. Accordingly, we propose the following relationship:

Hypothesis 1c: *There is an increasing convex relationship between fitting room traffic and store sales.*

2.4 Data and Methodology for Archival Data Analysis

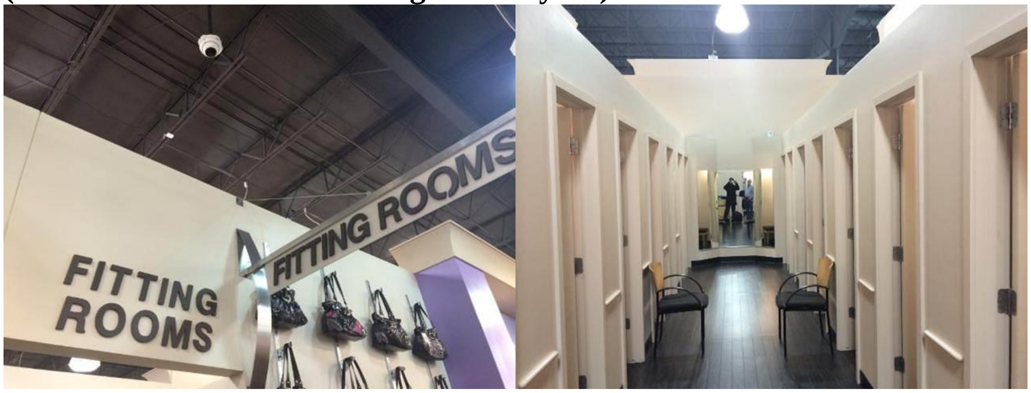
2.4.1 Data sources

We test our hypotheses using archival data obtained from RetailNext and one of its retail clients. RetailNext is a leading in-store analytics provider to retailers such as Sears. It collects traffic information from video cameras in retail stores to codify customer arrival patterns and customer pathways in the stores. In addition to traffic data, we further obtain POS and labor data from one of their clients, a large U.S.-based retailer (retailer A). Retailer A is an apparel retail chain that sells primarily women's, men's, and children's apparel, along with some home décor goods. We worked closely with both companies by interacting with the retailer's senior management, corporate planners, and store management team. The study period is from July 2012 to October 2013.

We obtain the following data for retailer A during the study period. First, we obtain all POS information recorded through scanner. We aggregate it to hourly level to possess store sales volume per hour ($). Second, we obtain labor data which allow us to calculate the number of employees in the store. Finally, we possess traffic data. Retailer A installed video cameras at store entrances to count the number of visitors. All retailer A's stores had this entrance camera during our study period. Only one store, however, had additional cameras installed within the store to track customer movement. We therefore focus on this store to study the impact of fitting room traffic on store sales. These cameras capture only the entrance of the fitting room area and nothing within fitting rooms' interiors. The first picture of Figure 2.2 shows the fitting room area at retailer A used in this study, located at the center of the store for ease of approach. The store has 16 fitting rooms in total in this area.

Figure 2.2: Fitting Room Areas

(Retailer A: Centralized fitting room layout)



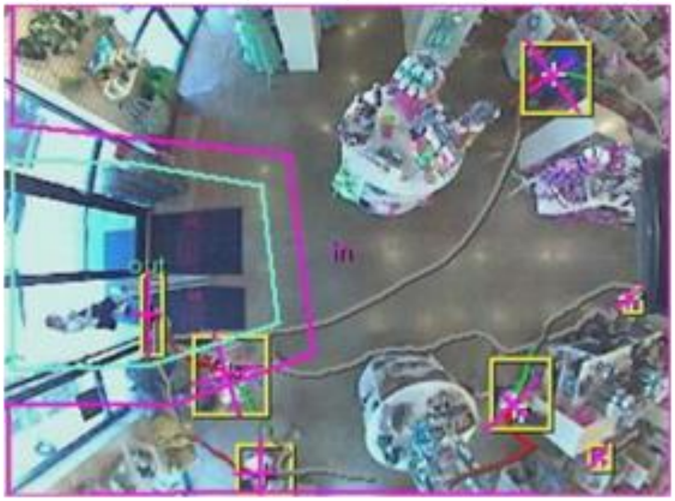
Note. Fitting rooms are located at the center of the store. This store has 16 fitting rooms in total.

(Retailer B: Decentralized fitting room layout)

Note. Fitting rooms are located near each brand. This fitting room area has 3 rooms in total.

Traffic cameras used in this study were able to differentiate between incoming and outgoing traffic by tracking the direction of customers' movements. Figure 2.3 shows how this technology can distinguish incoming from outgoing traffic. Each camera has two sensors, and if a customer goes through both, she is counted. The camera captures the direction of movement by determining the order in which a customer's motion is detected by the two sensors. Consider the two outlines around the entrance door in Figure 2.3. If a customer goes through the outside line (i.e., farther from the entrance) and then the inside line (i.e., closer to the entrance), in that order, then she is categorized as an “out-count,” and vice versa. We had this time-stamp records for every individual who passes the two sensors. We aggregate the time-stamp data to the hourly level, to match the other data. RetailNext audits the data regularly by manually counting the number of visitors and comparing that count to the numbers from the automated sensors, ensuring that the accuracy is at least 95%.

Figure 2.3: Distinguishing Incoming from Outgoing Traffic



2.4.2 Variables

We conduct our econometrics analysis at the hourly level.

2.4.2.1 Dependent variables

We measure the store performance on day t at hour h by sales in dollars ($Sales_{th}$). We find that store hours are not fixed. For example, before the Christmas holiday, the store is open until midnight. In order to avoid the spurious correlation that could arise between variables as a result of systematic differences in business hours, we use only data between 9 AM and 10 PM, which are the normal store hours.

2.4.2.2 Key variables of interest

We have in-count and out-count traffic measures for fitting room traffic ($Fit\_Traffic_{th}^{IN}$ and $Fit\_Traffic_{th}^{OUT}$). We use an average between the in- and out-counts ($A\_Fit\_Traffic_{th} = (Fit\_Traffic_{th}^{IN} + Fit\_Traffic_{th}^{OUT})/2$) as it would help mitigate concerns about measurement errors in the respective in- and out-count variables. As a robustness check, we repeat analysis by using in- or out-counts separately instead of their average and obtained consistent results.

2.4.2.3 Control variables

We next describe the rest of the controls used in our analysis. First, we need to control for store traffic to distinguish the impact on sales of fitting room traffic from overall store traffic. We use the average store traffic ($A\_Traffic_{th}$), similar to fitting room traffic, to control for the number of customers visited the store. Second, we control for store labor ($Labor_{th}$), defined as the number of employees working in the store, as that is known to affect sales (Perdikaki et al. 2012). We eliminate backroom labor from our main model since that does not affect sales directly, though our results are similar when we include backroom employees in $Labor_{th}$. Since the retailer collected time-stamp data on when each associate started work, $Labor_{th}$ can be fractional. For example, if one employee starts to work at 9:30 AM in the store, then we have a data point of 0.5 for $Labor_{th}$ at h = 9.

Third, store sales depend on promotions (Lam et al. 2001). We have information regarding the store's promotional activities. Based on the information, we create a dummy variable $Promotion_t$ that is set to one on days, t, when a promotion was ongoing, and set to zero, otherwise. Finally, we control for seasonality by introducing hourly dummies, day-of-the-week dummies, and monthly dummies.

We trim our data by excluding extreme values to ensure that our analyses are not influenced by extreme outliers and to obtain more robust statistics and estimators, though all results are consistent when we use full data. We remove all data with standardized residuals more extreme than 3. This resulted in a drop of 1.5% of observations. We perform all further analysis on this data set. We check the robustness of our analysis with cutoffs for standardized residuals at 2 and 2.5 and obtain consistent results.

Table 2.1 provides summary statistics for all variables used in our analysis. Subscript t denotes each date and h, ranging from 9 to 21, denotes each hour. The average hourly sales volume is $1,887. The average hourly store traffic is 111, while the average fitting room traffic is 66, indicating that majority of the customers use fitting rooms. Table 2.2 shows the Pearson correlation coefficients among all variables used in our analysis. Correlations between predictors are generally quite low, except for the correlation between $A\_Fit\_Traffic_{th}$ and $A\_Traffic_{th}$, which is quite high (0.79). This may pose multicollinearity problem. We explain how we deal with it in the next section. After taking care of it, we find the variance inflation factors (VIFs) to be below 10, indicating we are not likely to have multicollinearity problems.

Table 2.1: Summary Statistics of the Variables (Retailer A)

Variable name	**Mean**	**Std. dev.**	**P5**	**P25**	**P50**	**P75**	**P95**
$Sales_{th}$	1887.41	1317.21	271.37	986.5	1627.89	2480.47	4403.71
$A\_Traffic_{th}$	111.31	70.67	26	65.25	96.5	141.5	244
$A\_Fit\_Traffic_{th}$	65.97	41.13	16	37	57	86	143.5
$Labor_{th}$	8.64	3.41	5	6	8	10	14
$Promotion_{t}$	0.24	0.43	0	0	0	0	1

Note. Number of observations is 5312.

Table 2.2: Pearson Correlation Coefficients (Retailer A)

	(1)	**(2)**	**(3)**	**(3)**	**(4)**
(1) $Sales_{th}$	1.00				
(2) $A\_Traffic_{th}$	**0.87**	1.00			
(3) $A\_Fit\_Traffic_{th}$	**0.71**	**0.79**	1.00		
(4) $Labor_{th}$	**0.54**	**0.61**	**0.39**	1.00	
(5) $Promotion_{t}$	**0.09**	**0.14**	**0.10**	**0.20**	1.00

Note. Bold denote statistical significance at the 1% level.

2.4.3 Test of inverted-U relationship

There are many ways proposed in the literature to test for the inverted-U relationship, though none of them is perfect. The most common method is a quadratic functional form regression using the coefficient estimates of the linear and quadratic terms. Two confirmatory tests for the inverted-U relationship can be performed by ensuring the sign flips within the range of the data. One is to verify that the slopes of either side of the peak point are significant and of opposite signs (Aiken and West 1991) and the other is to confirm that the confidence interval of the peak point lies within the sample (Lind and Mehlum 2010). While these tests have been performed by thousands of papers, they suffer from the disadvantage that they are all based on the assumption of quadratic form. We may avoid this quadratic assumption and test for inverted-U relationship using spline regressions, a simple two-line test proposed in Simonsohn (2016), or including higher order powers. In this paper, we use all of these methods to test for the inverted-U relationship. Our primary results are based on the quadratic model and the robustness checks consider the various alternate approaches.

We propose the following model to relate store sales to fitting room traffic, with control variables:

$$\begin{aligned} Sales_{th} = \beta_0 + \beta_1 A\_Fit\_Traffic_{th} + \beta_2 A\_Fit\_Traffic_{th}^2 + \beta_3 A\_Traffic_{th} \\ + \beta_4 A\_Traffic_{th}^2 + \beta_5 Labor_{th} + \boldsymbol{W_{th}}'\boldsymbol{\beta_6} + \varepsilon_{th} \end{aligned} \tag{2.1}$$

where $\boldsymbol{W_{th}}$ is a column vector of control variables that includes promotion indicator, hourly dummies, day-of-the-week dummies, and monthly dummies. The availability of promotion variable is especially valuable as it enables us to overcome endogeneity concerns between traffic and sales.

An important consideration in the choice of our model specification is multicollinearity. This concern is especially significant in our model because we not only use store traffic and fitting room traffic which are highly correlated (0.79) but also use quadratic terms of these variables. An additional source of multicollinearity concern is that labor schedules are typically highly correlated with the hour of the day. Using all of these variables in (2.1) results in some variables' VIFs exceeding 10. So, we mitigate multicollinearity concerns in the following way. First, we mean center $A\_Fit\_Traffic_{th}$ and $A\_Traffic_{th}$ with their quadratic terms because mean centering can potentially alleviate multicollinearity issues (Aiken and West 1991). This significantly reduces VIFs, but some of them are still above 10. So, we further remove hourly dummies and replace with hour-block dummies. We clustered 9 AM–12:59 PM, 1 PM–5:59 PM, and 6 PM–10 PM based on common traffic patterns across hours in those blocks. The average store traffic across these three blocks of hours are 97.23, 147.49, and 79.87. So, the second block captures the store's peak hours. By using mean centered variables and hour-block dummies, we ensure VIF of all variables to be below 10 in all models. In addition, we interact the hour-block dummies with day-of-the-week dummies to capture the heterogeneity in customers across different days of the week during same periods. We finally ensure that our finding (i.e., inverted-U relationship between fitting room traffic and sales) is robust with hourly dummies (see Table 5.1.1 in the Appendix).

We test Hypothesis 1a-1c, the relationship between fitting room traffic and store sales, using estimates of coefficients β_1 and β_2. Since some customers may purchase without entering the fitting room, we control for store traffic in the model. We also add the quadratic term of store traffic, $A\_Traffic_{th}^2$, as Perdikaki et al. (2012) found non-linear relationship between store traffic and sales.

Finally, the coefficient of labor (β_5) is subject to endogeneity bias. So, we use an instrumental variable two-stage least squares (2SLS) technique to estimate this model where we consider two sets of instruments. The first instrument is a lagged labor variable from 7 days before. We argue lagged labor variables are valid instruments for the following reasons. First, the labor schedules do not change drastically from week to week so the current schedule is a good predictor of one-week ahead schedule for workers. Second, lagged labor variables satisfy the exclusion condition since they do not impact current period's sales. Lastly, lagged values of labor have been commonly used as instruments in many settings (Bloom and Van Reenen 2007; Siebert and Zubanov 2010; Tan and Netessine 2012). This instrument is not ideal in the present of common demand shocks that are correlated over time. So, we add a second set of instruments.

The second instrument is the labor variable of other store located in the same county under the same retail chain. As the labor cost within a county would be highly correlated, this market-based instrument serves as an exogenous cost-based labor-supply shifter. We chose county-level, as opposed to state-level because labor costs would be different across stores in different counties resulting in weak instruments. As stores in the same county face the same labor market condition, the labor schedule at one store could well predict the other store's schedule. This instrument also satisfies the exclusion restriction since it does not impact the focal store's sales. In addition, this market-based instrument has been commonly used in many settings (Nevo and Wolfram 2002). We assess the validity of the instrument and strength of it, and report the results in the next section.

2.5 Results: The Negative Impact of Congestion

First, we run (2.1) to test Hypothesis 1a-1c using ordinary least squares (OLS) methodology, which does not correct for endogeneity (columns (1) and (2) of Table 2.3).

Column (1) comprises only control variables. We then enter the fitting room traffic and its square in columns (2) so this model can serve as full model.

The results from column (2) support our conjecture that fitting room traffic has an inverted-U relationship with sales. In this model, we find that the coefficients of A_$Fit_Traffic_{th}$ (1.69, p<0.01) and $A\_Fit\_Traffic_{th}^{2}$ (-0.04, p<0.01) are both statistically significant. Since our variables are mean-centered, the coefficients imply that the peak point of sales is about 1.69/(2*0.04) ≈ 24.02 of fitting room traffic. This is about 58% of one standard deviation (41.13) above the sample mean (65.97) and lies well within the support of the data, indicating an inverted-U relationship. In other words, for low levels of fitting room traffic, sales increase with increasing fitting room traffic; however, beyond the peak point, we find that an increase in fitting room traffic is associated with lower sales. This reveals the negative impact of congestion in the fitting room. Too much traffic in the fitting room can hurt store sales (Hypothesis 1b).

We find that the estimated coefficients of the control variables are in the expected direction. In column (1) we find that store traffic has an increasing concave relationship with sales. This is consistent with Perdikaki et al. (2012). While Perdikaki et al. (2012) claim that decline in service quality is a driver of the diminishing return to store traffic, they do not provide any evidence. In our setting, when we add fitting room traffic and its square (columns (2) and (3)) we find that the quadratic term of the store traffic is no longer significant. It implies that the diminishing return to store traffic that was observed in column (1) is largely driven by the congestion-driven phenomenon in the fitting room area for this retailer. This further supports our claim that managing congestion in fitting rooms is crucial at this retailer.

Store labor is positively associated with sales (17.40, $p<0.01$). As we discussed earlier, coefficient of labor is biased due to endogeneity issues, so we will discuss it later using instruments. As expected, we find significant seasonality in store sales. For both models, we observe statistically significant monthly dummies and interactions between day-of-the-week dummies and hour-block dummies. We examine the individual dummy variables to endure that it is consistent with prior expectations of seasonality (available from the authors upon request). Consistent with prior literature (Perdikaki et al. 2012), promotions are negatively associated with sales (-106.81, $p<0.01$), meaning that for a given level of store traffic a randomly chosen shopper in the store is likely to spend less during promotional period. However, because of much higher traffic during the promotion, overall sales are higher. The adjusted R^2 is 82.04%, indicating that our model fit is good.

We now discuss the results from our instrumental variables regression (column (3)). Our main finding about the inverted-U shaped relationship between fitting room traffic and store sales is present, supporting Hypothesis 1b. All coefficients are quite similar with OLS estimation, but as expected, we find stronger impact of labor on store sales. After correcting for the endogeneity bias, the coefficient of labor variable is 112.85 ($p<0.01$).

To assess the validity of the instruments, we perform several statistical tests to examine whether they meet the relevance criteria. First, the R^2 value from the first stage regression of the endogenous labor variable is 0.69, indicating that the instruments have significant explanatory power. In this first stage regression, the coefficients of the instrumental variables are statistically significant with expected sign. We also check the simple correlation. The correlation between labor and other store's labor is 0.62 and correlation between labor and 7-day-lagged labor is 0.74. Second, the F-statistics of the excluded instruments in the first stage regression is well over 10 in our regressions, indicating that the instruments are not "weak" in the sense of Staiger and Stock (1997). Using lagged labor may not be an ideal instrument in the event of common demand shocks that are correlated over time. Although we use other instrument together with lagged labor, we adjust for these common demand shocks, which are basically trends (Villas-Boas and Winer 1999), in our models by adding monthly dummy variables, thus mitigating this concern. We further rule out serial correlation in our model by adding trend as a robustness check and obtain qualitatively the same results. Although not conclusive, these test statistics build our confidence that our instruments are valid.

Table 2.3: Inverted-U Relationship (Retailer A)

Dependent Variable:	**Sales ($Sales_{th}$)**		
	(1) OLS	**(2) OLS**	**(3) 2SLS**
$A\_Fit\_Traffic_{th}$		1.69***	1.53***
		(0.45)	(0.58)
$A\_Fit\_Traffic_{th}^2$		-0.04***	-0.03***
		(0.003)	(0.004)
$A\_Traffic_{th}$	14.33***	13.71***	11.81***
	(0.21)	(0.29)	(0.38)
$A\_Traffic_{th}^2$	-0.006***	0.001	-0.001
	(0.0008)	(0.001)	(0.001)
$Labor_{th}$	22.63***	17.40***	112.85***
	(3.49)	(3.48)	(12.50)
$Promotion_t$	-105.96***	-106.81***	-66.67**
	(24.29)	(24.03)	(28.71)
Controls	Yes	Yes	Yes
Observations	5312	5312	3696
Adjusted R^2	0.8156	0.8204	0.8294

Note. *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The following control variables were included in all of the regressions: interactions between hour-block dummies and day-of-the-week dummies, and monthly dummies. Column (3) has fewer observations due to paucity of data on instruments.

2.5.1 Robustness check: Alternative ways to test for inverted U-shaped relationship

Up to this point, our conclusion about the presence of inverted U-shaped relationship was based on two criteria: (1) negative and significant coefficient of $A\_Fit\_Traffic_{th}^2$ and (2) the peak point, defined as the value of fitting room traffic at which sales are highest, lies within the data range. However, a quadratic approximation of a concave unimodal relationship could be erroneous in the presence of extreme observations. So, as we mentioned in §2.4.3., we use various alternative ways to test for the inverted U-shaped relationship between fitting room traffic and sales. The first two methods are based on the quadratic assumption while the remaining three methods avoid such assumption.

First, we perform a robustness check based on Aiken and West's (1991) procedure for testing curvilinear relationship. Herein we compute the slope of the curve for different values of the variable and ensure that the slope differs significantly from zero and has different signs on either side of the peak point. In model (1), the slope and the standard error can be obtained by $\beta_1 + 2\beta_2 A\_Fit\_Traffic_{th}$ and $\sqrt{\sigma_{11} + 4A\_Fit\_Traffic_{th}^2\sigma_{22} + 4A\_Fit\_Traffic_{th}\sigma_{12}}$, respectively. Here σ_{11} and σ_{22} are the variance of β_1 and β_2, respectively, and σ_{12} is the covariance between β_1 and β_2. Table 2.4 shows that tests of the slopes for the (mean centered) fitting room traffic at the peak point, and ±1 SD, minimum, and maximum value in the sample. We find that the slopes are positive and significant in the region below the peak point while the slopes are negative and significant in the region above the peak point, confirming Hypothesis 1b.

Second, we calculate the confidence interval of the peak point and ensure that it lies within the sample, following Lind and Mehlum (2010). We use the Fieller (1954) method, which has been recommended by Staiger et al. (1997) as the delta method is severely biased for the finite sample. Since the 95% confidence interval lies well within the range of the sample, [12.44, 34.49], we again confirm the inverted U-shaped relationship.

Table 2.4: Robustness Checks for Inverted-U Relationship (Retailer A)

	Sales Equation		
	Value	**Slope**	***p*-value**
Minimum value	-64.65	6.24	0.000
Peak point – 1 SD	-17.11	2.89	0.000
Peak point	24.02	0	0.99
Peak point + 1 SD	65.15	-2.89	0.000
Maximum value	309.85	-20.11	0.000

Note. The p-values are based on one-tailed test on whether the slope is >0 or <0.

Third, we run spline regressions as a further validation of the inverted U-shaped relationship. Spline regressions allow us to check if sales increase first and then drop as fitting room traffic increases. In spline regressions, we choose knots that split $A\_Fit\_Traffic_{th}$ into equal-sized groups. For example, one knot splits $A\_Fit\_Traffic_{th}$ into two equal-sized groups, then estimate a spline regression to fit piecewise linear functions of $A\_Fit\_Traffic_{th}$ *1* (the lower 50%) and $A\_Fit\_Traffic_{th}$ *2* (the higher 50%). The same idea applies to two knots case. In one knot case (column (1) in Table 2.5), we find that the coefficient of the first spline is positive and significant (*p*<0.01), whereas the second is negative and significant (*p*<0.01). It implies that as fitting room traffic increases, sales first increase and then drop, supporting inverted-U. The conclusions are similar when we consider two knots, as shown in column (2).

Fourth, we perform two-line test (Simonsohn 2016). Herein, we estimate two separate regression lines, one for 'low' (i.e., below the peak point) and one for 'high' (i.e., above the peak point) values of $A\_Fit\_Traffic_{th}$. The inverted-U shape is present if the 'low' slope is positive and significant; and the 'high' slope is negative and significant. In column (3) of Table 1.5, we find that $A\_Fit\_Traffic_{th}$ *Low* is positive and significant (*p*<0.05) and $A\_Fit\_Traffic_{th}$ *High* is negative and significant (*p*<0.01). It again supports the inverted U-shaped relationship.

Table 2.5: Robustness Checks without Quadratic Functional From Assumption (Retailer A)

Dependent Variable:	**Sales ($Sales_{th}$)**			
	Spline Regressions		**(3)**	**(4)**
	(1)	**(2)**	**Two-Line Test**	**Higher Order**
	One knot	**Two knots**		**Power**
$A\_Fit\_Traffic_{th}$ *1*	4.30***	6.60***		
	(0.88)	(1.23)		
$A\_Fit\_Traffic_{th}$ *2*	-2.54***	0.02		
	(0.42)	(0.96)		
$A\_Fit\_Traffic_{th}$ *3*		-2.82***		
		(0.48)		
$A\_Fit\_Traffic_{th}$ *Low*			1.58**	
			(0.63)	
$A\_Fit\_Traffic_{th}$ *High*			-4.07***	
			(0.59)	
$High$			20.57	
			(31.06)	
$A\_Fit\_Traffic_{th}$				1.64***
				(0.46)
$A\_Fit\_Traffic_{th}^2$				-0.03***
				(0.006)
$A\_Fit\_Traffic_{th}^3$				-0.00002
				(0.00002)
Controls	Yes	Yes	Yes	Yes
Observations	5312	5312	5312	5312
Adjusted R^2	0.8176	0.8181	0.8175	0.8203

Note. The following control variables were included in all of the regressions: store traffic and its square, store labor, promotion indication, interactions between hour-block dummies and day-of-the-week dummies, and monthly dummies.
*, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Finally, we include cubic power for fitting room traffic in (1) to test for higher order effects in our model. In column (4) of Table 2.5, we still find that $A\_Fit\_Traffic_{th}^2$ is negative and significant (p<0.01) whereas $A\_Fit\_Traffic_{th}^3$ is insignificantly differentiated from zero, again supporting the inverted U-shaped relationship. Including cubic term does not improve adjusted R^2 indicating that the quadratic functional form is parsimonious. We do not add cubic power of store traffic in this model as the VIF increases above ten. Nonetheless, addition of this variable does not change our conclusions about the inverted-U relationship between fitting room traffic and sales.

In conclusion, we identify the negative impact of congestion by showing the inverted U-shaped relationship between fitting room traffic and store sales consistent with thwarting behavior.

2.5.2 Alternate explanations

While the archival data help us demonstrate the presence of an inverted-U relationship between fitting room traffic and sales, these data do not help us draw any causal inferences between thwarting behavior and sales. In fact, there are possible alternate explanations why we may observe the inverted-U relationship between fitting room traffic and sales even when there is no thwarting behavior. The rest of the store is likely to be congested when the fitting room is congested so the store sales could decline as a result of poor service in the rest of the store, phantom stockouts in the rest of the store, long waiting times in the checkout counter, or customers may decide to purchase less because of the crowded environment (Eroglu and Machleit 1990). It is impossible to rule out all alternative explanations without detailed data of all events concurrently occurring in the store during congestion. So, we conduct a field study and a field experiment to provide direct evidence for the thwarting behavior and to quantify its impact on store sales. In the field study we directly observe the behaviors of customers to determine if they exhibit thwarting behavior. While the field study could show if customers indeed change their behavior when they experience congestion consistent with thwarting behavior, it does not explicitly reveal the impact of this change in behavior on sales decline. As we expect thwarting behavior to decrease sales due to increase in waiting time for others leading to balking and reneging as well as phantom stockouts, we subsequently run a field experiment to isolate the impact of one of these factors on a decline in sales to show that thwarting behavior can indeed have a significant impact on store sales.

2.6 Field Study: Evidence of Thwarting Behavior

In this section, we describe a field study where we observe customer behavior when they experience congestion and when they do not in order to provide direct evidence for the presence of thwarting behavior. We use an observational study rather than a field experiment to study the impact of congestion on customer behavior due to the challenges involved in randomizing congestion across subjects (i.e., customers). We explain the ideal field experiment to study the problem and then present our research design, its limitations, and how we overcome potential selection biases in the next sections.

2.6.1 Theory: Ideal experiment

An ideal experiment would require randomizing the level of congestion perceived by each customer entering the fitting room area and their behavior noted. In other words, the treatment would be the level of congestion perceived by each customer and the outcome data would include the number of clothes taken inside the fitting room, amount of time spent in the fitting room, and the number of clothes left behind in the fitting room. We wish to measure the amount of time spent in the fitting room as it is unclear whether customers who take more clothes to fitting rooms speed up their trials as they are aware of waiting customers. This could cause waiting time for other customers to be unaffected even when more clothes are being taken into the fitting rooms. We follow Harrison and List (2004) to formalize the ideal experiment below. Let y_{i1} and y_{i0} be the outcomes of customer *i* under treatment ($T = 1$) and control ($T = 0$), respectively. Then the average treatment effect on the treated can be defined as $ATT = \mathbb{E}(y_{i1} - y_{i0}|T = 1)$. Since one of counterfactuals is missing, we do not observe $\mathbb{E}(y_{i0}|T = 1)$

and $\mathbb{E}(y_{i1}|T = 0)$ in our data. Therefore, the reported treatment effect on the treated is $Reported\ ATT = \mathbb{E}(y_{i1}|T = 1) - \mathbb{E}(y_{i0}|T = 0) = \mathbb{E}(y_{i1} - y_{i0}|T = 1) + \{\mathbb{E}(y_{i0}|T = 1) - \mathbb{E}(y_{i0}|T = 0)\}$. The *Reported ATT* differs from the true *ATT* due to the presence of selection bias, $\{\mathbb{E}(y_{i0}|T = 1) - \mathbb{E}(y_{i0}|T = 0)\}$. The normal approach used by experimentalists is to get rid of the selection bias using randomization.

In our case, we could not randomize congestion in the retail store. The decision on how many items to carry into the fitting room would be made based on whether an individual perceived congestion or not. It may depend upon many factors including checkout queues, the number of customers entering the store along with them, and the number of customers shopping in the same section. Manipulating all of these factors to change individual consumers' perception of congestion is hard for several reasons. First, the retailer we worked with insisted that we do not affect customer operations in anyway. Second, our analysis with archival data shows that congestion could lead to lower sales; hence, getting permission to run such an experiment would be difficult. So, we do not randomize congestion but let it be exogenously determined by store traffic. Since selection bias could be an issue in the absence of randomization of the treatment, we carefully design the field study to handle potential selection biases as explained in the next section.

2.6.1.1 Handling selection biases

A selection bias could occur when customers in the two groups, i.e., those who experience congestion and those who do not, are systematically different. This is possible as congestion is correlated with factors such as day-of-the-week and promotion as traffic tends to be higher during weekends and during promotional periods. Also, prior research shows that weekday (Monday-Friday) customers are different from weekend (Saturday-Sunday) customers

(Lam et al. 2001). So, we design our field study to observe customers under treatment and control conditions during the same day so that we remove any day-specific effects. In addition, it is possible that customers who arrive during peak hours, when there is congestion, could be different from those who arrive during non-peak hours, when it is less likely to have congestion. So, instead of defining congestion based on the time of arrival, we define it based on the state of congestion perceived by each customer when they decide to take clothes into the fitting room. The amount of congestion perceived by each customer could depend upon factors such as the number of visible customers in the store who were entering or leaving it, the length of the cashier line, and the number of customers who were shopping alongside them in the same area of the store. Since these factors are hard to measure, we adopt an alternate metric as a proxy for the level of congestion perceived by each customer in the store. We track the number of occupied fitting rooms when each customer enters the fitting room and use the utilization rate as a proxy for the level of congestion experienced by the customer. So, even within the same peak hour, as defined by the hour-of-the-day rather than traffic-based metrics, it is possible for some customers to experience congestion and for others to not. This measure of congestion also helps us mitigate another type of selection bias due to customer balking. If customers balk during highly congested periods, then the observed behavior is only of patient (or the most motivated) customers who may be different from the overall population. Since balking is more likely to be severe when there are long queues, we avoid an all-or-nothing approach to congestion but measure it at different levels based on room occupancy. Our assumption is that at lower levels of congestion, customers could still exhibit thwarting but they are less likely to balk.

It is also possible that customers shopping in typically congested stores might be different from customers shopping in typically uncongested stores due to heterogeneity across stores such as store management team, associates, and display of items. So, we observe customers in the two groups within the same store. Demographic differences such as gender and age might drive differences in customer behavior as well. So, we observe only a particular women's fitting room area where fitting room users are all female and in a similar ages. Finally, another source of selection bias could occur when customers shopping in different part of the store are different populations. For example, the store manager at Retailer B informed us that customers who shop for formal dresses are generally different from those who shop for swimsuits or casual wear. So, we observe differences in behavior between customers who experience congestion and those who do not in the same fitting room area.

2.6.2 Field study at retailer B

We could not conduct a field study at retailer A due to lack of access, so we sought out another retailer, B, to conduct this part of the study. Retailer B permitted us to position a female research assistant (RA) in the women's fitting room area who was able to collect customer-level and item-level data. After confirming the presence of the inverted-U relationship at retailer B (see Appendix), we use the observational data collected through our field study to provide direct evidence for thwarting behavior. Next, we provide background about retailer B and explain the field study design in detail.

We conducted the field study observing customer behavior at one of Retailer B's stores located in North Carolina in April and May 2015. Retailer B is a U.S.-based department store like retailer A, but larger. This retailer operated around 300 stores in the U.S. as of May 2015 with the average store size being 100,000 sq. ft. The store in which we conducted the field study had two floors. The first floor was mainly for men's, children's apparel, and home goods and the second floor was primarily for women's apparel, accessories, shoes, and cosmetics. The studied store of retailer B had multiple smaller fitting room areas (bottom picture in Figure 2.2) with 3–4 rooms inside each area. For example, Polo Ralph Lauren in the men’s apparel section had its own fitting room area, with 4 rooms.

We chose two fitting room areas, one on the first floor (4 rooms) and the other on the second floor (3 rooms), where the apparel categories and brands around these areas were similar. We chose weekends for our study to observe both customers who experienced congestion and those who did not because we found limited evidence of congestion during weekdays in our preliminary observation. This store had higher traffic on Saturdays compared to Sundays. Both fitting room areas had their own checkout counters (see the bottom picture of Figure 2.2), where we collect POS information.

2.6.2.1 Data collection

A female RA stationed in the fitting room area collected customer-level data during the hours 12 PM–6 PM for four days spanning three weekends. She observed customers entering the fitting room area on two Saturdays and two Sundays. She recorded the times of entry and exit of customers from the fitting room and the number of clothes they were carrying. Since the RA could not interact with customers, the number of clothes carried by customers was only an estimate based on visual inspection. Because customers typically bring few items of clothes and do so with their hangers made it easy to count the number of clothes.

On the contrary, it was difficult for us to track the number of clothes carried back by customers when they exited fitting rooms as they had typically removed clothes from their hangers. Furthermore, we could not count the clothes left-behind by each customer in the fitting room as this would require us to visit the fitting room each time a customer exited and cause disruption to other customers. Therefore, we use an alternative approach to obtain a proxy for the number of items left-behind by each customer. We track the number of items brought by associates from fitting rooms to a recovery rack from 12 PM to 6 PM for two days. Because we knew the number of customers who used the fitting rooms between two consecutive recovery operations by an associate, we could calculate the average number of items left-behind by each customer. Since this analysis precludes the identification of congestion level experienced by each individual customer, we put together customers into two groups; those who shopped on Saturday and those who shopped on Sunday, with an assumption that those who shopped on Saturday experienced congestion while those on Sunday did not. Prior literature (Lam et al. 2001) finds that weekend shoppers may be different from weekday shoppers but no difference between Saturday and Sunday shoppers have been identified so far. Hence we assume that the shoppers are similar so there is no selection bias as a result of our design.

We also measure the magnitude of phantom stockouts among items left behind in the fitting rooms to examine the impact of this thwarting behavior on lost sales. We do so by scanning each item left behind in the fitting rooms using a POS scanner to check the in-store availability of those SKUs.

2.6.3 Results

2.6.3.1 Thwarting behavior 1: Bringing more clothes into fitting rooms increases waiting time for other customers

We observe a total of 209 customers over 24 non-continuous hours spanning three weekends. Using the definition of congestion based on the occupancy rate of the fitting rooms as a proxy for the congestion perceived by each customer (explained in §2.6.1.1.), Table 2.6 shows results of the two-sample *t*-test with an assumption of both equal and unequal variances between two groups. We find that customers brought an average of 4.38 items into the fitting room when they did not experience congestion, whereas they brought an average of 5.76 items into the fitting room when they experienced congestion. So, customers who face congestion bring 1.38 additional clothes into the fitting room (*p*<0.01). It is possible that customers who enter fitting room area during congestion might speed-up their trials as they are aware of others waiting to use the room. However, our results show that it is not the case as they occupy fitting rooms 1.87 minutes longer (112 seconds, *p*<0.05) than those who enter the fitting rooms when there was no congestion. Thus, we find support that the waiting time for other customers increases during congested periods as customers take more clothes into the fitting rooms to try on.

Table 2.6: Impact of Congestion on the Number of Items and Time Spent in the Fitting Room (Retailer B)

		Mean (Std. dev.)	
	# obs (*N*=209)	**Number of items brought into the fitting room**	**Time spent in the fitting room (Seconds)**
Non-congestion	143	4.38 (2.87)	490.45 (373.58)
Congestion	66	5.76 (3.56)	602.61 (332.58)
Difference		-1.38	-112.16
t-test		***p*-value of H_1 (Diff < 0)**	
Equal (unequal) variances		0.0016 (0.0034)	0.0191 (0.0155)

Congestion: fitting room occupancy of 3 & 4 for the first floor and 2 & 3 for the second floor. Non-congestion: fitting room occupancy of 0, 1, & 2 for the first floor and 0 & 1 for the second floor.

We perform several robustness checks. First, even though we compare customers in the same store on the same day at the same hour, we also run a regression with hour and date controls to see whether our finding is affected by these factors. We still find that customers brought 1.42 extra clothes into the fitting rooms (*p*<0.01) and occupied fitting rooms 1.87 minutes longer (112 seconds, *p*=0.057) when they experienced congestion. This result indicates that our model free evidence does not suffer from selection biases due to hour and date because we carefully dealt with them as explained in §2.6.1.1.

Second, we conduct a two-sample *t*-test on each floor separately and obtain consistent results. Third, we compare the average number of items brought by customers into the fitting room and time spent by them in the fitting room by each state of fitting room occupancy. Compared to low fitting room occupancy (i.e., 0 or 1 occupied rooms when a customer entered a fitting room), we still find that customers who entered the fitting room at the high fitting room occupancy state (i.e., 3 or 4 for the fitting room area on the 1st floor and 2 or 3 for the fitting room area on the 2nd floor) carried more items and occupied fitting rooms longer. Finally, we simply compare the average number of clothes brought by customers into the fitting room and time spent by them in the fitting room on Saturday (high traffic) with those on Sunday (low traffic) and obtain similar results.

2.6.3.2 Thwarting behavior 2: Leaving behind more clothes in fitting rooms increases chance of lost sales due to phantom stockouts

We note that evidence of customers taking more clothes into the fitting rooms during congestion does not automatically imply that customers will leave behind more clothes as well. This is because customers could either choose to purchase more or decide to replace the clothes back in the shelves or recovery rack. In both cases the sales would be high and phantom stockouts due to misplaced inventory will be low.

We track an associate’s recovery operation of items left behind in the fitting room to the recovery rack from 12 PM to 6 PM on one weekend. So we have information about (1) the time at which an associate brought out misplaced items from the fitting room to the recovery rack and (2) the number of those items. We divide the total number of items that an associate brought out from the fitting rooms by the total number of customers who used fitting rooms during associate’s two consecutive cleanups to obtain the average number of items left behind in the fitting rooms by each customer for a given period. Then we compare the average of this metric on Saturday with that on Sunday using a two-sample *t*-test. Consistent with our thwarting behavior argument, we find that, on average, customers left behind 2.69 items more in the fitting room on Saturday than Sunday (4.98 vs. 2.29, *p*=0.069).

Having established that customers leave behind more clothes in the fitting room when they face congestion, we next turn our attention to the magnitude of phantom stockouts among those misplaced items. If the misplaced items in the fitting rooms happen to be the last available items in the store then lost sales can occur due to phantom stockouts. To check the magnitude of phantom stockouts among misplaced items in the fitting room, we scanned 559 items moved from fitting rooms to the recovery rack over another 12 hour period across two days (Table 2.7). Among 559 misplaced items, we find that 38.6% (216 items) were unique items in the store. In other words, 38.6% of the items left behind in the fitting rooms experienced phantom stockouts. We also find that 34.7% of these misplaced items had only one additional inventory in the store according to the POS system. Since inventory records are updated once a day, it is likely that many of these items might be the last inventory in the store, making our estimate of 38.6% phantom stockouts conservative.

Table 2.7: Phantom Stockouts (Retailer B)

# of items	**Total**	**# of available items in the store**				
(%)	**items**	**1**	**2**	**3**	**4**	**Over 5**
Day 1	223 (100%)	78 (34.98%)	88 (39.46%)	25 (11.21%)	16 (7.17%)	16 (7.17%)
Day 2	336 (100%)	138 (41.07%)	106 (31.55%)	55 (16.37%)	22 (6.55%)	15 (4.46%)
Total	559 (100%)	216 (38.64%)	194 (34.7%)	80 (14.31%)	38 (6.8%)	31 (5.55%)

Note. Day 1 is Sunday and day 2 is Saturday. Detailed information (e.g., price) is obtained in day 2.

The high volume of phantom stockouts that we observe can be especially costly to retailers if the misplaced items have high prices, so we next compare the prices of items misplaced in the fitting rooms with the prices of other items in the immediate sales area. We collect price information for items left behind in the fitting rooms over a six hour period in day 2. In addition, we also collect information on size, clearance item or not, and the number of swimsuits because the store manager suspected that phantom stockouts could be driven by these factors. Of the 336 misplaced items in day 2 (Table 2.7), we found 138 items experienced phantom stockouts. Among these 138 items, 7% were clearance items (10 items) and 15% were swimwear (21 items). In addition, phantom stockouts occurred in all sizes, so they were not restricted to extreme sizes. The average price of items left behind in the fitting rooms was $53.77 and the average price of phantom stockout items was $61.53. Using price data for all SKUs in the immediate selling floor near the fitting room area, we found the average price of 3,200 SKUs to be $46. This indicates that the items left behind in fitting rooms were indeed high-price items and the phantom stockout items are even expensive items with significant opportunity costs. This is consistent with our assumption on higher basket value for fitting room users in the conceptual model (Figure 2.1). Further details about phantom stockouts are available in the Appendix (see Tables 5.1.3 and 5.1.4).

To summarize, we find evidence for two mechanisms that impose negative externalities on other customers due to the thwarting behavior of customers. Our results show that customers who face congestion carry a few more extra items (1.38) into the fitting rooms and leave more items behind (2.69) compared to those who did not face congestion. Recall that customers take 5.76 clothes during congestion and leave behind 4.98 of them. These results suggest a double-whammy effect where not only do customers purchase less when they face congestion, as they are leaving behind almost everything that they take inside during congested periods, but also induce negative externality as they reduce the likelihood of purchase by other customers due to an increase in waiting time and phantom stockouts.

2.7 Field Experiment: Mitigating the Negative Impact of Phantom Stockouts

While the field study shows that customers exhibit thwarting behavior, it does not explicitly reveal the impact of this behavior on sales decline. As we explained earlier, there are several alternative explanations for sales decline so it is unclear whether thwarting behavior alone can cause significant drop in sales. In this section, we isolate the impact of thwarting behavior on store sales.

We expect thwarting behavior to decrease sales due to increase in waiting time for others leading to balking and reneging as well as phantom stockouts. Quantifying lost sales directly is hard as it requires us to observe all incidences of balking, reneging, and phantom stockouts due to thwarting behavior. Hence we take an alternative approach. We mitigate the negative consequence of thwarting behavior and measure the increase in store sales. We may do so by speeding up customer operations to reduce waiting time for others and by providing a backend recovery operation in fitting rooms to move misplaced inventory to the recovery rack where it can be re-shelved. Retailer B did not wish to pursue any options that involve direct interactions with the customers so we could not speed-up customer operations. Instead we focus on an intervention aimed at reducing phantom stockouts by recovering misplaced items in the fitting room using an associate.

While prior literature has shown the existence of phantom stockouts, its impact on lost sales has not been documented. This impact is not obvious because phantom stockouts do not necessarily lead to large lost sales as customers could contact store associates who might be able to expend effort and find the products. Our experiment to document the impact of phantom stockouts on lost sales can therefore be valuable.

2.7.1 Experimental design

The goal of our experiment is to demonstrate that mitigating phantom stockouts in the fitting rooms, through a backend recovery operation using labor, can significantly improve store sales. This will help us demonstrate that thwarting behavior that causes increase in phantom stockouts has significant detrimental impact on store sales. So, in this experiment, the treatment condition involves use of an associate in the fitting room area to restock the left-behind merchandise and the control condition does not have an associate in the fitting room area.

We choose two fitting room areas in the first and second floors to use one as the treatment and the other as the control during the same period. In other words, we track the sales during the intervention hours at both the treatment and control areas. This ensures that any time-specific effects such as weather, day, peak-hour, or promotion are similar across customers in the treatment and control conditions. By choosing treatment and control in the same store, our design also allows us to control for store-specific factors such as management team, assortment, and competition. We choose two fitting room areas among 20 alternative areas where the apparel categories and brands around them were similar, to mitigate area-specific factors. Nonetheless, it is still possible for the customers using the treatment fitting room area to be different from those using the control fitting room area. So, we rotate the treatment and control between the fitting room areas in the first and second floors to mitigate any remaining area-specific effect. An added advantage of this approach is that we can minimize any spillover effects that can happen with customers drifting from treatment to control or vice versa as we choose treatment and control areas in different floors.

2.7.2 Implementation and methodology

The implementation of our experiment presents a number of challenges. Enduring adherence to experimental design is harder to obtain in the case of experiments with labor as store associates may not follow the instructions we provide. For instance, we require the associate in our field experiment to perform a recovery operation of merchandise left behind in the fitting rooms but not interact with customers as it could confound our results. Further, we find that store managers placed greater emphasis on customer-facing tasks so they were tended to move the associate from fitting room area to tasks that involve directly helping customers. So, we had to be present in the store during the entire duration of the experiment to ensure adherence. This also limits the amount of time we can run the experiment.

We run the experiment for three weekends in April and May 2015 from 12 PM to 6 PM. On one of the days, we could find the additional associate for only 5 hours, so our total number of hourly (treatment) observations was 35. Table 1.8 shows the design of the experiment, which lasted for a total of four weeks, including post-experimental period. We collect our last week of data in the absence of the associate, to further confirm that we were not capturing an overall time-trend effect. We obtained hourly sales information for two fitting room areas from their own checkout counters.

Table 2.8: Experiment Design (Retailer B)

	Fitting room on the 1st floor	**Fitting room on the 2nd floor**
Week 1	Treatment	Control
Week 2	Control	Treatment
Week 3	Treatment	Control
Week 4	Control	Control

Control: Current staffing practice. Treatment: Having an associate restocking items left behind in the fitting room area.

We used the difference-in-differences (DiD) estimator to measure the impact of mitigating phantom stockouts on sales in the corresponding checkout counter. The critical assumption underlying the DiD estimator is the existence of a parallel trend. That is, the two groups would follow the same trend in the absence of treatment. In our context, because the control and treatment fitting room areas were very similar in terms of products displayed around them and they were located in the same store, it was highly likely that the parallel trend assumption held. We also found evidence of a parallel trend from the statistical tests of difference in trends across these fitting room areas.

If the null hypothesis is that, without treatment, sales generated by the fitting rooms on the 1st floor (or 2nd floor) over sales generated by the fitting rooms on the 2nd floor (or 1st floor) is constant, then we could use the following regression to examine the effect of treatment.

$$\begin{aligned} Sales_{fdh} = {} & \alpha_0 + \alpha_1 Treatment_{fdh} + \alpha_2 Promotion_d + \alpha_3 SecondFloor\ Dummy \\ & + \alpha_4 Hour\ Dumm\text{ies} + \alpha_5 Date\ Dummies + \varepsilon_{fdh} \end{aligned} \tag{2.2}$$

where subscript *f* denotes floor, *d* denotes date, and *h* denotes hour. $Treatment_{dh} = 1$ in the first floor for weekends in weeks 1 and 3 and for the second floor for a weekend in week 2 from 12 PM to 6 PM; otherwise, it equals zero. The coefficient of interest is α_1, which can be interpreted as the sales change due to treatment. We control for other factors such as hour-of-the-day, date, day-of-the-week, and month.

2.7.3 Results

Table 2.9 shows that an increase in sales for the checkout counter allocated to the treated fitting room area due to a backend recovery operation provided by an associate in the fitting room area varied from \$186.93 (*p*<0.01, Model 1) to \$177.7 (*p*<0.01, Model 2), depending on control variables included in the regression. We found that the \$186.93 increase in sales constituted a 22.6% increase in average hourly sales at this checkout counter during peak hours. Given that the wage rate was less than $15 per hour and the gross margin was about 40%, we show that this increase in labor in the fitting room area would be profitable. As a robustness check, we performed analysis for each floor and obtained consistent results.

Table 2.9: Impact on Sales of Recovery Operations by Fitting Room Associate (Retailer B)

Dependent Variable: Hourly Sales	**Model (1)**	**Model (2)**
Treatment	186.93***	177.70***
	(52.13)	(52.49)
Promotion	895.66***	942.02***
	(97.60)	(89.74)
Second Floor	-24.86	-25.18
	(20.22)	(20.57)
Hour dummies	Yes	Yes
Date dummies	Yes	No
Week, Day-of-the-week, & Month dummies	No	Yes
Number of observations	648	648
Number of treatments	35	35
Adjusted R^2	0.4404	0.4207

Note. *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

We find that mitigating phantom stockouts in the fitting rooms can improve store sales significantly. This shows that thwarting behavior leading to phantom stockouts can be detrimental to store performance.

2.8 Managerial Implications and Conclusion

To summarize, our archival data analysis shows an inverted-U relationship between fitting room traffic and store sales, consistent with thwarting behavior. Our field study shows that consumers indeed change behavior during congestion that could lead to increased waiting time for others and phantom stockouts. Finally, our field experiment reveals that these phantom stockouts cause significant lost sales in retail stores.

Our study has direct managerial impact at retailers. For example, we demonstrate the need for excellence in store execution by showing a large magnitude of phantom stockouts at retailer B when the fitting rooms were not well attended. Although retailer B’s management was aware of the challenges posed by misplaced inventory, they did not realize the large impact on lost sales. Retailer B had therefore followed a policy of re-shelving items on the recovery racks just before the store closed when store is not crowded. Our study shows that such a policy would lead to significant phantom stockouts during congested periods, leading to lower sales. In response to our findings, retailer B changed its policy to continuously monitor fitting rooms and to pay special attention to them during congested hours.

Other aspects of our findings can be further generalized to other retailers and other organizations in different industries. First, we identify thwarting behavior, defined as a change in customer’s behavior when they experience congestion in a way that induces negative externalities on other customers. Thwarting behavior is especially problematic if organizations have self-service operating systems with little or no monitoring, where customers determine their service speed. According to the recent report of Allied Market Research (Shende 2015), global self-service technology market will expand to $31.75B by 2020 with an annual growth rate of 13.98% during the forecast period 2015–2020. Some of the self-service environments include ticketing kiosks at movie theaters, bus and train stations, and parking lots; ATMs; food-ordering kiosks at restaurants such as McDonald’s and Chili’s; self-checkout kiosks at retail outlets; DVD rental kiosks at redbox; gasoline stations; and bike-sharing programs. Future research can test for the presence of thwarting behavior in these settings. For example, thwarting behavior in bike-sharing programs can manifest if customers had rather pay extra to retain bikes during congested periods rather than return and pick-up later for fear of losing bikes when they need them. In this particular case, while the sales may not be impacted in the short-term such strategic behavior may lead to long-term adoption problems. So, unlike recent studies (e.g., Gavett 2015 and references therein) show that self-service technologies can benefit sales, our study argues for caution and a need for understanding customer behavior especially during congestion.

Second, our paper offers insights on how labor can be used to boost store sales. Retailers tend to be budget-driven and view labor as a short-term expense rather than as a driver of sales (Ton 2009; Fisher and Raman 2010), resulting in understaffing chronically during peak hours (Mani et al. 2015). Under the environment of limited labor resource, using labor in a way to increase store sales is important. Our study demonstrates that the cost of having an associate in the fitting room area performing a backend recovery operation of misplaced merchandise in the fitting rooms could pay for itself and even more from the increase in sales. Using gross margin of 40% for retailer B and less than $15 hourly wage, we can conclude that having an associate in the fitting room area yields a positive ROI.

Like other empirical studies, our research has limitations related to data availability. Our store traffic and fitting room traffic data capture the total number of visits, not the number of visits by unique customers. In other words, if fitting room users make multiple trips to the store, they would be counted as multiple visits. Our results, however, are conservative, as we currently assume one visit per customer. If customers must make multiple visits before making their purchase, the impact of congestion on sales would be larger. Also, no extant technologies allow retailers to distinguish group shoppers, who may arrive with friends or family members to simply observe and offer opinions rather than to make purchases themselves, from single shoppers. This could cause measurement error in our variables for retailer A, but is not relevant for retailer B, where we observed every customer and identified group shoppers. Thus, our data analysis at retailer B relaxes some of these concerns.

Future research could have a number of venues. It would be interesting to understand the causes of thwarting behavior and ways to manage it. For example, how is the thwarting behavior linked to the patience threshold of customers? Researchers have pointed out that it is important for the OR and OM community to incorporate findings in behavioral literature in the design of queueing systems for service firms (Bitran et al. 2008). We hope that this new phenomenon identified in our paper, thwarting, is considered in the design of queueing systems.

CHAPTER 3
Can Incentives Improve Labor Scheduling Decisions? Evidence from a Quasi-Experiment

3.1 Introduction

Over the past several decades, the role of incentives in improving performance has been widely debated. On the one hand, theoretical literature on agency theory, expectancy theory, and goal-setting theory argue that incentives make agents exert more effort that leads to better performance (see Prendergast 1999 for an overview). These theoretical predictions have been validated by empirical research in economics, accounting, and marketing (for example, Banker et al. 2000, Lazear 2000, Paarsch and Shearer 2000, and Oettinger 2001). On the other hand, psychologists and behaviorists have provided a contrary viewpoint. Based on a meta-analysis of over one hundred experimental studies, Deci et al. (1999) conclude that in most instances extrinsic rewards have a negative effect on intrinsic motivation. Despite large empirical research around pay-for-performance incentives in other streams, there is limited empirical research in operations management (OM) that has examined the role of pay-for-performance incentives in improving operational outcomes with field data.

There are multiple reasons why examining the impact of pay-for-performance incentives on operational decisions would be valuable. First, large organizations often invest heavily in software based solutions to aid decision making. This trend is expected to increase with greater use of data analytics techniques such as machine learning and artificial intelligence for decision making. So, it is unclear whether financial incentives for managers are even required to obtain better decisions in organizations. Second, lab-based experiments have shown that the impact of incentives on performance depends upon the type of tasks (Libby and Lipe 1992) and complexity of the task (Pelham and Neter 1995). Since operational tasks are often complex it is unclear whether incentives can improve operational outcomes. Finally, there is a long history of analytical research that designs optimal supply chain contract through economic incentives (see Cachon 2003 for an overview). This literature generally assumes that agents are self-interested and rational and react strongly to incentives. Examining how individual agents make operational decisions in the face of incentives is critical to substantiate the underlying assumptions of this large body of literature.

Using a novel dataset from a quasi-experimental setting at a retailer, we examine if incentives are effective in improving labor scheduling outcomes at a retailer that changed its incentive plan for store managers from strong financial incentives to meeting labor budget to almost zero financial incentives to meeting labor budget. Managers had to stay close to the labor budget as being under the labor budget could signal lower customer service due to understaffing and going over the labor budget implies overstaffing. Using weekly sales (actual and forecasts from manager and software) and labor (actual, planned, and budget) data for 75 stores over 47 weeks, we test whether introduction of new weakened financial incentive plan is associated with worsening of overall labor scheduling outcomes and their underlying decisions (forecasting, labor planning, and execution).

Our field research site provides a nice quasi-experimental setting in which to address the research questions. In the old incentive plan, managers were paid a bonus based on two components. The first component is whether quarterly sales exceed sales in the same quarter in the previous year. For every 1% year-over-year sales growth rate, store managers are eligible to receive 1% of their annual salary as a bonus. The second component was based on the deviation of actual labor from labor budget. The labor budget was calculated based on multiple factors such as actual sales, size of the store, and labor productivity. If store managers exceed their labor budget (after normalized by actual sales) by 0.1%, then their eligible bonus from the first step will be reduced by 10%. In the new incentive plan, stores became eligible for bonuses only when the firm made a profit in that quarter. Since the profit for this retailer had significantly dropped in the previous 5 quarters and became negative in the previous quarter, we argue that this new threshold significantly reduced financial incentives for managers. Thus, the quasi-experiment can be described as a change from a strong financial incentive plan tied to meeting store sales target and labor budget to a weak one[2]. Moreover, tasks, technology, management, monitoring of stores, and other operational practices were the same under both incentive schemes.

Our primary findings are as follows. First, we find that store managers make more accurate labor scheduling decisions under the strong incentive scheme, indicating that incentives for store managers indeed help improve labor scheduling decisions. We measure total error in labor outcomes by the absolute percentage error between actual labor and labor budget which can capture store managers' effort on scheduling labor to be closer to labor budget as both understaffing and overstaffing are worse-off. Overstaffing can result in smaller amount of bonus while understaffing can draw district managers' attention as it can signal lower customer service level. The total error in labor outcomes is 2.44% larger under the weak incentive scheme compared to the strong incentive scheme period. This is a substantial difference as the average total error in labor outcomes is 5.65% in our data. So, the weak incentives have resulted in 43.19% worsening of labor scheduling decisions by managers. Similar to economics literature that has

[2] In the Appendix, we describe an alternative interpretation for the change in incentive and provide an analysis supporting our interpretation.

found huge impact of incentive on productivity (see survey in Prendergast 1999)[3], our surprising finding is not so much that labor decisions become worse with weakening of pay-for-performance incentives but by how much.

Second, we find that the impact of financial incentives on labor scheduling decisions occurs mainly due to change in effort instead of selection mechanism. By considering stores without change in managers during study period, the total error in labor outcomes is 2.29% larger under the weak incentive scheme compared to the strong incentive scheme. Again, comparing to the average total error on labor outcomes of 5.37% for these stores, the weak incentives have resulted in 42.64% worsening of labor scheduling decisions. By showing that labor scheduling decisions become worse-off even after removing selection mechanism, our results demonstrate that incentives play a vital role in making managers exert more effort towards labor scheduling decisions.

Third, by decomposing outcomes of overall labor scheduling into three underlying decisions, we observe that financial incentives have differential impact on forecasting, labor planning, and execution. We find no evidence that incentives improve the accuracy of managers' sales forecast. However, we find that stronger incentives are associated with higher efforts towards forecasting although this higher effort does not translate into more accurate forecasting. The labor planning error, measured by the absolute percentage error between planned labor and labor budget, is 1.3% larger under the weak incentive scheme. Comparing to 5.6% of the average labor planning error, this result shows that weakening incentives have resulted in 23.21% worsening on labor planning decisions. Finally, we find that managers adjust their labor plan to the new information available in the last mile – in the time period between when plan is

[3] For example, Prendergast (2011) says "the surprising thing in this literature is not that productivity seems to go up when there is pay for performance, it is that it goes up by so much. It is not unusual to see productivity numbers going up by 25 percent to 35 percent."

generated and executed – during the strong incentive period whereas they do not adjust much during the weak incentive period. This suggests lower efforts during plan execution under the weak incentive scheme.

This paper contributes to the OM literature in the following ways. Our paper shows the importance of financial incentives for managers even when they are aided by underlying software. According to the recent report[4], the global market of workforce management solutions is estimated to grow from $4,880.3 million in 2015 to $7,725.8 million by 2020, at a compound annual growth rate of 9.6%. Even with such large investments, the role of the store manager in making labor decisions is not diminished as these software decisions may potentially be improved by store managers.

Second, our paper shows that the impact of financial incentives on operational decisions depends upon the type of task. We observe an improvement in labor planning and execution decisions but do not see any improvement in forecasting decisions. Overall, our results show that financial incentives have differential impact on forecasting, labor planning and execution decisions of store managers as each task may require different types of effort and have different sensitivity to effort (Libby and Lipe 1992) and complexity (Pelham and Neter 1995).

Finally, our paper is one of the few studies that empirically examine the impact of incentive on operational outcomes as well as underlying decisions driving those outcomes. The study of pay-for-performance incentive on operational decision making is of great interest. Yet, empirical investigations in the OM literature are scarce partly due to the paucity of data. Notable exceptions include DeHoratius and Raman (2007), Guajardo et al. (2012) and Chan et al. (2014).

[4] See "Workforce Management Market by Solution, by Service, by Deployment, by Organization Size, by Vertical, by Region - Global Forecast to 2020" by marketsandmarkets.com, November 2015.

However, even these papers only examine the impact of incentives on outcomes while we are able to also study the impact on underlying decisions.

3.2 Literature & Hypotheses

3.2.1 Related literature

Agency theory, among others such as expectancy theory and goal-setting theory, is one paradigm that suggests an effectiveness of performance-based incentive. In the basic agency model (for example, Basu et al. 1985, among others), a principal designs a contract to motivate an agent to exert unobservable efforts in a production process that is characterized by uncertainty. Given the performance-based incentive, the agent encounters a trade-off between his dis-utility for effort and his expected utility for higher compensation resulting from improved performance and risk bearing.

Prior literature studied the impact of performance-based incentive on productivity in top executives level as well as front-line employee level. Studied by Abowd (1990), Brooks et al. (2001), Morgan and Poulsen (2001) find that the introduction of incentive for top executives that pays based on their performance results in increased performance. Research concerning front-line employees also documents similar productivity gains from bonus-based incentive scheme consistent with effort effects. Banker et al. (1996) find an immediate increase in sales upon the adoption of bonus incentive plan for sales consultants. Lazear (2000) shows a rise in worker productivity upon the introduction of a piece-rate compensation plan for line employees at an installation facility. Paarsch and Shearer (2000) estimate higher productivity for workers under piece-rate at a tree-planting company. Oettinger (2001) finds that higher commission piece rates for vendors results in higher performance. Yet, the impact of performance-based incentive for middle-class managers such as store managers has studied little. As store managers play an important role in retail context, understanding the value of incentive for store managers is vital for retail operations.

One exception that studied store manager's incentive scheme is DeHoratius and Raman (2007). Using a quasi-experiment setting where retailer changed the incentive scheme to store managers, DeHoratius and Raman (2007) studied the impact of incentive on store managers' effort allocation between sales and inventory shrinkage. They found that when incentive is changed in a way that sales became more important than inventory shrinkage, store managers exert more effort on sales than inventory shrinkage. Our work is different from DeHoratius and Raman (2007) in the following aspects. First, while DeHoratius and Raman (2007) have a one-way incentive, we have a two-way incentive. DeHoratius and Raman (2007) has a penalty if store has inventory shrinkage, thereby store managers try to minimize inventory shrinkage. Similarly, our setting has a penalty if store manager schedules too many labor than labor budget (i.e., overstaffing). In addition to a penalty, however, in our setting the bonus is contingent on sales (compared to its target) and understaffing could drive sales down. The district managers also monitor store managers' performance and discuss it when it's abnormal (e.g., severely understaffed or overstaffed). Hence, store managers try to be closer to the labor budget as both understaffing and overstaffing are inefficient. Second, while DeHoratius and Raman (2007) studied the impact of incentive on allocating efforts among multiple tasks, we estimate not only the impact of incentive change on labor outcomes, but also underlying decisions that led to those outcomes.

Our paper is also related to empirical papers on the subject of supply chain contracting in the OM literature although it is scarce. In the context of maintenance service contract, Guajardo et al. (2012) provide empirical evidence that the performance-based contract promotes reliability improvement more than resource-based contract does using a dataset from a major commercial aircraft engine manufacturer. Chan et al. (2014) study similar issue of how different contracting mechanisms influence maintenance outcomes, but they focus on maintenance service contracts for medical equipment. Our work is different in two aspects. First, they study firm-to-firm setting, where customers pick contract from alternative contract options (e.g., performance-based contract vs. resource-based contract), while we consider within firm setting, where the incentive scheme for individuals changed. Second, they focus on the impact of *contract* change on operational outcomes such as reliability whereas we examine the impact of *incentive* change on both operational outcomes as well as its underlying decisions driving those outcomes.

Finally, scholars have empirically studied a variety issues pertaining to workforce management in retail operations. Excellence in labor management has been identified as one of the key drivers of execution issues in store operations (Raman et al. 2001 and Fisher et al. 2007). Store labor has been found to increase profit (Ton and Huckman 2008, Kesavan et al. 2014, and Mani etal. 2015); sales (Kesavan et al. 2014, Tan and Netessine 2014, and Mani et al. 2015); basket value (Netessine et al. 2010); customer satisfaction (Ton and Huckman 2008); and service speed (Tan and Netessine 2014). Store labor also has been shown to decrease expenses (Kesavan et al. 2014) and phantom stockouts (Lee et al. 2017). Although numerous papers have examined labor issues in OM, there is no empirical research that has studied the role of incentives for store managers in their labor scheduling outcomes and their underlying operational decisions such as forecasting, labor planning, and execution.

3.2.2 Hypotheses

Store managers play a key role in managing labor in the retail industry. They are expected to hire the right person, train them, and schedule them to meet customer demand. Of these activities, labor scheduling decision is particularly very important for retailer. Too little labor can result in lost sales and poor service, while too much labor can increase store expenses. Studies show that understaffed stores can lose 8.56% of sales (Mani et al. 2015) while overstaffed stores can lead to significant expenses as labor related expenses account for over 85% of the controllable store expense in one of the studies (Kesavan et al. 2014). In our research setting, store managers may expend considerable endeavor to maximize their bonus by scheduling the right amount of labor as labor scheduling decision is included in the incentive as a penalty. Once store managers hit the sales target, if they spend on labor exceeding labor budget, the amount of eligible bonus is reduced. Also the performance-based incentive plan in our research setting changed from strong to weak in the middle of the study period. Under the strong incentive scheme, 50% of eligible bonus was paid based on store performance and the remaining 50% is paid based on retail chain performance, whereas under the weak incentive scheme, 100% of eligible bonus was paid based on retail chain performance. It is significant worsening because managers eligible to get the bonus due to higher store performance might not get the bonus if the retail organization does not perform well. Therefore, as growing evidence shows that performance-based incentive contracts lead to improved financial performance (Prendergast 1999, Banker et al. 2000, Lazear 2000, Paarsch and Shearer 2000, and Oettinger 2001), the weakening of the bonus-based incentive plan for store managers will reduce managers’ desire to exert effort to improve outcomes of overall labor scheduling decisions.

While the basic agency model of incentives predicts the negative impact of weakening the incentive on performance, store managers might continue to be motivated to keep their performance in a good shape as district managers provide (imperfect) monitoring and feedback of store managers' performance. Moreover, arguments from organization theory and cognitive psychology support this viewpoint. In a meta-analysis of over one hundred experimental studies, Deci et al. (1999) conclude that in most cases rewards have a negative effect on intrinsic motivation. Some recent economic literature addresses the "hidden cost of incentives", focusing on agents' extrinsic rewards may "crowd out" any intrinsic motivation the agent has to complete the tasks (see Benabou and Tirole (2003) and Gneezy and Rustichini (2001) among others). Camerer and Hogarth (1999) also show that incentives hurt in difficult judgment or decision tasks when incentives make decision makers self-conscious about. Scheduling labor is not an easy task as it contains the numerous nuances. For example, in a case study of the implementation of labor scheduling software at Belk Inc., Bernstein et al. (2014) find that over 70% of the schedules are over-written by store managers. Thus, the incentive for store managers might not work to improve labor scheduling decisions. This alternative explanation argues against any negative impacts from the weaker incentive. From these conflicting arguments, we propose the following hypothesis favorable to the agency theory:

Hypothesis 1 (Labor scheduling outcomes): *Store managers make overall labor scheduling decisions closer to the labor budget under the strong incentive scheme than under the weak incentive scheme.*

Economic theory suggests that performance-based incentive contracts can increase an organization’s overall performance (1) by attracting and retaining more productive employees (selection effect) and (2) by inducing employees to increase or to better allocate their effort (effort effect). The selection effect can occur because a performance-based incentive can act as a screening device that attracts and retains the most productive employees and discourages the least productive employees (Milgrom and Roberts 1992). Thus, the weakening of the bonus-based incentive plan for store managers at our research site could result in high ability managers leaving the firm because their expected future wages under the weak incentive scheme are lower than their prior wages and more low ability managers remaining and joining the firm. Several experimental studies have provided supporting evidence by documenting that high skilled individuals select performance-based incentives when given a choice between piece rate and fixed pay. For example, in a laboratory setting, Cadsby et al. (2007) document that the sorting effect of performance-based contracts was more important in explaining performance than the effort effect. Using a survey of sales managers, Lo et al. (2011) show that sales agents with greater selling ability are associated with more high-powered incentives. Lazear (2000) finds that performance pay results in the hiring of higher ability workers. Banker et al. (2000) document that incentive pay results in attracting of workers who are more productive than those that leave. Thus, evidence is accumulating on the sorting effect of financial incentives.

Sorting mechanism may play a key role when the firm hires many employees from a large pool of applicants and when the entrance barrier is low so that many people frequently join and leave. Changing store managers, however, is more difficult and rare compared to changing sales-force employees. As a result, in our setting, selection effect might not play a significant role to improve store managers' performance due to incentives. To test this idea, we propose the following hypothesis which quantifies the effort effect by excluding selection effect:

Hypothesis 2 (Effort): *Store managers make overall labor scheduling decisions closer to the labor budget under the strong incentive scheme than under the weak incentive scheme when same managers make labor decisions in both periods.*

The outcomes of overall labor scheduling decisions by store managers consist of three component tasks: forecasting, labor planning, and execution. In order to have the right amount of labor in the store, it is vital to have accurate sales forecasts, plan associates appropriately, and execute to the plan. The performance-based incentive scheme may drive store managers exert more effort for each of three tasks. With the strong performance-based incentive, store managers may forecast sales more accurately; plan labor closer to its target (i.e., labor budget); and exert execution effort to adjust their initial plan according to the new information available in the last mile such as demand pattern in the previous week. Accordingly, we propose the following three hypotheses:

Hypothesis 3 (Forecasting): *Store managers forecast sales closer to the actual sales under the strong incentive scheme than under the weak incentive scheme.*

Hypothesis 4 (Labor planning): *Store managers plan labor closer to the labor budget under the strong incentive scheme than under the weak incentive scheme.*

Hypothesis 5 (Execution): *Store managers exert more effort on adjusting initial plan according to the new information in the last mile under the strong incentive scheme than under the weak incentive scheme.*

Even though we hypothesize the impact of incentives on managers' effort for each task, we do not think the impact is identical. The performance-based incentive scheme may have differential impact on each task for labor scheduling decisions as each task may require different types of effort, have different sensitivity to effort, and have different task complexity. First, store managers need to exert different types of effort to fulfill each of three tasks in order to make the right labor scheduling decisions. For example, forecasting and labor planning may need for cognitive effort while execution may demand for relational effort. Forecasting requires managers to predict sales in the future under the uncertain environments which may need more cognitive-intensive effort. Once the forecasting is made, labor planning requires managers to take pre-existed constraints into account, which may demand cognitive effort (possibly less intensive than forecasting). In contrast, execution requires managers to keep good relationship with associates, which can help them find back-up associates when someone is absent or have on-call scheduling smoothly if needed. Hence the execution effort is more relational than cognitive. Second, each component task for labor scheduling decision might have different sensitivity on effort. In the lab setting, prior literature (Libby and Lipe 1992) shows that tasks have different sensitivity on effort and the impact of incentive on performance is contingent on this effort-sensitivity. In our context, it is possible that effort might have a different marginal effect on performance. For instance, by exerting the same amount of effort on labor planning and forecasting, the improvement in performance can be much higher for planning task than forecasting task due to uncertain nature.

Finally, different complexity in each task may affect the relation among incentives, effort, and performance. Broadly speaking, task complexity refers to the amount of attention or processing a task requires as well as the amount of structure and clarity the task provides. Thus, task complexity increases as the required amount of processing (i.e., task difficulty) increases and as the level of structure (i.e., task structure) decrease (Wood 1986 and Campbell 1988). Empirical findings from lab settings that use both cross-task and within-task definition of task complexity have shown that an increase in task complexity attenuates the positive effect of incentives on performance. For example, Pelham and Neter (1995) found that subjects performed better with incentives for the easy tasks while they did not perform well for the complex tasks. In our context, forecasting is a complex task as it is difficult and less structured. Even though managers get help from software tools, they still need to make a decision under uncertainty. Labor planning is a less complex task than forecasting as it is difficult but more structured. Labor planning typically requires managers follow pre-determined list of constraints. Execution is also a complex task as it is difficult and less structured. Managers should be capable of identifying signals from noise and responding accordingly. While both forecasting and execution are complex tasks, they require different type of effort. Forecasting is more cognitive-intensive whereas execution is more interpersonal relation intensive.

3.3 Data & Methodology

3.3.1 Data sources

Our data were obtained from a large U.S. based retailer with annual revenues exceeding $600 million. This company has over 100 stores in 22 states as of March 2017, with more than 5 million square feet of selling space. More than 5,300 associates work on a daily basis in its stores. As a discount department store, this retailer sells primarily women’s, men’s, and children’s apparel, along with accessories, fragrances, and home furnishing goods. Its target customers are 35- to 54-year-old working mothers. The study period was from February 2014 to December 2014, where the retailer had 75 stores.

We obtained the following data for the study period. First, we obtained weekly labor data from payroll database. It contains planned labor ($), actual labor ($) and labor budget ($). Labor budget is defined as the amount of labor the store should have had and calculated by a workforce software tool based on many factors such as actual sales, size of the store, labor productivity, marketing promotional activities, employee training, and cost of backend labor activities. Second, we possessed weekly actual and forecasted sales data. We have two different sales forecast for each week: (1) sales forecast from the software and (2) sales forecast adjusted by store managers. The retail chain invested on the software tool where it generates sales forecast based on historical data. Given the software forecast, store managers make their own sales forecast taking into account, among many other factors, the current trend and local knowledge. Having these two different sales forecasts help us separate managers’ forecast decisions from their reference (i.e., software forecast).

Now we provide details about labor scheduling process at this retailer and the incentive structure for store managers in the following subsections before defining variables and explaining empirical methodology.

3.3.2 Labor scheduling process

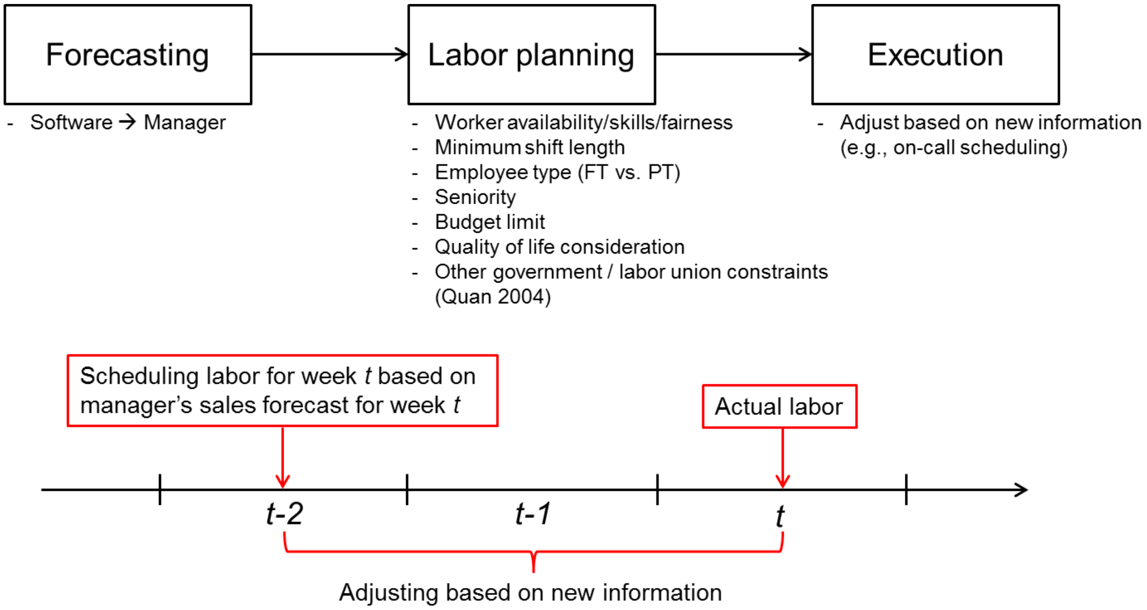
We illustrate store managers’ labor scheduling process at this retailer which consists of three main tasks: forecasting, labor planning, and execution. Figure 3.1 describes the labor scheduling process along with the time line for each decision. First, the store manager needs to accurately forecast sales. The retail chain provides store managers the sales forecast generated by the software tool taking into account mainly historical sales performance and seasonality. The retail chain ensures the software sales forecast to be available to store managers two weeks in advance. The store managers may update sales forecast based on their local knowledge such as store specific characteristics, sales trends, upcoming promotions, and new product introductions. It is also possible that store managers do not change software sales forecast if they believe the forecast is good enough or if they do not want to exert effort to improve forecast accuracy.

Second, the store manager needs to plan the appropriate number of associates to ensure coverage based on their sales forecast. Typically, retailers provide a scheduling software tool to help managers plan the right number of associates in the store. However, managers still need to exert effort to adjust plans recommended by the software tool as they often fail to capture numerous nuances of scheduling labor. For example, in a case study of the implementation of labor scheduling software at one of the large department stores in the U.S., Bernstein et al. (2014) find that over 70% of the schedules are over-written by store managers. Labor planning requires the manager to match the availability of workers with labor requirements after taking into account a number of pre-existed constraints such as minimum shift length, employee type (part-time vs. full-time), seniority, fairness, distribution of workers' skills, budget limit, quality of life consideration, and other government or labor union constraints (Quan 2004). These plans are generally posted a week or more in advance so that associates can plan accordingly.

Lastly, the store manager needs to execute plan flawlessly in order to have the right number of associates in the store. Since labor plans are typically generated a week or more in advance (as shown in the bottom of Figure 3.1), it is possible that store managers obtain new information in the last mile – in the time period between when plan is generated and executed – either about demand or supply side. Hence managers need to react to the new information and ensure that the right number of associates show up for work at the right time. For the demand side information, managers may observe demand signal and use on-call[5] scheduling. They call unscheduled associates to work in a few hours when they encounter an unexpectedly high demand whereas they let scheduled associates go home without working when they face an unexpectedly low demand. For the supply side information, managers may encounter different associates' availability. Associates might call off for several reasons including illness, either associate's or her child's, transportation issues, and conflicts with scheduling a second job (Lambert and Henly 2010). The store manager exerts effort to find back-up associates to maintain the right number of associates in the store.

[5] For recent debate on "on-call" shift, see "A push to give steadier shifts to part-timers" by Greenhouse S., The New York Times, July 15, 2014.

Figure 3.1: Labor Scheduling Process


Forecasting
Labor planning
Execution
- Software → Manager
- Worker availability/skills/fairness
- Minimum shift length
- Employee type (FT vs. PT)
- Seniority
- Budget limit
- Quality of life consideration
- Other government / labor union constraints (Quan 2004)
- Adjust based on new information (e.g., on-call scheduling)
Scheduling labor for week *t* based on manager's sales forecast for week *t*
Actual labor
t-2
t-1
t
Adjusting based on new information

3.3.3 Incentive scheme

We provide the details of incentive scheme for store managers used in this retail chain. This retail chain implemented a new incentive scheme at the beginning of second half of fiscal year 2014 (i.e., August). This quasi-experimental setting allows us identify the impact of incentive on store managers' overall outcomes of labor scheduling and its underlying decisions that led to those outcomes.

The store manager's bonus depends upon their store performance as well as the overall company performance. The store performance is measured based on the variance of actual sales from prior year's sales and the variance of actual labor to labor budget. Both variances are measured in percentages of sales. The overall company performance is determined based on whether the company meets its EBITDA (Earnings Before Interest, Tax, Depreciation and Amortization) goal or not. The way the company accounted for the store performance and company performance, and the frequency of payout changed during 2014 as explained below.

In the first two quarters of 2014, store managers were paid quarterly bonuses calculated in the following way. First, the year-over-year (YOY) sales growth rate was computed for each store in that quarter. Stores that had a decline in YOY sales growth did not receive any bonus. However, managers of stores with positive growth rate received 1% of their salary as a bonus for every 1% YOY growth up to 6% YOY growth. Beyond 6% YOY growth, the managers would get 1.3% of salary as an incremental bonus. If a store's YOY sales growth rate was 107%, then a manager whose annual salary is $100,000 would be eligible to receive $7,300 as a bonus.

Second, the deviation of actual labor from labor budget was calculated. Labor budget is calculated based on a workforce scheduling software that determines the amount of labor the store should have had based on many factors including actual sales it made in a given period, size of the store, labor productivity, marketing promotional activities, employee training, and cost of backend labor activities. In general, managers are assumed to spend too much expense on labor when actual labor is more than labor budget (i.e., overstaffing). The incentive scheme for store managers has a direct impact on bonus in case of overstaffing. For every 0.1% negative deviation between actual labor and labor budget (after divided by actual sales), the bonus calculated in the first step is reduced by 10%. From the same example above, if a store's labor budget was 9% of sales and the store's actual labor was 9.4% of sales, then the manager would be eligible to obtain only $4,380 as bonus for that quarter. On the contrary, when actual labor is less than labor budget (i.e., understaffing), managers are assumed to provide lower customer service. This might result in lost sales. Although understaffing does not tie with the bonus amount, it can have negative consequences to store managers. The district manager monitors store managers' performance on a daily basis and discusses it if the performance is abnormal such as understaffing. As district manager can influence on store managers' promotion, managers tend to avoid understaffing. Thus, both overstaffing and understaffing could have a negative impact on store managers' utility.

The managers were paid 50% of the quarterly bonus 45 days after the quarter ended and the remaining 50% was paid at the end of the year depending on whether the company met its EBITDA goal or not.

In the second half of 2014, the company made two changes. First, it decided to pay out bonuses annually as opposed to making any quarterly payments. Second, it made the manager's bonus contingent on whether the company met its goals. So, if the company did not make its goals, then a manager would not receive any payment even if they exceeded sales targets and their payroll was less than the target amount. For example, a store manager who was eligible for $4,380 as bonus every quarter would have received $8,760 as bonus in that year according to the previous plan paid at the end of the quarter even if the company did not meet its EBITDA target but now would not receive any bonus if the company fails to do so. Since the profit for this retailer had significantly dropped in the previous 5 quarters and became negative in the previous quarter, we argue that this new threshold significantly reduced, if not eliminated completely, financial incentives for managers. No other organizational change took place during the study period.

Monitoring of store manager performance

While the financial incentives are paid out quarterly or annually, monitoring of store performance was done at daily, weekly, and monthly levels. Store managers reported to district managers who monitored the performance of the stores closely. Daily reports of store performance were sent to district managers. Unless the performance was abnormal, district managers did not discuss the daily reports. However, it was common to have a call every week to review the previous week's performance. The sales performance and payroll performance were discussed in detail during these calls. Greater emphasis was placed in monthly review calls when the performance for the previous month was discussed.

3.3.4 Variables

For the tests of Hypotheses 1 and 2, total error in overall labor scheduling decisions is measured as absolute deviation between actual labor and labor budget in percentages of labor budget: $APE\ actual\ labor_{it} = \left|\frac{Actual\ labor_{it} - Labor\ budget_{it}}{Labor\ budget_{it}}\right|$ for store i and week t. This metric captures store managers' effort on scheduling labor to be closer to labor budget as both overstaffing and understaffing can cause negative consequences. Overstaffing can result in reducing the amount of eligible bonus whereas understaffing can draw district managers' attention as it can signal lower level of customer service. For the test of Hypotheses 3, 4, and 5, we further disaggregate total error in overall labor scheduling decisions into its component tasks: forecasting, labor planning, and execution. Forecasting error is measured as absolute deviation between manager sales forecast and actual sales in percentages of actual sales: $APE\ manager\ SF_{it} = \left|\frac{Manager\ SF_{it} - Actual\ sales_{it}}{Actual\ sales_{it}}\right|$. Labor planning error is measured as absolute deviation between planned labor and labor budget in percentages of labor budget: $APE\ planned\ labor_{it} = \left|\frac{Planned\ labor_{it} - Labor\ budget_{it}}{Labor\ budget_{it}}\right|$. To understand store managers' underlying decisions for each task, we also use weekly sales (actual and forecasts from manager and software) and labor (actual, planned, and budget) variables.

The strong incentive period consists of first two quarters in 2014 (fiscal week 1 to 26). The weak incentive period consists of the remaining weeks in 2014 until the end of the study period (fiscal week 27 to 47). $Weak\ incentive_{it}$ is a dummy equal to one when weak incentives are in place and zero when strong incentives are in place. Table 3.1 provides descriptive statistics for the entire period (Panel A), fore the strong incentive period (Panel B) and for the weak incentive period (Panel C). The typical store in the studied retail chain during the study period generates on average \$136,647 of sales per week and spends \$6,864 for labor per week, indicating labor productivity of \$19.91. That is, managers expect to have about \$20 of sales for every \$1 spending on labor. As the second period contains holiday seasons, the average sales and labor are higher than those during the first period. So, we exclude holiday seasons (i.e., after Thanksgiving holiday) from our analysis as a robustness check and obtain the same results.

Table 3.1: Summary Statistics

Panel A: All sample (N=3,248, 75 stores, week 1-47)

Variable	Mean	SD	Median	Min	Max
$Planned\ labor_{it}$ (\$)	6,847.13	2,004.39	6,324.5	4,102	14,987
$Actual\ labor_{it}$ (\$)	6,864.43	2,027.91	6,337.5	4,119.27	14,861
$Labor\ budget_{it}$ (\$)	6,945.98	2,108.04	6,397.5	4,073	16,655
$Actual\ sales_{it}$ (\$)	136,646.6	57,031.64	123,828	48,045	379,656
$Software\ SF_{it}$ (\$)	136,373.2	58,432.71	122,756.5	42,938	439,466
$Manager\ SF_{it}$ (\$)	135,390.7	57,949.81	122,277	37,540	422,796
$APE\ actual\ labor_{it}$ (%)	0.057	0.050	0.042	0.00	0.27
$APE\ planned\ labor_{it}$ (%)	0.056	0.056	0.038	0.00	0.62
$APE\ software\ SF_{it}$ (%)	0.098	0.093	0.074	0.00	0.82
$APE\ manager\ SF_{it}$ (%)	0.084	0.066	0.069	0.00	0.37

Panel B: Under the strong incentive scheme (N=1816, 75 stores, week 1-26)

$Planned\ labor_{it}$ (\$)	6,038.79	1,245.50	5,772	4,102	1,2139
$Actual\ labor_{it}$ (\$)	6,047.23	1,285.33	5,756.4	4,119.27	13,558.64
$Labor\ budget_{it}$ (\$)	6,064.65	1,201.55	5,797	4,073	11,503
$Actual\ sales_{it}$ (\$)	123,935	40,925.38	118,790.5	48,503	339,154
$Software\ SF_{it}$ (\$)	123,615.7	40,838.64	117,795	45,800	321,002
$Manager\ SF_{it}$ (\$)	122,749.1	40,772.31	116,543	40,836	365,166
$APE\ actual\ labor_{it}$ (%)	0.046	0.040	0.035	0.00	0.26
$APE\ planned\ labor_{it}$ (%)	0.050	0.049	0.035	0.00	0.33
$APE\ software\ SF_{it}$ (%)	0.109	0.102	0.085	0.00	0.82
$APE\ manager\ SF_{it}$ (%)	0.088	0.068	0.073	0.00	0.37

Panel C: Under the weak incentive scheme (*N*=1432, 75 stores, week 27-47)					
$Planned\ labor_{it}$ (\$)	7,872.22	2,295.16	7,307.5	4,211	14,987
$Actual\ labor_{it}$ (\$)	7,900.78	2,305.12	7,345	4,133.4	14,861
$Labor\ budget_{it}$ (\$)	8,063.64	2,452.83	7,366	4,299	16,655
$Actual\ sales_{it}$ (\$)	152,767	69,216.09	133,604.5	48,045	379,656
$Software\ SF_{it}$ (\$)	152,551.8	71,857.97	131,203.5	42,938	439,466
$Manager\ SF_{it}$ (\$)	151,422.1	71,073.15	131,545	37,540	422,796
$APE\ actual\ labor_{it}$ (%)	0.070	0.058	0.052	0.00	0.27
$APE\ planned\ labor_{it}$ (%)	0.063	0.063	0.042	0.00	0.62
$APE\ software\ SF_{it}$ (%)	0.085	0.077	0.066	0.00	0.58
$APE\ manager\ SF_{it}$ (%)	0.079	0.062	0.066	0.00	0.35

3.3.5 Empirical methodology

We first identify the effect of the change in incentives from strong to weak on outcomes of overall labor scheduling decisions by store managers. We estimate the accuracy of store managers' labor decisions ($APE\ actual\ labor_{it}$) of store i on week t using the following panel data regression:

$$APE\ actual\ labor_{it} = \beta_i + \gamma Weak\ incentive_{it} + \delta Z_{it} + \epsilon_{it}. \quad (3.1)$$

By including proper controls, we estimate the treatment of the treated, $\hat{\gamma}^{TTE}(x)$, using before-after data (*first differences*, Ichniowski and Shaw 2009) as follows:

$$\hat{\gamma}^{TTE}(x) = \bar{Y}^1_{Post} - \bar{Y}^1_{Pre} \quad (3.2)$$

where Y represents performance in labor scheduling decisions ($APE\ actual\ labor_{it}$). In words, the treatment of the treated effect is the difference in the conditional means of the treated group before (the strong incentive period), and after (the weak incentive period), the treatment (incentive change). Thus, the parameter of interest throughout is γ, namely the effect of the move from strong incentive to weak incentive on outcomes of store managers' overall labor scheduling decisions.

We include the following controls (x in (3.2.)). Store fixed effects (β_i) capture location-specific, time-invariant factors. For example, store managers may make a different labor schedule based on the local labor market condition. If it is very hard to find new employees, store managers may end up scheduling labor lower than their desired level resulting in larger error in outcomes of overall labor scheduling decisions. As the new weakened incentives are introduced simultaneously across all stores, it is not possible to control for week fixed effects. Instead, we control for time-varying factors at the store (Z_{it}) level.

Using a quasi-experimental setting, our identification of the effect of change in incentives from strong to weak on labor scheduling decisions arises from a comparison over time of the same stores. As initiating financial incentive for only a subset of stores is hard in practice, many papers have explored similar before-and-after change in incentive (for example, see Lazear 2000, Bandiera et al. 2005, and Anderson et al. 2010). The estimated effect γ is then biased downward to the extent that it captures factors that cause the accuracy in labor scheduling decisions to decrease in the second period regardless of the change in incentive schemes. For example, if all stores performed badly due to some unexpected demand shock in the second period, then our model would over-estimate the treatment effect. In this case we are unable to disentangle the effect of incentive from the effect of unexpected demand shock. We overcome this issue by controlling for the accuracy of software sales forecast relative to actual sales ($APE\ software\ SF_{it}$), defined as the absolute deviation between software sales forecast and actual sales in percentages of actual sales. This variable shows the relative performance of actual sales (i.e., surprising component) to their expectation (i.e., sales forecast). As this is the key consideration in our model, we investigate alternative ways to capture unobserved surprising component in demand in the robustness check.

Finally, the disturbance term ϵ_{it} captures unobservable determinants of managers' labor scheduling decisions at the store-week level. Labor decision observations within the same store are unlikely to be independent since stores face similar labor market conditions. We account for this by clustering standard errors at the store level in all regressions. To be specific, we estimate the variance-covariance matrix allowing *any* correlation pattern within store over time. This estimator for the variance-covariance matrix is given by:

$$W = (V'V)^{-1}\left(\sum_{j=1}^{N} u'_j u_j\right)(V'V)^{-1} \tag{3.3}$$

where N is the total number of stores, V is matrix of independent variables and u_j is defined for each store to be:

$$u_j = \sum_{t=1}^{T} e_{jt} v_{jt} \tag{3.4}$$

where e_{jt} is the estimated residual for store j at week t and v_{jt} is a row vector of dependent variables (including the constant). This methodology can overcome an issue of over-estimating t-statistics and significant levels because of serial correlation (Bertrand et al. 2004).

3.4 Results

3.4.1 Can incentive improve outcomes of overall labor scheduling decisions?

3.4.1.1 Model free evidence

Here we first provide the model free evidence. Our data show that financial incentives for store managers seem to help them make better overall labor scheduling decisions as store managers have actual labor closer to its target, i.e., labor budget, under the strong incentive scheme. Figure 3.2 provides the histogram of accuracy in weekly labor scheduling decisions, measured by $\frac{Actual\ labor_{it} - Labor\ budget_{it}}{Labor\ budget_{it}}$, under the strong incentive scheme (left figure) as well as that under the weak incentive scheme (right figure). The deviation between actual labor and labor budget is more concentrated on zero under the strong incentive period, indicating better outcomes of overall labor scheduling decisions.

Figure 3.2: Histogram of Labor Scheduling Decisions.

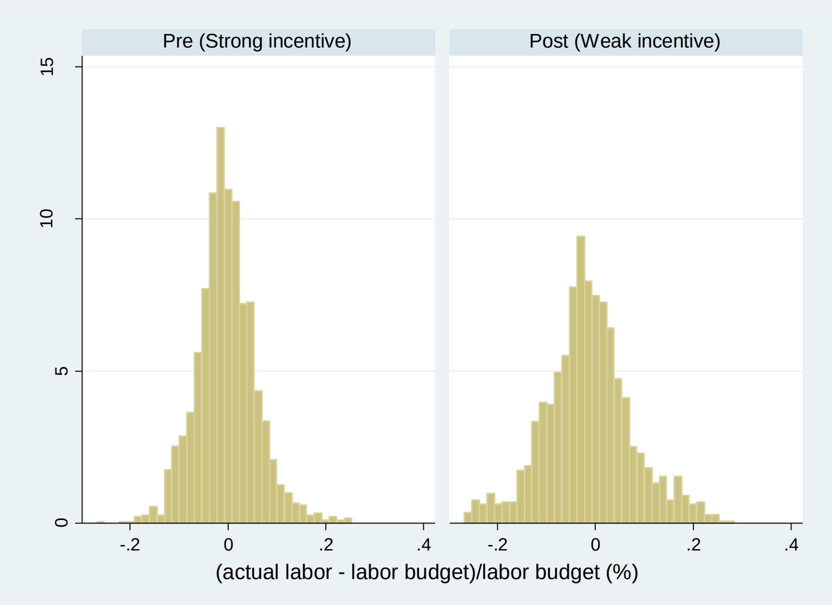


Figure 3.3 shows the mean of $APE\ actual\ labor_{it}$ over time. With the introduction of the weaker incentive the total error in outcomes of overall labor scheduling rose and remained at a higher level. Figure 3.4 provides similar observation. Each dot represents the coefficient of week fixed effect from the following regression:

$Log\ APE\ actual\ labor_{it} = \beta_i + \lambda_t + \delta Log\ APE\ software\ SF_{it} + u_{it}$. The pattern is very similar from the one observed in Figure 3.3, indicating worsening in labor scheduling outcoms after the transition from strong incentive to weak one.

Figure 3.3: Accuracy in Labor Scheduling Decisions Over Time.

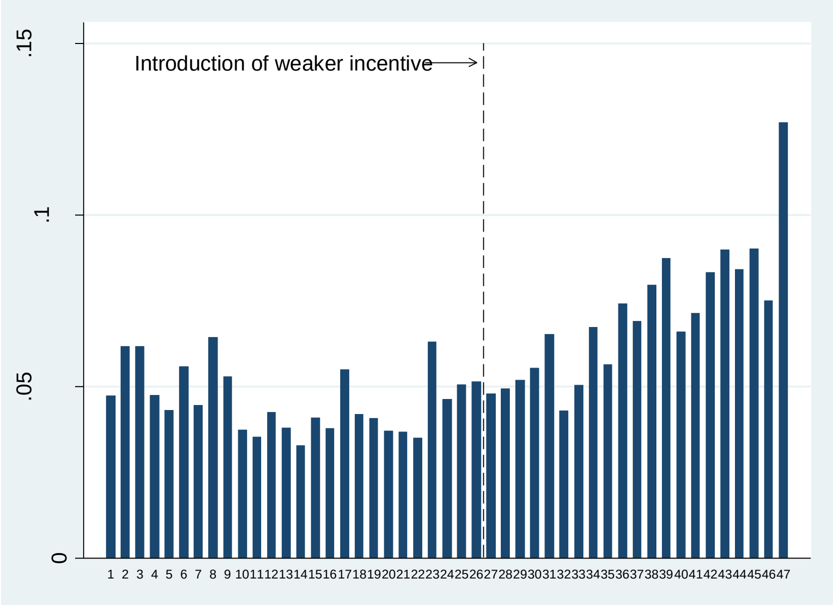
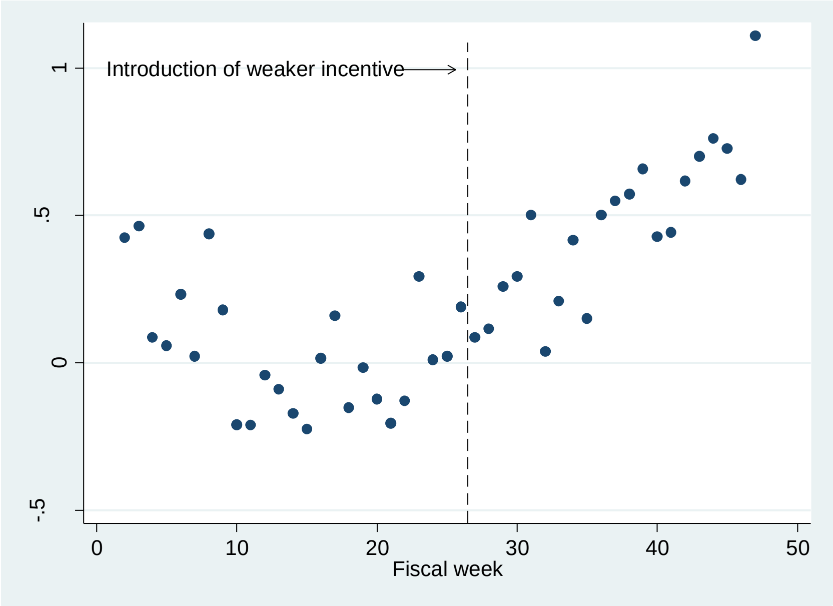
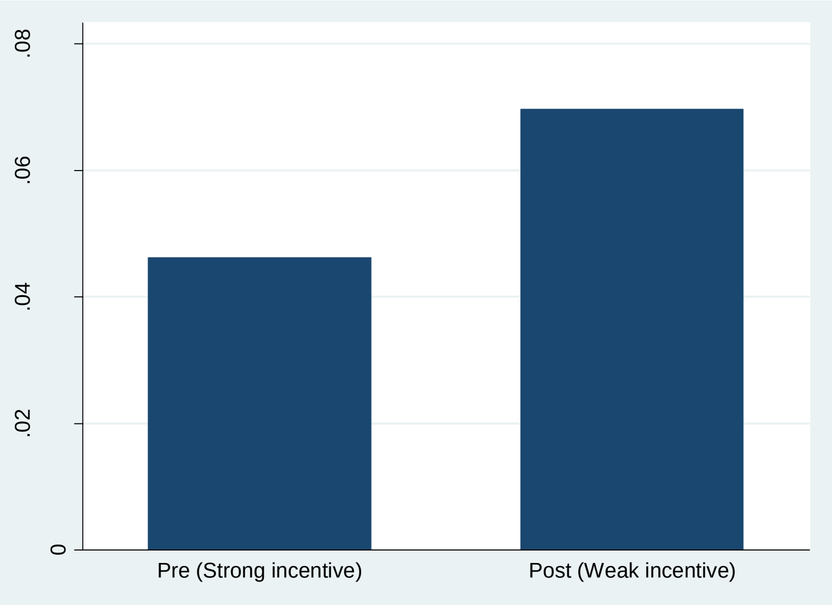


Figure 3.4: Accuracy in Labor Scheduling Decisions over Time (Alternative Method).



Finally, Figure 3.5 shows that the average of weekly $APE\ actual\ labor_{it}$ is much smaller under the strong incentive scheme compared to the weak incentive scheme (4.62% vs. 6.96%, *diff*=-2.34%, *p*<0.0000). It indicates 50.65% (-2.34/4.64) worsening in outcomes of overall labor scheduling due to incentive change. This model free evidence is consistent with Hypothesis 1.

Figure 3.5: Average of Accuracy in Labor Scheduling Decisions.



Although we find model free evidence consistent with our hypothesis, it may be confounded by other factors. For example, if all stores under-perform due to unexpected demand shock during the second half of the fiscal year 2014, data would show the same result. In addition, above observation is possibly driven by few extreme stores, but not other stores. We next show the results from econometrics model with proper controls to rule out these potential alternative explanations.

3.4.1.2 Main results

To examine whether the bonus-based incentive plan for store managers help them improve outcomes in their overall labor scheduling decisions, we first run a regression without any control (Column (1) in Table 3.2), clustering standard errors by store. Then, we subsequently add each control variable one at a time from column (2) to column (3) to deal with potential alternative explanations; hence, column (3) is a full model. In column (1), which is equivalent to the model free evidence (Figure 3.5), we observe significant impact of incentive change from strong to weak on the accuracy of overall labor scheduling decisions.

Table 3.2: Incentive Affects Overall Labor Scheduling Decisions

Dependent variable:	$APE\ actual\ labor_{it}$			
	(1) **Unconditional**	**(2)** **Store heterogeneity**	**(3)** **Unobservable surprise**	**(4)** **No selection**
$Weak\ incentive_{it}$	0.023*** (0.003)	0.024*** (0.003)	0.024*** (0.003)	0.023*** (0.004)
$APE\ software\ SF_{it}$			0.033*** (0.01)	0.042** (0.02)
Store effect	No	Yes	Yes	Yes
Observations	3248	3248	3248	1420
Adjusted R^2	0.0533	0.1063	0.1096	0.0963

Note. *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Standard error, clustered by store, is reported in parenthesis.

Column (2) controls for store fixed effects, so that only variation within a store over time is exploited. Controlling for store heterogeneity improves the fit of the model considerably. Store fixed effects double the explained variation in accuracy of store managers' overall labor scheduling decisions. The estimated effect of the change in incentives on total error in labor scheduling remains significant and of similar magnitude as in column (1).

Column (3) controls for other time-varying factor of store managers' overall labor scheduling decisions at the store level. We include a measure of surprising sales relative to the expectation to capture the unexpected demand shock over time. The estimated coefficient of $Weak\ incentive_{it}$ is still positive and significant (0.024, p<0.01), indicating 2.44% larger error under the weak incentive scheme compared to the strong incentive scheme period. This is a substantial difference as the average total error in scheduling labor is 5.65% in our data (Table 3.1). So, the weak incentives have resulted in 43.19% worsening of labor scheduling by managers. It shows that store managers make more accurate labor scheduling decisions under the strong incentive scheme, supporting Hypothesis 1. Our surprising finding is not so much that outcomes of scheduling labor become worse with weakening of pay-for-performance but by how much.

The coefficients' estimates of the control variables are in the expected direction. The accuracy of software sales forecast ($APE\ software\ SF_{it}$) is positively associated with the accuracy of labor scheduling decisions. As one of the important factors for store managers to schedule labor is the sales forecast, we expect the positive association between accuracy of software sales forecast and accuracy of labor scheduling decisions. Moreover, this variable plays a key role in our model to account for the unexpected demand shock by capturing a surprising portion in realized sales relative to the expectation. As expected, we find significant heterogeneity across stores.

3.4.1.3 Effort vs. selection

Our main finding based on the before-and-after comparison in the quasi-experimental setting has a potential alternative explanation due to selection in managers. As the retailer changes bonus-based incentive for store managers from strong to weak, some store managers may leave the organization looking for an opportunity with a better outside option. Those managers who decide to quit the job are likely to be high-ability managers as they have feasible outside options with a higher expected wage when they exert effort. As a result, the remaining managers are likely to be low-ability managers who cannot schedule labor as good as high-ability managers. In addition, the newly hired store managers after the high-ability managers quit the job might be low-ability managers as retailer's incentive is weak. This selection mechanism can potentially explain our main result.

In order to rule out this alternative explanation, we perform a regression analysis with only stores where managers stay without any turnover during the study period. We obtain annual manager turnover data. Among 75 stores, 33 stores (i.e., 44%) have no manager turnover in the fiscal year 2014. Column (4) in Table 3.2 shows the result with those 33 stores without manager change. We still find that the estimated coefficient of $Weak\ incentive_{it}$ is positive and significant (0.023, p<0.01), indicating 2.29% larger error in labor scheduling decision under the weak incentive scheme compared to the strong incentive scheme. Comparing to the average total error on overall labor scheduling decisions of 5.37% for these stores, the weak incentives have resulted in 42.64% worsening in outcomes of overall labor scheduling. Compare to the overall sample, we find that even after removing selection effect, we continue to observe significant effect of incentive on outcomes of labor scheduling decisions. This result supports Hypothesis 2 indicating that our main result is mainly due to effort effect, not due to the selection in managers.

To sum up, we provide evidence for the impact of financial incentives for store managers on outcomes in their overall labor scheduling decisions. We find that store managers make more accurate labor scheduling decisions under the strong incentive scheme. Unlike prior literature (Lazear 2000 and Banker et al. 2000) where studies the impact of incentive on individual workers' productivity and reports significant effect of sorting mechanism in addition to effort effect, we show that labor scheduling decisions become worse-off mainly due to decreasing effort. This indicates that incentives play a vital role in making managers exert more effort towards labor scheduling decisions.

3.4.1.4 Robustness checks

Table 3.3 presents a series of robustness checks. As the change in incentives from strong to weak occurs at the same time in all stores, the main concern in our analysis is to capture the unexpected demand shock which can confound our result. In the main model, we use the accuracy of software sales forecast relative to actual sales ($APE\ software\ SF_{it}$) to capture the surprising component of actual sales relative to their expectation (i.e., sales forecast). Alternatively we can calculate the surprising part in sales with respect to its expectation based on manager forecast (column (1), $APE\ manager\ SF_{it}$). The estimated effect of the change in incentives on total error in overall labor scheduling remains significant and of similar magnitude as in column (3) in Table 3.2, indicating that our result is robust with alternative control.

Table 3.3: Selective Robustness Checks

Dependent variable:			$APE\ actual\ labor_{it}$		
	(1) **Manager SF**	**(2)** **Software SF**	**(3)** **Manager SF**	**(4)** **Traffic**	**(5)** **DV by level**
$Weak\ incentive_{it}$	0.024*** (0.003)	0.024*** (0.003)	0.024*** (0.003)	0.025*** (0.003)	323.16*** (24.16)
$APE\ software\ SF_{it}$					409.79*** (90.54)
$APE\ manager\ SF_{it}$	0.064*** (0.01)				
$APE\ software\ SF_{it}$ (w.r.t. last year sales)		0.031** (0.01)			
$APE\ manager\ SF_{it}$ (w.r.t. last year sales)			0.015 (0.01)		
$Abs\ traffic\ growth_{it}$				0.004 (0.013)	
Store effect	Yes	Yes	Yes	Yes	Yes
Observations	3248	3248	3248	2922	3248
Adjusted R^2	0.1130	0.1081	0.1066	0.1131	0.1776

Note. *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Standard error, clustered by store, is reported in parenthesis.

Another concern is a reverse causality between actual labor and actual sales. One of our controls in the main model is $APE\ software\ SF_{it}$, which contains actual sales, and our dependent variable contains actual labor. The staffing level could drive sales; hence the concern of reverse causality arises. To mitigate this reverse causality concern (although it is not our key variable of interest and just one of controls), we replace actual sales by last year's sales to proxy unexpectedness relative to software sales forecast (column (2)) and manager sales forecast (column (3)). Again, the estimated effect of the change in incentives on total error in overall labor scheduling remains significant and of similar magnitude as in column (3) in Table 3.2, mitigating the reverse causality concern.

There is another way to control for the unexpectedness in demand using exogenous traffic information. This retail chain installed video cameras at store entrances to count the number of store visitors. All stores of this chain had this entrance camera during our study period. The third party video technology company responsible for the installed technology audited the data regularly by manually counting the number of visitors and comparing that count to the number from the automated sensors, ensuring the accuracy of at least 95%. Using traffic information, we measure absolute traffic growth, defined as the absolute difference between average daily traffic for a given week in 2014 and that in 2013 in percentages of the average daily traffic in 2013 ($Abs\ traffic\ growth_{it} = \left|\frac{Average\ daily\ traffic\ 2014_{it} - Average\ daily\ traffic\ 2013_{it}}{Average\ daily\ traffic\ 2013_{it}}\right|$), to account for the surprising component in demand. In column (4) of Table 3.3, we again find that weakening of incentive has resulted in significant worsening in labor scheduling outcomes (*p*<0.01) We note that the number of observations is reduced due to the paucity of data in last year's weekly store traffic.

In addition to capture the unexpectedness in demand using alternative ways, we also perform other robustness checks. First, we use alternative approach to address an issue of over-estimating t-statistics and significance levels because of serial correlation. In addition to the method that we used in the main model (Bertrand et al. 2004 call it as arbitrary variance-covariance matrix), Bertrand et al. (2004) proposed to use block bootstrapping which maintains the auto-correlation structure by keeping all the observations that belong to the same group (e.g., store) together. Using 200 bootstrap replications, we obtained very similar result. The clustered standard error for $Weak\ incentive_{it}$ is 0.0027 with t-statistics of 8.91 while the bootstrap standard error is 0.0028 with t-statistics of 8.65. This result indicates that our main results are not suffering an issue due to serial correlation.

Second, we examine the stability of the results using different performance measure. In the main model, we measure accuracy in store managers' overall labor scheduling decisions as absolute percentage error. We can alternatively use absolute error in level instead of percentage. All results are very similar to the ones obtained with percentage measure (column (5)). This shows that our substantive results are no artifact of the specific measure chosen in the analysis.

Third, we test the validity of our results without holiday seasons. As a large portion of sales (20-30% of annual sales, according to National Retail Federation) is generated in this short period of time, store operation is often very different during holiday seasons. So, we repeat our analysis after excluding holiday seasons (i.e., after Thanksgiving holiday). The main conclusions remain unchanged, indicating that our results are not affected by holiday seasons (available from the authors upon request).

Lastly, some variables in our analysis such as $APE\ actual\ labor_{it}$ and $APE\ planned\ labor_{it}$ are both non-negative (minimum achieved at zero), and have a large right-skew (signified by a maximum that is typically many times larger than the mean). For example, the average number of absolute percentage error for software sales forecast ($APE\ software\ SF_{it}$) in a week across stores is 9.8%. However, there are occasions when the software forecasting error is high. The observed maximum is 82%, which is more than eight times the average. As a robustness check, we transform those non-negative right-skewed variables into their natural logarithms ensuring that our hypothesis test statistics follow t-distribution. Our main conclusions remain qualitatively the same, indicating that our results are not driven by specific functional form (available from the authors upon request).

3.4.2 Mechanisms: Underlying decisions for scheduling labor

As explained in section 3.3.2, labor scheduling process consists of three component tasks. They are forecasting, labor planning and execution. In this section, we delve into these three component tasks and test for Hypotheses 3, 4, and 5 examining whether the financial incentives have an impact on each of them, driving outcomes of overall labor scheduling.

3.4.2.1 Forecasting

We first examine whether the financial incentive for store managers has an impact on their forecasting task for labor scheduling decisions (Hypothesis 3). We use the accuracy of manager sales forecast ($APE\ manager\ SF_{it}$) as a dependent variable. In column (1) of Table 3.4, we do not include the reference forecast from software for store managers. The estimated coefficient of $Weak\ incentive_{it}$ is negative and significant (-0.009, p<0.01). It looks like manager sales forecast is improved under the weak incentive scheme. However, when we add the accuracy in software sales forecast ($APE\ software\ SF_{it}$) in column (2), the estimated coefficient of $Weak\ incentive_{it}$ becomes insignificantly differentiable from zero. This indicates that managers adjust sales forecast provided by software tool and the financial incentive does not affect this adjustment. We do not find support for Hypothesis 3. This result is consistent with prior literature (Remus et al. 1998) which finds no evidence that incentives improve forecast accuracy in the lab setting.

Table 3.4: Mechanism: (1) Forecasting (No impact on accuracy)

Dependent variable:	$APE\ actual\ labor_{it}$	
	(1)	**(2)** **Relative to software SF**
$Weak\ incentive_{it}$	-0.009***	0.001
	(0.002)	(0.002)
$APE\ software\ SF_{it}$		0.404***
		(0.03)
Store effect	Yes	Yes
Observations	3248	3248
Adjusted R^2	0.0161	0.3339

Note. *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Standard error, clustered by store, is reported in parenthesis.

There are, at least, two explanations for the above finding. It is possible that store managers do not exert effort to improve forecasting regardless of the financial incentive. It is also possible that managers exert effort to improve forecast, but it does not convert into higher accuracy in forecasting. Prior literature (see Bonner and Sprinkle 2002 and references therein) from lab settings shows that the higher effort due to incentive may not translate into higher performance when the task is complex. We examine whether store managers exert more effort on forecasting under the strong incentive scheme than the weak incentive scheme by measuring the extent for managers to change sales forecast from software forecast (i.e., $\left|\frac{Manager\ SF_{it} - Software\ SF_{it}}{Actual\ sales_{it}}\right|$). This variable can capture managers' effort to improve forecast as they can simply take software forecast without changing it if they do not want to exert effort. We compare this metric before and after the incentive change.

Table 3.5 shows the result. In column (1), the estimated coefficient of $Weak\ incentive_{it}$ is negative and significant (-0.027, p<0.01), indicating that managers change forecast more from the one recommended by the software during strong incentive period compared to weak incentive period. Together with the previous finding of no incentive impact on forecasting accuracy, it appears that store managers exert more effort on improving forecasting, but this higher effort does not translate into higher accuracy in forecasting. This finding compliments previous literature on task complexity in a laboratory setting (Pelham and Neter 1995) where finds that subjects perform better with incentives for the easy version of a task while they do not for the complex version.

Table 3.5: Mechanism: (1) Forecasting (More effort)

Dependent variable:	$\left\lvert\frac{\boldsymbol{Manager\ SF_{it} - Software\ SF_{it}}}{\boldsymbol{Actual\ sales_{it}}}\right\rvert$	
	(1)	**(2)**
$Weak\ incentive_{it}$	-0.027*** (0.003)	-0.015*** (0.003)
$APE\ software\ SF_{it}$		0.491*** (0.034)
Store effect	Yes	Yes
Observations	3248	3248
Adjusted R^2	0.0795	0.3753

Note. *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Standard error, clustered by store, is reported in parenthesis.

There could be an alternative explanation. If the software forecast becomes more accurate in the second period, then managers do not need to modify it. This might confound our results. However, the forecasting algorithm of software has remained the same throughout the study period, mitigating the concern of systematic change in software forecast. Nevertheless, we control for the accuracy of software sales forecast ($APE\ software\ SF_{it}$) in column (2). We still find the same conclusion, indicating that the systematic change in software forecast is not an issue in our analysis.

The forecasting error comprises with bias and noise. We further examine the impact of financial incentive on the forecasting bias. In column (1) of Table 3.6, we find that store managers are in general optimistic in forecasting as the estimated coefficient of $Manager\ SF_{it}$ is 0.911 (p<0.01), indicating that actual sales are smaller than manager sales forecast. When we add $Weak\ incentive_{it}$ and the interaction between $Weak\ incentive_{it}$ and $Manager\ SF_{it}$ in column

(2), we find that store managers are more optimistic in forecasting sales under the strong incentive scheme compared to the weak incentive scheme. When managers expect sales to be \$1,000, the actual sales is on average \$809 under the strong incentive scheme whereas that is \$906 under the weak incentive scheme. It is conceivable as strong incentive ties managers' bonus to sales performance. We obtain similar results when we further add weekly store traffic in column (3) as a proxy for the store demand. This result indicates that managers' effort on forecasting due to incentive in fact increases optimistic bias. Together with previous findings, it further indicates that the incentive reduces noise in forecasting as the overall forecasting error remains the same.

Table 3.6: Mechanism: (1) Forecasting (Optimistic bias)

Dependent variable:		$\boldsymbol{Actual\ sales_{it}}$	
	(1)	**(2)**	**(3)**
$Manager\ SF_{it}$	0.911***	0.809***	0.546***
	(0.008)	(0.03)	(0.05)
$Manager\ SF_{it}$		0.097***	0.073***
$\times\ Weak\ incentive_{it}$		(0.03)	(0.03)
$Weak\ incentive_{it}$		-8636.08***	-6565.21**
		(2868.73)	(2821.72)
$Store\ traffic_{it}$			8.63***
			(1.76)
Store effect	Yes	Yes	Yes
Observations	3248	3248	3248
Adjusted R^2	0.0795	0.3753	0.9487

Note. *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Standard error, clustered by store, is reported in parenthesis.

To summarize, we find that under the strong incentive scheme store managers exert effort on forecasting by making changes on software forecast based on their local knowledge, but this effort does not convert into more accurate forecast. In addition, managers tend to have more optimistic bias on forecasting sales during the strong incentive period compared to the weak incentive period, indicating that stronger incentive might increase bias but reduce noise; hence, the forecasting error is unaffected.

3.4.2.2 Labor planning

Here we examine whether the financial incentive for store managers has an impact on their labor planning task for labor scheduling decisions (Hypothesis 4). To estimate the impact of incentive on labor planning error, we control for the accuracy in managers' sales forecasting decision as labor planning is based on manager forecast. This variable is also important to account for the unexpected demand shock. Table 3.7 contains the results.

Table 3.7: Mechanism: (2) Labor planning

Dependent variable:	$\boldsymbol{APE\ planned\ labor_{it}}$			
	Full	**Enough labor case**		
	(1)	**(2)** $\boldsymbol{Actual\ labor_{it} > Planned\ labor_{it}}$	**(3)** $\boldsymbol{Planned\ labor_{it} > Planned\ labor_{i,t-1}}$	**(4)** $\boldsymbol{Planned\ labor_{it} > Actual\ labor_{i,t-1}}$
$Weak\ incentive_{it}$	0.013***	0.019***	0.0099***	0.01***
	(0.003)	(0.005)	(0.003)	(0.003)
$APE\ manager\ SF_{it}$	0.103***	0.155***	0.078***	0.064***
	(0.017)	(0.025)	(0.02)	(0.021)
Store effect	Yes	Yes	Yes	Yes
Observations	3248	1647	1832	1804
Adjusted R^2	0.0773	0.1322	0.0416	0.0479

Note. *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Standard error, clustered by store, is reported in parenthesis.

In column (1), the estimated coefficient of $Weak\ incentive_{it}$ is 0.013 (p<0.01), indicating that the labor planning error is on average 1.3% larger under the weak incentive scheme compared to the strong incentive scheme. This is a substantial difference as the average labor planning error is 5.6% in our data (Table 3.1). So, the weak incentives have resulted in 23.21% worsening on labor planning decisions by managers. It indicates that the financial incentive for store managers improves their labor planning decisions, supporting Hypothesis 4.

It is possible, however, that store managers may face a shortage in labor from the local labor market. It may exacerbate the labor planning error as managers cannot find enough labor to plan at the desired level. If this happens more severely during the second period, then our result could be confounded with this labor supply issue. To mitigate this concern, we consider the case where managers had enough labor using three alternative ways. First, if actual labor appears to be higher than planned labor, it could indicate that managers had enough labor to plan when they make a planning decision. In column (2), we find that the impact of incentive is even stronger. The labor planning error is 1.9% (p<0.01) larger during the weak incentive period than the strong incentive period. Second, if planned labor to be higher than planned labor or actual labor last week, it could indicate that managers had enough labor to plan this week. In columns (3) and (4), we still find the similar results. These results show that our substantive results are no artifact of the supply shortage issue in labor.

3.4.2.3 Execution

Finally, we examine whether the financial incentive for store managers has an impact on their execution task for labor scheduling decisions (Hypothesis 5). As we illustrated in section 3.3.2, there are time gap between planned labor and actual labor. Since labor plans are typically generated a week or more in advance, it is possible that store managers obtain new information in the last mile (see Figure 3.1). Hence managers need to exert execution effort on reacting to the new information and ensuring that the right number of associates shows up for work at the right time. Two types of information might be available in the last mile. One is the demand information. Managers may observe demand signal and adjust staffing accordingly. They may call unscheduled associates to work in a few hours when they encounter an unexpectedly high demand whereas they may let scheduled associates go home without working when they face an unexpectedly low demand. Another one is the supply information. Managers may encounter different worker availability and need to adjust staffing accordingly. Associates might call off for several reasons including illness, either associate’s or her child’s, transportation issues, and conflicts with scheduling a second job (Lambert and Henly 2010). So, store manager needs to exert effort to find back-up associates according to the new information about associate availability to maintain the right number of associates in the store. We have data on demand side information while we do not possess supply side information. So we focus on the new information in the last mile on the demand side shock.

In the queueing theory, the number of employees is determined by so-called *square root staffing formula* (Feldman et al. 2008): $s = d + \alpha\sqrt{d}$, where s denotes the number of staffs, d denotes workload (i.e., demand), and α indicates the quality of service (i.e., service level). Accordingly, store managers might adjust their staffing based on two demand information: mean demand and volatility of demand. First, managers might increase staffing level when they see higher demand (than expectation) as s increases when d increases. We use store traffic as a proxy for the demand. We measure $Avg\ traffic\ change_{i,t-1}$ as a difference between average daily traffic for week $t-1$ in 2014 and that in 2013 to capture additional information in terms of average demand relative to last year’s mean demand. We use difference in mean demand relative to the last year’s mean demand because store managers typically set their expectation based on last year’s information. Thus, the difference measure could be a good proxy to capture the unexpected part in the mean demand. Second, managers might increase staffing level when they observe more volatile demand (than expectation). To keep the service level (α) constant, managers need more employees to satisfy high demand when demand fluctuates a lot. We measure $CV\ traffic\ change_{i,t-1}$ as a difference between coefficient of variation (CV, defined as standard deviation divided by mean times 100) of daily traffic for week $t - 1$ in 2014 and that in 2013 to capture additional information in terms of demand volatility relative to last year. For the same reason with mean demand, we use difference in demand volatility relative to the last year's volatility.

Table 3.8 shows the results. We use $Labor\ adjustment_{it}$, defined as $Actual\ labor_{it} - Planned\ labor_{it}$, as a dependent variable to capture the extent for managers to adjust labor from initially planned one. We only have $Avg\ traffic\ change_{i,t-1}$ in column (1). We do not find significant adjustment on actual labor according to the traffic growth information in the last week. However, when we add $Weak\ incentive_{it}$ and the interaction between $Weak\ incentive_{it}$ and $CV\ traffic\ change_{i,t-1}$ in column (2), we find that the average traffic change information in the last week is significantly associated with labor adjustment this week (0.374, *p*<0.01) under the strong incentive scheme. When there are 100 more store visitors in the last week, store managers on average increase labor adjustment this week by $37.4 under the strong incentive scheme. On the contrary, managers do not exert effort on adjusting initial labor plan according to the new available information on average traffic (0.374-0.552=-0.178, *p*=0.261) under the weak incentive scheme. We test other information regarding traffic variability in columns (3) and (4) and find similar results. In column (4), we find that the change of variability in traffic information in the last week is significantly associated with adjusting labor this week (3.78, *p*<0.01) under the strong incentive scheme. In contrast, managers do not exert effort on adjusting initial labor plan according to the new available information on variability in traffic (3.78-5.87=-2.07, *p*=0.282) under the weak incentive scheme. When we have both information on mean demand as well as volatility in demand in column (5), we obtain directionally similar results, but the interaction between weak incentive indicator and change of variability in traffic is marginally insignificant (*p*=0.131). These results provide support for Hypothesis 5 that the financial incentive for store managers make them exert more execution effort on adjusting in the last mile according to the new information in the demand side.

Table 3.8: Mechanism: (3) Execution (Last mile adjustment effort)

Dependent variable:	$Labor\ adjustment_{it}$				
	Average traffic change		**CV traffic change**		**Both**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**
$Planned\ labor_{it}$	-0.065***	-0.103***	-0.066***	-0.102***	-0.102***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
$Avg\ traffic\ change_{it}$	0.056	0.374***			0.39***
	(0.1)	(0.1)			(0.1)
$Weak\ incentive_{it}$		-0.552***			-0.536**
$\times\ Avg\ traffic\ change_{it}$		(0.18)			(0.22)
$CV\ traffic\ change_{it}$			1.07	3.78***	4.03***
			(1.12)	(1.34)	(3.05)
$Weak\ incentive_{it}$				-5.87**	-4.67$^+$
$\times\ CV\ traffic\ change_{it}$				(2.39)	(3.05)
$Weak\ incentive_{it}$		200.61***		231.27***	204.4***
		(31.32)		(32.69)	(31.31)
Store effect	Yes	Yes	Yes	Yes	Yes
Observations	2866	2866	2866	2866	2866
Adjusted R^2	0.9300	0.9325	0.9300	0.9322	0.9327

Note. *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Standard error, clustered by store, is reported in parenthesis. $^+$ indicates *p*=0.131

To summarize, we provide empirical evidence for the differential impact of financial incentive on each of three tasks in store managers' labor scheduling decisions: forecasting, labor planning, and execution. We find that incentives make store managers exert more effort on all of three tasks, but the consequences and procedures are different. The weakening of incentive does not affect the accuracy in forecasting although managers exert more effort on forecasting by changing it from the software forecast. In addition, managers are more optimistic on sales forecasting under the strong incentive scheme, indicating that incentive increases optimistic bias but decreases noise. The new weakened incentive has resulted in 23.21% more error in labor planning. Finally, managers exert effort on last mile adjustment under the strong incentive scheme whereas they do not exert such effort under the weak incentive scheme. These results might be because each task requires different types of effort, has different sensitivity to effort (Libby and Lipe 1992), and has different complexity (Pelham and Neter 1995).

3.4.2.4 Robustness check

We perform several robustness checks for the mechanism as follows. We examine the stability of the results for store without manager change to rule out selection mechanism. The high ability managers, who exert more effort, might quit the job due to weakened financial incentive and the low ability managers, who do not exert effort as high as the high ability managers do, might remain on duty resulting in reduction in overall effort level. By considering only 33 stores without any manager change during the study period, we still find qualitatively the same results for forecasting, labor planning, and execution (available from the authors upon request). It supports our argument based on change in store manager’s behavior towards exerting more effort due to incentive and mitigates the possibility of selection explanation.

Similar to the previous robustness checks for the main result on outcomes of overall labor scheduling, we also test the validity of our results without holiday seasons and alternative model specifications using log transformation. The main conclusions remain unchanged, indicating that our results are not affected by holiday seasons as well as our choice of model specification (available from the authors upon request).

3.5 Conclusion

Our analysis suggests that performance-based incentive for store managers improves outcomes of overall labor scheduling decisions by managers. Our estimates indicate that, by comparing the strong incentive to the weak one, there is a positive and statistically significant effect of incentive on the accuracy of overall labor scheduling decisions and that its magnitude is

42.64%. We also show that this result is mainly driven by change in effort than selection mechanism, indicating that incentives play a vital role in making managers exert more effort towards labor scheduling decisions. More importantly, we find that the financial incentive has differential impact on three underlying tasks (i.e., forecasting, labor planning, and execution) for overall labor scheduling decisions. This finding complements previous literature in lab settings where finds that the impact of incentive on performance is dependent upon types of task as each task requires different types of effort, has different sensitivity to effort (Libby and Lipe 1992) and complexity (Pelham and Neter 1995). The results obtained from our analysis provide a first step toward understanding the overall impact of incentive on operational outcomes (specifically, labor scheduling) as well as underlying decisions that lead to those outcomes.

The main results presented in this paper have a number of implications. First, by providing empirical evidence that financial incentives improve store managers' labor scheduling decisions, we show the importance of financial incentives for managers even when they are aided by underlying software. Many organizations have invested heavily in software-based solutions under the current trend of using data analytics techniques such as machine learning and artificial intelligence for decision making. The role of managers, however, is not diminished as they can potentially improve the decisions made by software. Hence, incentives still play a key role to motivate managers to expend the effort to make the right decisions.

Second, our result shows the importance of considering task-sensitivity in the effect of incentive on performance when organizations design it. By studying underlying decisions on the overall labor scheduling, we find that the financial incentive improves labor planning and execution decisions, but not forecasting decisions. It is possible that the incentive tied with labor planning or execution is more effective on improving performance than the one tied with forecasting. In general, organizations can maximize the effect of incentive by considering the differences across tasks such as sensitivity of effort and task complexity.

Although our analysis provides firm evidence that supports these conclusions, it is not free of limitations. As the change in incentives from strong to weak occurs at the same time in all stores, we do not have a comparison group. It is very hard, however, in practice to find cases where a financial incentive is effective for only a subset of managers. In fact numerous papers have investigated similar before-and-after change in incentive (for example, see Lazear 2000, Bandiera et al. 2005, and Anderson et al. 2010). The key challenge in this case is to control for the unexpected time-related shocks which can potentially confound the treatment effect. We overcome this issue by using various alternative controls to capture unexpectedness in demand. We consider a surprise component in sales relative to their expectation (defined by two different sales forecasts from software and manager), current expectation relative to last year's sales, and exogenous traffic growth rate. Another limitation arises due to data availability. We have relatively short period of data before and after the incentive change, which makes us focus on immediate impact of incentive. An interesting avenue of future research is to analyze the long-term effect of incentive on operational outcomes and their underlying decisions by incorporating, for example, learning effect. Despite the limitations, we believe that our study provides valuable insights to practitioners who are continuously striving to improve operational decisions using pay-for-performance incentives.

In this paper we present one of the few studies that empirically estimate the impact of performance-based incentive on operational outcomes as well as their underlying decisions. Although we cannot claim that the conclusions obtained in this study are applicable to all other firms, our findings are of interest, not only to the retail industry but also to all industries in which financial incentives are an important driver of improving operational decisions that leads to better firm performance.

CHAPTER 4
Determinants of Excess Inventory Announcement and Stock Market Reaction in the Retail Sector

4.1 Introduction

On May 28, 2015 Abercrombie and Fitch Co. (Ticker: ANF) announced that it had accumulated excess inventory so it was writing-down $26.9 million of inventory. Abercrombie's management blamed sluggish sales as the reason for its excess inventory. Abercrombie's management was not alone; blaming sluggish sales for the buildup of excess inventory is the most popular reason given by retail managers. Hendricks and Singhal (2009) note that 67.59% of firms that provided a reason for their excess inventory announcements during 1990 to 2002 claimed sluggish demand as the primary reason. If unexpected softening of demand is the primary reason for excess inventory announcements, then likelihood of retailers making such announcements should be independent of whether they are operationally competent or not.

Prima facie, there is theoretical basis behind such a claim. In the commonly used newsvendor model, left-over inventory occurs when realized demand is lower than the forecasted demand. So, excess inventory announcements may be an inevitable outcome of demand uncertainty. Furthermore, we observe that excess inventory announcements are made even by firms that are renowned for their operational excellence. Toyota, for example, announced that it cut-down production in order to reduce excess inventory in 2009 (Linebaugh 2009) while Wal-Mart announced that excess inventory had accumulated in its apparel products' category (McWilliams and Dodes 2007). So, it is not clear whether there is a link between operational competence and excess inventory announcements. Our first research question explores the presence of this link.

More importantly, it is unclear whether the stock market believes the retailers' explanation for excess inventory. In other words, does the stock market treat these announcements as the outcome of a bad draw from a random demand and penalize all excess inventory announcements similarly or is the stock market response to excess inventory announcements conditional on operational competence of the announcing firm? If it is the latter, then does the stock market penalize highly competent firms more for the disappointing announcement or does it penalize the low competent firms more as they are skeptical of their ability to manage the excess inventory and take appropriate actions to overcome the problem? Our second research question examines if the stock market reaction to such announcements is conditional on the operational competence of the announcing firm.

We study these two questions for the following reasons. Excess inventory announcements have generated considerable interest in the business press and academic research because of their large negative impact on firm performance (Hendricks and Singhal 2009). Since excess inventory announcements are the result of supply-demand mismatches, many researchers have emphasized operational improvements to reduce their occurrences (Fisher et al. 2000; Billington et al. 2002; Chopra and Sodhi 2004; Narayanan and Raman 2004; Tang 2006). For example, Fisher et al. (2000) show that four elements – forecasting; supply-chain speed; inventory planning; and accurate, available data – form the foundation of rocket science retailing to achieve "Right Product in the Right Place at the Right Time for the Right Price." Yet a direct link between operational performance and excess inventory announcements has not been established so far. Such a link, if present, needs to be established so we could demonstrate the value of operational improvements to managers.

In addition, we scrutinize the stock market response to excess inventory announcements to glean insights on whether the stock market holds a premium for operationally competent firms. If so, we would expect a sharper decline in stock market valuation when an operationally competent firm announces excess inventory as the market is likely to be more disappointed than if a less competent firm made such an announcement.

Recent literature has pointed to several disadvantages of using inventory turnover (IT) as a metric of inventory productivity (Gaur et al. 2005) so we use a different measure of operational competence to study these two questions. We measure operational competence using the total factor productivity (TFP) metric. We choose TFP as our primary measure of operational competence as it is a well-studied metric across multiple fields and there exists strong evidence that higher TFP is associated with better operations through various internal drivers such as management practices, employee knowledge, and information technology (see Syverson 2011 for a review). Using data from 245 stores of a UK retailer, Siebert and Zubanov (2010) find that different skills of store manager explain about 27-35% of variation in store-level TFP. In addition, we use inventory turnover, adjusted inventory turnover (AIT) metric from Gaur et al. (2005), and gross margin return on inventory (GMROI) as alternate metrics of operational competence.

We focus on the U.S. retail sector and collect data from three separate databases to perform our analysis. The first, obtained from Compustat through Wharton Research Data Services (WRDS), includes firm level financial data during 1962 to 2011 such as sales; operating income before depreciation and amortization; the number of employees; gross, property, plant, and equipment; and capital expenditure. We collect our sample from 1962 because Compustat data for earlier than 1962 have a serious selection bias (Fama and French 1992). We supplement this data with output and investment deflators from the Bureau of Economic Analysis (BEA) and annual average wage index from the Social Security Administration (SSA). Estimated firm level TFP, using Compustat data, is merged with two other datasets to investigate whether the firm level TFP is associated with excess inventory announcement. The second dataset, obtained from Factiva, collects 85 excess inventory announcements made by publicly traded U.S. retailers in the Wall Street Journal (WSJ) and Dow Jones News Service (DJNS) during 1990 to 2011. It allows us to use the event study methodology to examine the stock market's reaction on excess inventory announcement. In order to conduct the event study methodology, we used the last dataset which gathered information on daily stock prices from the Center for Research on Stock Prices (CRSP).

Our primary findings are as follows. We find support for our argument that operationally competent retailers have fewer excess inventory announcements. High TFP retailers, those in the top 90th percentile of TFP, are 3.53 times less likely to report excess inventory in the following year compared to low TFP retailers, those in the bottom 10th percentile. Therefore, we conclude that operationally competent retailers manage inventory better, thus having fewer excess inventory announcements than their less competent peers. We obtain consistent results even when we measure operational competence using IT but TFP appears to be a better predictor of excess inventory announcements than IT. These results cast doubt on managers' attribution of excess inventory to sluggish sales.

Our stock market response analysis yields the following results. Consistent with Hendricks and Singhal (2009), we find that excess inventory announcements are associated with

2.53% decline in stock market valuation over a two-day period (the day of the announcement and the day before the announcement) in the retail sector. However, we find that the market penalizes excess inventory announcements made by high TFP retailers much more severely than those made by low TFP retailers. Our analysis shows that an increase in one-year-lagged mean-adjusted TFP by one-standard-deviation is associated with -4.14% in the stock return over a two-day period. This result contrasts with prior findings that the stock market does not penalize high and low IT retailers differently (Hendricks and Singhal 2009). Our results suggest that when high TFP retailers announce excess inventory, the market might be more disappointed as it had higher expectations from these firms. Thus, we find that the stock market does not fully believe excess inventory announcements to be the results of a bad draw of a random demand, as claimed by retail managers.

Interestingly, we observe that over 47% of retailers had positive increase in stock price following the excess inventory announcement. In other words, there is considerable heterogeneity in the market's response to the announcement. To explain this anomaly, we also considered the information provided by the retailers for the reasons for excess inventory accumulation and actions that they have taken or plan to take to handle the excess inventory. Our analysis shows that providing follow-up actions moderates the negative association between firm's operational competence and abnormal stock returns due to excess inventory announcement. Retailers in the top 50th percentile of TFP face a -3.78% (median) decline in stock returns when they announce excess inventory without providing follow-up actions but face a 1.74% (median) increase in stock returns when they provide follow-up actions. We conjecture that the market trusts the competent companies to turn-around their operations when provided with a definite plan-of-action. In contrast, we do not find any difference in stock market reaction to whether retailers in the bottom 50th percentile of TFP provide follow-up actions or not to fixing the excess inventory problem.

This study makes several contributions to the literature. First, we undertake, to the best of our knowledge, the first empirical examination on determinants of excess inventory. Although excess inventory has been well studied, what drives excess inventory has been remained unclear. In this paper, we attempt to fill this gap by showing empirical evidence that operationally competent retailers have fewer excess inventory announcements than their less competent peers. This finding is important as it shows that excess inventory is not a random phenomenon merely driven by demand uncertainty, which is typically harder for managers to control, but by management practices.

Second, our paper contributes by expanding the literature on the relationship between operational performance and financial performance. Previous literature shows that excess inventory has a negative financial impact on stock market valuation (Hendricks and Singhal 2009). We find heterogeneity in such negative financial impact of excess inventory announcement on stock returns: Specifically, high TFP retailers are impacted more negatively by such an announcement compared to low TFP retailers.

Third, recent literature has shown that investments based on inventory turns yield higher abnormal stock market returns (Kesavan and Mani 2013; Alan et al. 2014). There are two possible explanations offered for this finding. The stock market might not be fully incorporating inventory information in pricing stocks (Kesavan et al. 2010; Kesavan and Mani 2013) or high IT retailers could be riskier than the low IT retailers as they have higher returns (Alan et al. 2014). Our study provides evidence for the former and against the latter. By contrasting with TFP, our study finds that the stock market reaction differs across high TFP and low TFP retailers but there is no significant difference in market reaction to announcements from high IT and low IT retailers. This result suggests that the stock market may not be distinguishing between high IT and low IT retailers, leading to abnormal returns in the future. In addition, we show that the low IT retailers are potentially riskier than high IT retailers because they have a greater likelihood of announcing excess inventory compared to high IT retailers.

The rest of the paper is organized as follows. The relevant literature is presented in §4.2. In §4.3, we develop our main hypotheses of the paper. Section 4.4 deals with the first main research question: the determinants of excess inventory announcement. It contains the estimation of the firm level TFP, data description with econometrics model, and results. The second research question about the market reaction to excess inventory announcement is investigated in §4.5. It includes event study methodology, model specification, and results. Section 4.6 concludes the paper.

4.2 Literature Review

We first look at the firm-level productivity literature. One of two common findings in the productivity literature (we discuss both findings in the U.S. retail sector in detail in the Appendix) is that large and persistent differences in estimated productivity levels across firms are ubiquitous. This finding has fueled diverse research questions in a number of fields: microeconomics, industrial organization, trade, and labor in economics literature (Syverson 2011); information technology (Brynjolfsson and Hitt 1996; Dewan and Kraemer 2000), organizational change (Bertschek and Kaiser 2004), and inventory (Lieberman and Demeester 1999) in business literature. The main focus of such productivity literature has been shifted from “what?” question to “why?” question since Bartelsman et al. (2000) first surveyed the micro-data productivity literature. Drivers of productivity, also, have been well documented and classified by internal and external factors (Syverson 2011). Especially, in the internal factors, previous literature shows that productivity is closely related to management practices (Bloom and Van Reenen 2007); managerial ability (Bertrand and Schoar 2003); worker's education (Ilmakunnas et al. 2004); and information technology (Brynjolfsson and Hitt 1996; Dewan and Kraemer 2000). These factors, especially information technology, have also been identified as drivers of good inventory management (e.g., Barua et al. 1995; Mukhopadhyay et al. 1995). However, the link between productivity and better inventory control has been sparsely studied. So, we examine if productive (i.e., high TFP) firms have fewer excess inventory announcements compared to low TFP firms.

We aware of one paper that shows the link between labor productivity and better inventory control. By studying 52 Japanese automotive companies, Lieberman and Demeester (1999) show that reducing work-in-process (WIP) inventory increases labor productivity. Unlike Lieberman and Demeester (1999), we do not study the antecedents of productivity but whether high productivity firms have fewer excess inventory announcements. We also consider retail sector, as opposed to manufacturing that was studied by Lieberman and Demeester (1999), and use total-factor-productivity (TFP) which is commonly regarded as a more appropriate measure of productivity rather than single-factor-productivity measures such as labor productivity.

There has been huge interest in empirically showing the link between inventory performance and financial performance measures. Rumyantsev and Netessine (2007), for instance, observe that inventory responsiveness is positively associated with profitability, but not inventory leanness by analyzing panel data for a sample of more than 700 firms. Some research has used stock returns as a financial performance metric. For example, Thomas and Zhang (2002) and Chen et al. (2005) investigate the relationship between long-term stock returns and levels of inventory turnover. They analyze the long-term stock returns based on annual data covering more than 20 years. In contrast, other researchers analyze the short-term stock returns around the time when firms announce some events by conducting the event study methodology (e.g., Hendricks and Singhal 2009; Thirumalai and Sinha 2011). For example, Hendricks and Singhal (2009) find significant negative impact of excess inventory announcement on stock returns. Based on a sample of 276 excess inventory announcements made during 1990-2002, they find -6.79% to -6.93% abnormal returns over a two-day period.

To answer our second research question we also examine the market reaction to excess inventory announcement by using the event study methodology. However, our work is different from Hendricks and Singhal (2009) in the following aspects. First, Hendricks and Singhal (2009) include growth prospect, firm size, and IT in their econometrics model to see the different market reaction to excess inventory announcement under a certain condition. On the contrary, in addition to growth rate, firm size, and IT, we also contain TFP as a proxy of firm's operational competence to examine whether the market reacts differently to excess inventory announcement based on the announcing firms' operational competence. Second, our analysis considers the moderating role of specific information (i.e., action) in the announcement to the adverse association between TFP and abnormal returns due to excess inventory announcement. Although Hendricks and Singhal (2009) explore the main effect of various actions and reasons on abnormal returns, they do not consider moderating impacts on the relationship between the announcing firm's operational competence and abnormal returns. Third, the sample used in this paper is different from that in Hendricks and Singhal (2009). While they use a sample of 276 excess inventory announcements from all sectors during 1990-2002, we use a sample of 85 excess inventory announcements from only retail sector during 1990-2011.

Our paper is also related to the body of literature that has shown that investments based on IT yield higher abnormal stock market returns (Kesavan and Mani 2013; Alan et al. 2014) although Gaur et al. (2005) documented several disadvantages of using IT as a metric of inventory productivity. There are two possible explanations offered for this finding. One is information-based argument: the stock market might not be fully incorporating inventory information in pricing stocks (Kesavan et al. 2010; Kesavan and Mani 2013). The other is risk-based argument: high IT retailers could be riskier than low IT retailers as they have higher returns (Alan et al. 2014). Such a risk-based argument, in the opposite direction, has been offered for the abnormal returns observed when investing in TFP based portfolios where low TFP retailers yield higher returns and are expected to be riskier compared to high TFP retailers (Imrohoroglu and Tuzel 2014). By studying the predictors of excess inventory announcement and the market response to the announcement, we provide evidence for the information-based argument and against the risk-based argument for investment in IT; but our paper supports for the risk-based argument for investment in TFP.

By using quarterly firm-level data of 183 U.S. retailers between 1985 and 2012, Kesavan et al. (2016) have shown that low IT retailers have larger abnormal inventory growth compared to high IT retailers. However, the paper does not use excess inventory announcements. Our paper shows that low IT retailers are more likely to announce the buildup of excess inventory in the following year, which is consistent with the finding of Kesavan et al. (2016).

4.3 Hypothesis Development

4.3.1 Determinants of excess inventory announcement

We derive our null hypothesis from theoretical operations management literature and observations in practice. In the commonly used newsvendor model, left-over inventory occurs when realized demand is lower than the forecasted demand. So, excess inventory announcements may be an inevitable outcome of demand uncertainty. This is supported by anecdotal evidence from practice where we observe majority of retailers claiming sluggish demand as the reason for excess inventory announcements. So, high and low operationally competent retailers have similar likelihoods of excess inventory announcements.

Alternatively, it is possible that operationally competent retailers have fewer excess inventory announcements. Excess inventory announcements are the result of supply-demand mismatches and many researchers have emphasized operational improvements to reduce their occurrences (Fisher et al. 2000; Billington et al. 2002; Chopra and Sodhi 2004; Narayanan and Raman 2004; Tang 2006). For example, Fisher et al. (2000) show that four elements – forecasting; supply-chain speed; inventory planning; and accurate, available data – form the foundation of rocket science retailing to achieve “Right Product in the Right Place at the Right Time for the Right Price.” However, a direct link between operational competence and excess inventory announcements has not been established so far. Accordingly, we develop our first hypothesis as follows:

Hypothesis 1 (H1) *High operationally competent retailers have fewer excess inventory announcements compared to low operationally competent retailers.*

4.3.2 Market response on excess inventory announcement

The negative impact of excess inventory announcement on the stock returns has been well documented (Hendricks and Singhal 2009). If the stock market considers excess inventory announcements to be outcomes of randomness in demand, then its negative reaction should not vary across high and low competent retailers. On the contrary, if the market believes such announcements to signal operational (in)competence of the announcing firm, then its reaction could vary based on the type of firm. We argue for two possible reactions based on the type of firm.

One may argue that the market penalizes excess inventory announcements made by high operationally competent retailers more severely than those made by their less competent peers. This is because the stock market on average could have a high expectation for operationally competent retailers and a low expectation for less competent retailers. So, when operationally competent retailers announce excess inventory, the market may be surprised by it, resulting in a sharp drop in the stock price. In contrast, when their less competent peers announce excess inventory, the market may have been expecting it so the stock price decline may not be too severe.

Alternatively, it is possible that the market might react more negatively to excess inventory announcements made by low competent retailers for the following reason. Excess inventory announcement only implies that the firm has excess inventory. The market may expect operationally competent retailers to recover from excess inventory problem sooner than their less competent peers. Hence the market does not penalize severely when high operationally competent retailers announce excess inventory while it may severely penalize when low operationally competent retailers announce an accumulation of excess inventory. Following above explanations, therefore, we develop two competing hypotheses as follows:

Hypothesis 2A (H2A) *The stock market reaction to excess inventory announcements will be more negative for high operationally competent retailers compared to low operationally competent retailers.*

Hypothesis 2B (H2B) *The stock market reaction to excess inventory announcements will be less negative for high operationally competent retailers compared to low operationally competent retailers.*

4.4 Determinants of Excess Inventory Announcement

To study the impact of operational competence on excess inventory announcement, we first explain why we measure operational competence by TFP, and then how to estimate it in the following section.

We choose TFP as our primary measure of operational competence because it is a well-studied metric across multiple fields and there exists strong evidence that higher TFP is associated with better operations through various internal drivers such as management practices, employee knowledge, and information technology (Syverson 2011). Using data from 245 stores of a UK retailer, Siebert and Zubanov (2010) find that different skills of store manager explain about 27-35% of variation in store-level TFP. As the firm-level TFP is a sum of store-level TFP, we expect to capture retailer's operational competence such as a capability of store management team by considering the firm-level TFP. In addition, we use inventory turnover, adjusted inventory turnover (AIT) metric from Gaur et al. (2005), and gross margin return on inventory (GMROI) as alternate metrics of operational competence.

4.4.1 Estimation of TFP

Total factor productivity (TFP) is a measure of overall efficiency in operations: how much output (e.g., revenue) is obtained from a given set of inputs such as capital, labor, and intermediate materials. It is also called multifactor productivity, which is conceptually opposite to single-factor productivity. Unlike single-factor productivity, TFP does not suffer from the different intensity problem of excluded input usage. For example, suppose that we estimate labor productivity for firm A and B, where firm A is a high technology firm while firm B is not. In this case, although firm A and B have exactly same labor level, labor productivity is affected by both labor and different intensity of excluded input such as technology because the importance of labor for firm A is relatively smaller than that of firm B.

The most popular way to measure the firm level TFP is to get a residual, which is the deviation between observed output and predicted output, from the Cobb-Douglas production function estimated by ordinary least square (OLS). However, such estimation may suffer from simultaneity bias (Marschak and Andrews 1944). Olley and Pakes (1996, abbreviated as OP) and Levinsohn and Petrin (2003, abbreviated as LP) introduce methods to control such bias so that allowing us to estimate consistent parameters of the production function, and thus obtain reliable TFP estimates.

The main difference between two approaches is that OP use investment while LP use intermediate inputs like energy and materials used in operations to control for correlation between inputs (i.e., explanatory variables) and the unobserved productivity shock (i.e., error term). Investment is a good proxy for the firm which has positive investment, but as LP pointed out there is a "zero investment" problem. In this case, investment proxy may not smoothly respond to the productivity shock, violating the consistency condition. Therefore, we use LP as a main method to estimate the firm level TFP although we report all result with OP as a robustness check.

We estimate the production function based on labor and physical capital as two main inputs. The production function is given by:

$$y_{it} = \beta_0 + \beta_l l_{it} + \beta_k k_{it} + \epsilon_{it} \tag{4.1}$$

$$\epsilon_{it} = \Omega_{it} + \eta_{it} \tag{4.2}$$

where y_{it} is the log of value added for firm *i* in period *t* (We use value added, total output – intermediate materials, as an output. Therefore, we exclude intermediate materials from a set of inputs.); l_{it} and k_{it} are log of labor and capital inputs; and ϵ_{it} is an error term. The error term is a sum of two errors: Ω_{it}, the TFP, and η_{it}, the an unexpected idiosyncratic productivity shock. Both LP and OP assume that TFP, Ω_{it}, is observed by the decision maker in the firm before the firm makes its input decisions, which gives rise to the simultaneity problem. That is, inputs are correlated with the realization of the TFP. Specifically, labor, l_{it}, is the only variable input, thus its value can be affected by current TFP, Ω_{it}, while capital, k_{it}, is a fixed input at time *t*, and its value is only affected by the conditional distribution of Ω_{it} at time *t* – 1. Therefore, Ω_{it} is a state variable which has an impact on firms' decision making. For example, firms that observe a positive productivity shock in period *t* will consume more intermediate inputs, m_{it}, and hire more labor, l_{it}, in that period. Note that OP uses investment instead of intermediate inputs with the similar logic.

Demand for the intermediate input, m_{it}, is assumed to depend on the firm's state variables, Ω_{it} and k_{it}:

$$m_{it} = m(\Omega_{it}, k_{it}) \tag{4.3}$$

This intermediate input equation is based on the assumption that future TFP is strictly increasing in current TFP, Ω_{it}, so firms which observe a positive productivity shock in period *t* will require more intermediate inputs in that period, for any capital, k_{it}. This assumption is supported by the fact that TFP is not fleeting. For example, autoregressive coefficient of TFP is 0.64 in the U.S. retail sector (in the Appendix). Since m_{it} is strictly positive, we can write the inverse function for the unobserved productivity shock, Ω_{it}, as

$$\Omega_{it} = m^{-1}(m_{it}, k_{it}) = h(m_{it}, k_{it}) \tag{4.4}$$

which is strictly increasing in m_{it}. The unobservable TFP is now expressed solely as a function of two observed inputs, m_{it} and k_{it}.

LP further assume that TFP is governed by a first-order Markov process

$$\Omega_{it} = \mathbb{E}\left[\Omega_{it}\middle|\Omega_{i,t-1}\right] + \xi_{it} \tag{4.5}$$

where ξ_{it} is an innovation to TFP that is uncorrelated with k_{it}, but not necessarily with l_{it}; this is one source of the simultaneity bias.

Using equation (1), (2), and (4), we can obtain

$$y_{it} = \beta_l l_{it} + \phi_{it}(m_{it}, k_{it}) + \eta_{it} \tag{4.6}$$

where $\phi_{it}(m_{it}, k_{it}) = \beta_0 + \beta_k k_{it} + \ h(m_{it}, k_{it})$, and approximate ϕ_{it} with a third-order polynomial series in capital and intermediate inputs. Approximation with a higher order polynomial does not significantly change the results. This first stage estimation results in an estimate for $\hat{\beta_l}$ which controls for the simultaneity problem.

However, the first stage does not identify β_k. To do that, we begin by computing the estimated value for ϕ_{it} using

$$\widehat{\phi_{\iota t}} = \widehat{y_{\iota t}} - \widehat{\beta_l} l_{it} = \widehat{\delta_0} + \sum_{l=0}^{3}\sum_{j=0}^{3-l} \widehat{\delta_{l\jmath}}\, k_{it}^{l} m_{it}^{j} - \widehat{\beta_l} l_{it} \tag{4.7}$$

For any candidate value $\widetilde{\beta_k}$, we can compute (up to a scalar constant) a prediction for Ω_{it} for all periods t using

$$\widehat{\Omega_{\iota t}} = \widehat{\phi_{\iota t}} - \widetilde{\beta_k} k_{it} \tag{4.8}$$

Using these values, a consistent (nonparametric) approximation to $\mathbb{E}\left[\Omega_{it}\middle|\Omega_{i,t-1}\right]$ is given by the predicted values from the regression

$$\widehat{\Omega_{\iota t}} = \gamma_0 + \gamma_1 \Omega_{i,t-1} + \gamma_2 \Omega_{i,t-1}^2 + \gamma_3 \Omega_{i,t-1}^3 + \upsilon_{it} \tag{4.9}$$

which LP call $\mathbb{E}\big[\widehat{\Omega_{\iota t}\big|\Omega_{\iota,t-1}}\big]$.

Given $\widehat{\beta_l}$, $\widetilde{\beta_k}$, and $\mathbb{E}\left[\widehat{\Omega_{\iota t}\middle|\Omega_{\iota,t-1}}\right]$, LP write the sample residual of the production function as

$$\widehat{\xi_{\iota t}+\eta_{\iota t}} = y_{it} - \widehat{\beta_l}l_{it} - \widetilde{\beta_k}k_{it} - \mathbb{E}\left[\widehat{\Omega_{\iota t}|\Omega_{\iota,t-1}}\right] \tag{4.10}$$

We can estimate $\widehat{\beta_k}$ by solving

$$\min_{\widetilde{\beta_k}} \sum_i \sum_t \Big(y_{it} - \widehat{\beta_l}l_{it} - \widetilde{\beta_k}k_{it} - \mathbb{E}\left[\widehat{\Omega_{\iota t}|\Omega_{\iota,t-1}}\right]\Big)^2 \tag{4.11}$$

Finally, the TFP is estimated by:

$$TFP(LP)_{it} = \exp(y_{it} - \widehat{\beta_0} - \widehat{\beta_l}l_{it} - \widehat{\beta_k}k_{it}) \tag{4.12}$$

By using all data available up until that year, we estimate the production function parameters every year to eliminate a potential look-ahead bias in the TFP estimates. We calculate the firm level TFP for each year using that year's data (y_{it}, l_{it}, k_{it}, and m_{it}) and the corresponding production function parameters for that year ($\widehat{\beta_l}$ and $\widehat{\beta_k}$). For example, to calculate TFP values for 2010, we use all data up to and including 2010 to estimate parameters and then use the 2010's data to calculate TFP value for each firm. These values would then be used to predict the likelihood of announcing excess inventory in 2011.

Our estimation of TFP is consistent with prior literature that has shown large and persistent differences in productivity (Syverson 2011). For the large differences in TFP, we find a significant dispersion in the firm-level TFPs in the U.S. retail sector. The overall 90-10 TFP ratio in the U.S. retail sector is 2.54 (2.04 in OP). Prior study has shown the similar differences. For example, Syverson (2004) finds the average 90-10 TFP ratio of 1.92 in the U.S. manufacturing sector and Imrohoroglu and Tuzel (2014) report the same measure of 1.8 in all U.S. sectors. For the persistent differences in TFP, we find the autoregressive coefficient of 0.64

(0.73 in OP) in the retail sector. It is robust with productivity literature, which ranges autoregressive coefficients between 0.6 and 0.8 (e.g., Abraham and White 2006; Foster et al. 2008). We provide details in the Appendix.

4.4.2 Data and variables

We use three different databases to test our hypotheses.

4.4.2.1 Firm level TFP

The first main data source for estimating TFP (explained above in §4.4.1) is Standard and Poor's Compustat from Wharton Research Data Services (WRDS). We use the Compustat fundamental annual data from 1962 to 2011. Our sample for production function estimation is comprised of all U.S. retail firms by SIC code between 5200 and 5999. The sample is an unbalanced panel with approximately 1,773 distinct retail firms; the total number of firm-year observations is approximately 18,281. This is only for estimating TFP. The sample size reduces after we merge the Compustat data with other datasets such as Factiva and daily stock price to test our hypotheses.

The key variables for estimating the firm level TFP are the value added (y_{it}), employment (l_{it}), and physical capital (k_{it}). Firm level financial data is supplemented with three additional data: 1) price index for Gross Domestic Product (GDP) as deflator for the value added; 2) price index for private fixed investment as deflator for capital, both from the Bureau of Economic Analysis (BEA); and 3) national average wage index from the Social Security Administration (SSA).

We use revenue-based measure of TFP, which is highly correlated with physical quantity-based measure (Foster et al. 2008), so value added (y_{it}) is calculated as *Sales* – *Materials*, deflated by the GDP price index. *Sales* is net sales (SALE in Compustat), which is gross sales minus cash discounts, returned sales, etc. *Materials* (m_{it}) is measured as *total expenses* minus *labor expenses* where total expenses is approximated as *Sales* minus *Operating Income Before Depreciation and Amortization* (OIBDP in Compustat) and labor expenses is calculated by multiplying the number of employees (EMP in Compustat) by average wage index from the SSA. Thus our value added definition is proxied by *Operating Income Before Depreciation and Amortization* plus *labor expenses*.

The labor input (l_{it}) is computed by the number of employees (EMP in Compustat). The capital stock (k_{it}) is measured by *gross property, plant, and equipment* (PPEGT in Compustat) and deflated by the price index for private fixed investment (Brynjolfsson and Hitt 2003).

4.4.2.2 Excess inventory announcement

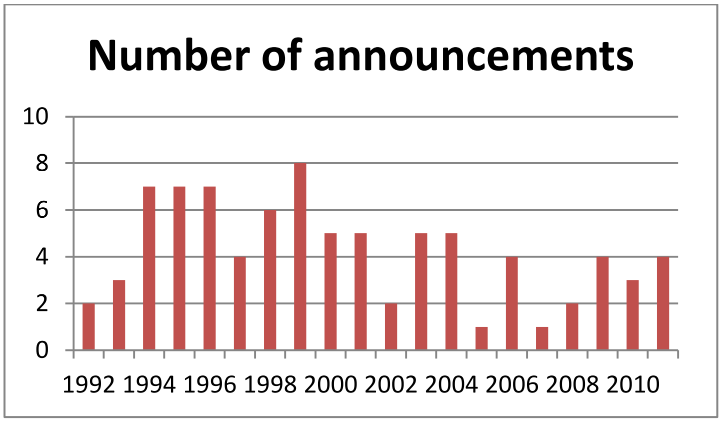
The second data set, obtained from Factiva, collects excess inventory announcements made by publicly traded U.S. retailers in the *Wall Street Journal* (WSJ) and *Dow Jones News Service* (DJNS) during 1990 to 2011. We use a set of keywords to search for announcements regarding excess inventory based on Hendricks and Singhal (2009). Specifically, we collect announcements which have the word "inventory or inventories" within five words of terms such as "obsolete, excess, glut, buildup, reduce, bloated, charge, adjust, loss, write-off, write-down, liquidate, accumulate, or revalue." More details about searching algorithm which used in this paper are available on the Online Supplement of Hendricks and Singhal (2009). We obtain exactly the same number of announcements, 4612 (as reported in Hendricks and Singhal 2009), when we follow their algorithm to check consistency.

The final sample consists of 85 excess inventory announcements (73 unique firms) for retail firms in the U.S. between 1990 and 2011. We found 95 announcements initially, but after merging with other datasets, we end up having 85 announcements because of missing data, bankruptcy, etc. Here are some examples of announcement:

- Best Buy Co. has started 12-month no-interest financing specials to trim PC inventory which will become obsolete when new technology arrives. (*WSJ*, 19 December 1996)
- Gap Inc. cut prices to spark slow sales and reduce inventory which results in flat gross profit margins. (*WSJ*, 5 May 2000)
- Wal-Mart's inventories jumped 10.3% in the fiscal first quarter, ended April 30, to $35.2 billion from a year earlier, driven by unsold apparel, home decor and outdoor products. (*WSJ*, 21 May 2007)

Figure 4.1 shows the number of announcements by year. Nearly 52% of announcements are made during 1990’s and 48% are during 2000’s. Further investigation of the timing of announcement indicates that more announcements are reported during the first and the fourth quarters (i.e., from October to March), with 32.94%, 18.82%, 18.82%, and 29.41% in the first, second, third, and fourth quarters, respectively. This result differs from an observation made by Hendricks and Singhal (2009) which has nearly equally distributed across the four quarters. Different fiscal year across industries may cause this variation. Considering retailer’s fiscal year, which ends in general at the end of January and earnings report date is couple of months later, our sample represents retailers’ tendency to announce excess inventory before earnings report date. Based on the National Bureau of Economic Research, our sample embraces two recession periods: 1) March 2001 to November 2001; and 2) December 2007 to June 2009. Our data show that 9.41% of our announcements (8 out of 85 announcements) are made during the recession period. We test the effect of recession on our main models as a robustness check, but we do not find any statistically significant difference between recession and non-recession periods.

Figure 4.1: Distribution of Excess Inventory Announcements by Year



Note: This time period includes two recession periods according to National Bureau of Economics Research: 1) March 2001 to November 2001 and 2) December 2007 to June 2009. The number of sample in each period is 2 and 6, respectively.

Our sample consists of firms in retail sector from eight different industries based on two-digit standard industrial classification (SIC) codes between 5200 and 5999. Table 4.1-a illustrates the distribution of excess inventory announcements by industry. Since our sample does not have any excess inventory announcement in SIC 58, we exclude it from our analysis. Miscellaneous retail (SIC 59) represents 32.94% of the sample while Food stores (SIC 54) represents 1.18% of the sample.

Some announcements provide detail information including the reasons of accumulating excess inventory (Table 4.1-b) and the actions that a firm has taken or plans to take to deal with excess inventory (Table 4.1-c). 35.42% of the sample mentions sluggish sales and 5.21% of the sample mentions obsolete and discontinued inventory as a main reason of excess inventory buildup. Interestingly, many firms (35.42%) blame external factors such as sluggish sales as a main reason of building excess inventory whereas very few firms (4.71%) mention internal factors such as internal inefficiency and poor execution as a main reason of excess inventory, suggesting that the randomness in demand is the key driver of excess inventory, but not operational incompetence. For action information, 38.61% of firms states markdowns and promotions and 23.76% of firms states inventory write-down as their primary actions to cope with excess inventory problem. As action is verifiable by the market, but reason is not, the stock market may respond differently by action but not reason. We explore this idea in §4.5.4.

Table 4.1-a: Distribution of Excess Inventory Announcements by Industry

SIC 2 digit		# of announcement	% of sample
52	Building materials & garden supplies	4	4.71%
53	General merchandise stores	12	14.12%
54	Food stores	1	1.18%
55	Automotive dealers & service stations	5	5.88%
56	Apparel and Accessory stores	21	24.71%
57	Furniture and home furnishings stores	14	16.47%
59	Miscellaneous retail	28	32.94%
		85	100.00%

Note: Since our sample does not have any excess inventory announcement in SIC 58, Eating and drinking places, we exclude it in the analysis.

Table 4.1-b: Reasons of Excess Inventory Announcement

Reasons	# of announcement	% of sample
Sluggish sales	34	35.42%
Obsolete and discontinued inventory	5	5.21%
Other reasons	20	20.83%
No reasons given	37	38.54%
	96	100.00%

Note: 74 and 11 announcements provide single and multiple reasons, respectively.

Table 4.1-c: Follow-up Actions to Excess Inventory Announcement

Actions	# of announcement	% of sample
Inventory write down	24	23.76%
Markdowns and promotions	39	38.61%
Other actions	23	22.77%
No action indicated	15	14.85%
	101	100.00%

Note: 69 and 16 announcements provide single and multiple reasons, respectively.

4.4.2.3 Daily stock price

Final data set is obtained from CRSP, which provides daily stock prices for all public firms. Using them, we estimate the expected return during the event window for each of announcing firms when they announce excess inventory. See the details about the event study methodology in §4.5.1.

4.4.2.4 Variables

To test our hypothesis, we generate the following variables for our analysis.

Dependent Variables:

- $EI_{it} = \begin{cases} 1, & \text{if firm } i \text{ announces excess inventory in year } t \\ 0, & \text{otherwise} \end{cases}$

Independent Variables:

- $TFP(LP)_{it} = \text{estimated TFP by LP for firm } i \text{ in year } t$
- $TFP(OP)_{it} = \text{estimated TFP by OP for firm } i \text{ in year } t$
- $IT_{it} = \text{inventory turns for firm } i \text{ in year } t = \frac{Cost\ of\ Goods\ Sold_{it}}{Annual\ Total\ Inventory_{it} + LIFO\ reserve_{it}}$
- $Sales\ Growth_{it} = \text{sales growth for firm } i \text{ in year } t = \frac{Sales_{it}}{Sales_{i,t-1}}$
- $Market\ Value_{it} = \text{market value of the firm's common equity for firm } i \text{ in year } t$

  $= \text{closing stock price (PRCC\_F in Compustat)}$

  $\times \text{ the number of shares outstanding (in millions of shares, CSHO in Compustat)}$

We use mean-adjusted variables for the measure of operational competence such as TFP and IT for the following reasons. First, different industries within the U.S. retail sector have different levels of TFP and IT. Our data show that the estimated firm level TFP (both LP and OP) varies significantly by two-digit SIC (Table 5.2.1 in the Appendix). Overall mean TFP by LP is 2.54 with standard deviation of 0.35. Heterogeneity in IT across industries in the U.S. retail sector is well documented by Gaur et al. (2005). Second, firm's TPF and IT may vary over time even in the same industry by, for example, introducing new technology or changing the top management team. To account for heterogeneities both across industries and across years, we calculate the mean TFP and IT for each year in each industry, then subtract them from raw values. Hence, we use mean-adjusted variables (i.e., $MeanAdj\_TFP_{it}$ and $MeanAdj\_IT_{it}$) in our analysis. For a robustness check, we also report the results with (1) median-adjusted and (2) raw values of TFP and IT without any adjustment.

Summary statistics and the Pearson correlation coefficients among all variables used in our analysis are provided in Tables 4.2-a and 4.2-b, respectively. Mean-centered variables are used to compute the correlation coefficients because of panel structure of our data. We note that correlation between mean-adjusted TFP estimated by LP and that by OP is 0.95, indicating that both estimates are robust each other. We trim the top 1% and bottom 1% of observations based on TFP by LP and OP, IT, sales growth, and market value of the firm’s common equity. This approach ensures that our analyses are not influenced by extreme outliers.

Table 4.2-a: Summary Statistics

Variable		**Mean**	**Std. dev.**	**Min**	**Max**
$ExcessInv_{it}$	Overall	0.018	0.133	0	1
	Between		0.081	0	1
	Within		0.123	-0.482	0.968
$MeanAdj\_TFP(LP)_{i,t-1}$	Overall	-0.007	0.169	-0.422	0.821
	Between		0.169	-0.415	0.796
	Within		0.088	-0.673	0.799
$MeanAdj\_TFP(OP)_{i,t-1}$	Overall	-0.003	0.123	-0.386	0.524
	Between		0.115	-0.320	0.469
	Within		0.071	-0.433	0.459
$MeanAdj\_IT_{i,t-1}$	Overall	-1.181	5.110	-9.790	36.307
	Between		5.553	-8.970	35.161
	Within		2.196	-18.130	28.553
$Sales\ Growth_{i,t-1}$	Overall	1.123	0.195	719	2.332
	Between		0.168	0.807	2.145
	Within		0.159	0.562	2.243
$Market\ Value_{i,t-1}$	Overall	1734.193	4045.649	1.533	35615.55
	Between		2772.876	1.771	19854.83
	Within		2178.522	-12967.22	21529.26

Note: Outliers are removed (<1% & >99%). SIC 58 is not included in our sample since it does not have an excess inventory announcement. The number of observations is 4739 and the number of firms is 578 for all variables except $MeanAdj\_TFP(OP)_{i,t-1}$ which has 4711 observations with 576 firms. Average length of years is 8.20.

Table 4.2-b: Pearson Correlation

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
$(1) ExcessInv_{it}$	1					
$(2) MeanAdj\_TFP(LP)_{i,t-1}$	**-0.0470**	1				
$(3) MeanAdj\_TFP(OP)_{i,t-1}$	**-0.0417**	**0.9514**	1			
$(4) MeanAjd\_IT_{i,t-1}$	-0.0258	*0.0342*	**0.0413**	1		
$(5) Sales\ Growth_{i,t-1}$	0.0230	**0.1059**	**0.1299**	-0.0210	1	
$(6) Market\ Value_{i,t-1}$	0.0117	0.0296	*0.0336*	0.0279	0.0013	1

Note: All bold coefficients have *p*<0.01 and italicized coefficients have *p*<0.05. All variables are mean-centered except $ExcessInv_{it}$.

4.4.3 Model specification and analysis

This section presents the econometrics model to test Hypothesis 1 about the impact of operational competence on excess inventory announcement. We propose the following probit model:

$$\Pr(EI_{it} = 1|\boldsymbol{X}) = \Phi(\boldsymbol{X}'\beta) \qquad (13)$$

where Φ denotes standard normal cdf and ***X*** is a vector of covariates including two-digit SIC industry fixed effect (b_I), year fixed effect (a_t), main variables of interest to measure firm's operational competence ($MeanAdj\_TFP_{i,t-1}$ and $MeanAdj\_IT_{i,t-1}$), and two control variables ($Sales\ Growth_{i,t-1}$ and $Market\ Value_{i,t-1}$).

In our empirical model specification, we include year specific dummies (a_t) and two-digit SIC industry specific dummies (b_I) to account for time- and industry-specific unobserved heterogeneity (i.e., selection on observables). By using all explanatory variables in a lagged form (i.e., imposing pre-existing condition), we can minimize the concern of simultaneous problem and estimate conditional probability of announcing excess inventory at time *t* given all the information available at time *t*-1. Two control variables other than time fixed and industry fixed effects are used: sales growth controls for different growth rate across retail firms; and market value of firm's common stock controls for firm size.

We run a random effect probit model instead of fixed effect for the following reasons. First, our dependent variable, an indicator of excess inventory announcement, has only 85 observations (73 unique firms) with value 1. If we use a fixed effect model, we cannot use the firms that only have one value of dependent variables (i.e., no time variation over time). For example, if firm A does not report any excess inventory announcement, firm A is omitted from our sample to run a fixed effect model because all observations of dependent variable in firm A are 0. Thus, we have to use a small sample of 73 unique firms if we run a fixed effect model, which result in a huge loss in sample size. Second, recent papers in economics literature (Arellano and Honore 2001; Hahn 2001; Laisney and Lechner 2003; Greene 2004; Cerro 2007), by analyzing the fixed effects model on binary choice dependent variable (e.g., probit model), show that the fixed effect estimator is inconsistent and substantially biased away from zero. Third, an analysis of decomposed overall variation into between and within variations supports our use of random effect model. In Table 4.2-a, all explanatory variables have larger between-variation than within-variation. We use probit model as a main model; however, we also check the robustness of our result with different model specifications such as logit and complementary log-log models.

4.4 Results

4.4.1 Results: Determinants of excess inventory announcement

Table 4.3 presents regression results to unveil determinants of excess inventory announcement. We report the result with only control variables in column (1); then add TFP and IT in columns (2) and (3), respectively. Finally both TFP and IT are entered in column (4).

Table 4.3: Regression Result for Productivity and Excess Inventory Announcement (Probit Model)

DV: $\boldsymbol{Excess\ Inventory_{it}}$	(1)	(2)	(3)	(4)
$MeanAdj\_TFP_{i,t-1}$		-1.24***		-1.22***
		(0.35)		(0.36)
$MeanAdj\_IT_{i,t-1}$			-0.03**	-0.03*
			(0.017)	(0.016)
$Sales\ Growth_{i,t-1}$	0.07	0.31	0.08	0.32
	(0.24)	(0.24)	(0.24)	(0.24)
$Market\ Value_{i,t-1}$	7.39×10^{-6}	-6.71×10^{-6}	9.63×10^{-6}	-5.28×10^{-6}
	(1.2×10^{-5})	(1.31×10^{-5})	(1.21×10^{-5})	(1.32×10^{-5})
Year fixed effect	Yes	Yes	Yes	Yes
Industry fixed effect	Yes	Yes	Yes	Yes
# of observations	4739	4739	4739	4739
# of firms	578	578	578	578
Log likelihood	-410.03	-402.96	-407.22	-400.57

Note: *** p<0.01, ** p<0.05, * p<0.1.

Consider the results from column (2) that support our conjecture that operationally competent retailers have fewer excess inventory announcements. The coefficient of one-year-lagged mean-adjusted TFP captures the impact of operational competence on the probability of announcing excess inventory in the following year. As we expected, we find that one-year-lagged mean-adjusted TFP has negative impact on the probability of announcing excess inventory (-1.24, p<0.01). High TFP retailers, those in the top 90th percentile of TFP, are 3.53 times less likely to announce excess inventory than low TFP retailers, those in the bottom 10th percentile. Thus, we find evidence supporting Hypothesis 1 and conclude that TFP is a predictor of excess inventory announcement. While the direction of this result is intuitive, it is still useful to not only document this finding but, more importantly, the large magnitude we observe.

When we use IT as a measure of operational competence as shown in column (3), we find that one-year-lagged mean-adjusted IT has negative impact on the probability of announcing excess inventory (-0.03, p<0.05). High IT retailers, those in the top 90th percentile of IT, are 2.33

times less likely to report excess inventory than low IT retailers, those in the bottom 10th percentile. When we use IT and TFP in the same model as shown in column (4), we find that the coefficients of both IT and TFP remain similar indicating that this pair of variables have low correlation, confirmed in Table 4.2-b, but we find that significance of IT reduces to 10% level. So, it appears that TFP is a better predictor of excess inventory announcement than IT.

We note that the coefficients' estimates of the control variables are in the expected direction. The year fixed effect and two digit SIC industry fixed effect are significant in the entire models, implying that unobserved year specific error and industry specific error (i.e., selection on observables) should be controlled. The market value of the firm's common stock, which is the proxy for firm size, and sales growth are not significantly associated with the chance of reporting excess inventory in our data.

4.4.4.2. Robustness checks

We perform a number of tests to show the robustness of our result, namely the negative impact of operational competence on excess inventory announcement (Table 4.4).

Table 4.4: (Robustness Check) Regression Result for Productivity and Excess Inventory Announcement

DV: $Excess\ Inventory_{it}$	Alternative Model Specifications		Alternative Adjustment		Alternative TFP	Alternative Inventory Efficiency	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Logit	Complementary log-log	Median adjusted	Raw	OP	AIT	GMROI
$MeanAdj\_TFP_{i,t-1}$	-2.95^{***}	-2.91^{***}				-1.20^{***}	-1.20^{***}
	(0.87)	(0.86)				(0.38)	(0.35)
$MedianAdj\_TFP_{i,t-1}$			-1.14^{***}				
			(0.35)				
$TFP_{i,t-1}$				-1.25^{***}			
				(0.36)			
$MeanAdj\_TFP(OP)_{i,t-1}$					-1.36^{***}		
					(0.45)		
$MeanAdj\_IT_{i,t-1}$	-0.07^{*}	-0.07^{*}			-0.03^{*}		
	(0.04)	(0.04)			(0.016)		
$MedianAdj\_IT_{i,t-1}$			-0.04^{**}				
			(0.02)				
$IT_{i,t-1}$				-0.03^{**}			
				(0.02)			
$MeanAdj\_AIT_{i,t-1}$						0.02	
						(0.11)	
$MeanAdj\_GMROI_{i,t-1}$							-0.03
							(0.03)
$Sales\ Growth_{i,t-1}$	0.78	0.76	0.31	0.33	0.30	0.24	0.31
	(0.54)	(0.53)	(0.24)	(0.24)	(0.24)	(0.25)	(0.24)
$Market\ Value_{i,t-1}$	-1.28×10^{-5}	-1.27×10^{-5}	-3.51×10^{-6}	-5.60×10^{-6}	8.57×10^{-7}	-2.99×10^{-6}	-6.51×10^{-6}
	(3.08×10^{-5})	(3.03×10^{-5})	(1.31×10^{-5})	(1.33×10^{-5})	(1.26×10^{-5})	(1.31×10^{-5})	(1.31×10^{-5})
Year fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# of observations	4739	4739	4739	4793	4711	4534	4734
# of firms	578	578	578	578	576	571	577
Log likelihood	-400.75	-400.78	-400.77	-399.89	-401.88	-384.72	-402.23

Note: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Alternative Model Specifications: We examine the stability of the results using different model specifications. In the main model, we use probit model (equation (4.13)) where we assume the cumulative standard normal distribution as a link function. We can alternatively use logit model where a link function is assumed by the cumulative logistic distribution (i.e., $\Pr(EI_{it} = 1|\boldsymbol{X}) =$

$\left[1+e^{-\boldsymbol{X}'\beta}\right]^{-1}$). As the binary dependent variable in our data is asymmetric, we can also use a complementary log-log model specification which assumes $\Pr(EI_{it}=1|\boldsymbol{X})=1-e^{-e^{\boldsymbol{X}'\beta}}$as a link function. Columns (1) and (2) in Table 4.4 show the results with logit model and complementary log-log model, respectively. The results are very similar to the ones obtained with probit model (Table 4.3). This shows that our substantive results are no artifact of the specific model chosen in the analysis.

Alternative Data Adjustments: We test the validity of our theoretical model using alternative adjustments for TFP and IT. In the main model, we adjust TFP and IT by their mean values calculated by each industry in each year. It helps us control for both heterogeneities across industries and across years. The mean, however, is sensitive to extreme values and median is preferable in this case. Hence we use median-adjusted TFP and IT instead of mean-adjusted TFP and IT. As a further robustness check, we also report the result using raw variables without any adjustment. As seen in columns (3) and (4) in Table 4.4, the conclusions remain unchanged, indicating that our main result is not affected by extreme values.

Alternative TFP: We examine the sensitivity of the results to an alternative measure of TFP introduced by Olley and Pakes (1996). As we discussed in §4.4.1, OP is an alternative way of measuring the firm level TFP. Since the correlation between mean-adjusted TFP estimated by OP and that by LP is 0.95 (shown in Table 4.2-b), we expect to see robust results. The main conclusions do not change with TFP estimated by OP (column (5) in Table 4.4).

Alternative Inventory Efficiency: We repeat our analysis with alternative proxies of inventory efficiency. We use two additional metrics other than IT: growth margin return on inventory (GMROI) and adjusted inventory turnover (AIT). See Alan et al. (2014) for details of definition and computation. The results are reported in columns (6) and (7) in Table 4.4 for mean-adjusted

AIT and GMROI, respectively. Our main finding, the negative impact of TFP on the probability of announcing excess inventory, is still consistent. Interestingly, the both proxies for inventory efficiency are not significantly associated with the likelihood of announcing excess inventory any more. Using other adjustments such as median-adjusted or raw variables still estimate insignificant coefficients for the proxies of inventory efficiency.

Recession Period: We rerun the regression with recession dummies instead of year specific dummies. The main result is unchanged and recession dummies are not statistically significant (table is omitted), implying that the probability of announcing excess inventory does not differ from recession period to non-recession period.

To sum up, we find that excess inventory announcement is not just a function of randomness in demand; it is systematically correlated with the announcing firm's operational competence. We also find that TFP is a better metric of overall operational competence than IT.

4.5 Market Reaction

4.5.1 Event study methodology

Our main measure for the second research question about market reaction is the short-term abnormal returns (AR) accruing from excess inventory announcements to the focal firm, estimated by the event study methodology (see Brown and Warner 1985 for a review of this methodology). Using stock price (i.e., shareholder value) as a performance metric has several advantages: It is forward looking, integrates multiple dimensions of performance, and is less easily manipulated by managers than other measures (Gielens et al. 2008). Event studies usually enable 1) to test for the existence of information effects of event (e.g., the impact of the excess inventory announcement on market value of stock price) and 2) to identify factors that enlighten changes in market value of stock price (e.g., announcing firm's operational competence).

Consistent with the approach used in many event studies (Brown and Warner 1985), we measure abnormal returns over a two-day event period (i.e., the day of the announcement and the day before the announcement date). If the excess inventory announcement is made before 4 PM Eastern Standard Time (EST), the event window includes the day of announcement and the preceding trading day to account for the possibility that the information about the event may have been released the day before the announcement. If the excess inventory announcement is made after 4 PM EST, then the event window consists of the day of the announcement and the trading day after the announcement to account for the fact that the market cannot act until the next trading day. We translate calendar days into event days as follows. For announcements made before 4 PM EST, the announcement calendar day is Day 0 in event time, the next trading day is Day +1, and the trading day before the announcement is Day -1, and so on. For announcements made after 4 PM EST, the announcement calendar day is Day -1 in event time, the next trading day is Day 0, and the trading day before the announcement is Day -2, and so on. In addition to a two-day event period, we also use a three-day event window (from Day -1 to 1) as a robustness check.

The information effects of excess inventory announcements are assessed by computing the difference between the observed return, R_{id}, on the event date and the expected return, $\mathbb{E}[R_{id}]$, estimated on a benchmark model.

$$AR_{id} = R_{id} - \mathbb{E}[R_{id}] \tag{4.14}$$

The observed return, R_{id}, is expressed as the percentage change in stock price:

$$R_{id} = \frac{P_{id} - P_{i,d-1}}{P_{i,d-1}} \tag{4.15}$$

where P_{id} is the closing stock price for announcing firm i on day d. The price P_{id} incorporates the long term impacts of the additional information becoming public on the day d. It follows the “efficient market” (or “rational expectation”) paradigm which assumes a complete and immediate investor response to any available information. Consistent to the literature (e.g., Kalaignanam et al. 2013; Hendricks et al. 2014), we estimate the expected return, $\mathbb{E}[R_{id}]$, using the Fama-French four-factor model that includes the three factors identified by Fama and French (1993) and the momentum factor identified by Carhart (1997):

$$\mathbb{E}[R_{id}] = \left[\hat{\alpha}_i + \hat{\beta}_i R_{md} + \hat{\gamma}_i SMB_d + \hat{\delta}_i HML_d + \hat{\sigma}_i UMD_d\right] \tag{4.16}$$

where R_{md} is the stock return of the benchmark market portfolio, SMB_d is the difference between rate of returns of small and big stock firms, HML_d is the difference in returns between high and low book-to-market ratio stocks, and UMD_d is the momentum factor defined as the difference in returns between firms with high and low past stock performance.

We estimate the expected daily stock returns for each firm using OLS regression over the estimation period from day -220 to day -21. In estimating the parameters we require that a firm must have a minimum of 40 stock returns during the estimation period of 200 trading days. Abnormal returns are estimated as the difference between the observed returns, R_{id}, and the expected returns, $\mathbb{E}[R_{id}]$:

$$AR_{id} = R_{id} - \mathbb{E}[R_{id}] = R_{id} - \left[\hat{\alpha}_i + \hat{\beta}_i R_{md} + \hat{\gamma}_i SMB_d + \hat{\delta}_i HML_d + \hat{\sigma}_i UMD_d\right] \tag{4.17}$$

The abnormal returns are aggregated for a firm over an event window $[-d_1, d_2]$ and are given by

$$CAR[-d_1, d_2] = \sum_{d=-d_1}^{d_2} AR_{id} \tag{4.18}$$

When information leakage (for d_1 days before the event) and/or dissemination over time (for d_2 days after the event) occur, the abnormal returns for a firm are aggregated over the “event window” $[-d_1, d_2]$ into a cumulative abnormal return (CAR). Because the event study is conducted across *N* different events, the individual CARs can be averaged into a cumulative average abnormal return (CAAR).

$$CAAR[-d_1, d_2] = \sum_{n=1}^{N} \frac{CAR_n[-d_1, d_2]}{N} \tag{4.19}$$

4.5.2 Model specification

This section presents the econometrics model to test our hypotheses about the market response to excess inventory announcements. We propose the following model:

$$CAR_i[-1, 0] = a_t + b_I + \theta_1 MeanAdj\_TFP_i + \theta_2 MeanAdj\_IT_i \tag{4.20}$$

$$+\theta_3 Sales\ Growth_i + \theta_4 lnSale_i + \varepsilon_i$$

Similar to the previous model in equation (13), we include two fixed effects. First, year fixed effect (a_t) is included to account for the unobservable yearly shock which can impact on multiple announcements made in a specific year. Second, industry fixed effect (b_I) is added to account for the unobservable industry shock which can impact on announcements made in a specific industry. Two control variables are used: sales growth controls for different growth rate across retailers; and natural logarithm of sales for firm size. Note that we ensure that all explanatory variables are in the most recent fiscal year completed before the date of the excess inventory announcement (i.e., typically one-year-lagged form).

4.5.3 Results

4.5.3.1 Results: Main effect of excess inventory announcement on stock price

We examine the cumulative average abnormal returns for the 85 excess inventory announcements across different event windows. We find a statistically significant abnormal return on the announcement day (-2.18%, *p*<0.01). Table 4.5 shows the results for four different event windows using the four-factor model with statistics such as the cross-sectional variance-adjusted Patell test statistic. Notice that all four event windows show significantly negative mean abnormal returns. For example, $CAAR[-1, 0]$ is -2.53% (p<0.01), meaning that the stock market reflects the information of holding accumulated inventory by penalizing the announcing firm's stock price. Although 47.06% of the companies are positively affected, the CAR is negative for 52.94%. To reduce the influence of outliers, we supplement the *t*-statistic with non-parametric test, the Wilcoxon signed-rank test. It shows that the median abnormal return is statistically different from zero (p<0.10).

Table 4.5: Cumulative Average Abnormal Returns across Different Event Windows

Window	**Mean Abnormal return (%)**	**Patell *t*-statistic**	**% Positive[a]**	**Rank test *Z*-statistic**
[-1, 0]	-2.53	-4.21***	47.06	-1.56*
[0, +1]	-2.06	-3.42***	37.65**	-0.92
[-1, +1]	-2.41	-3.27***	44.71	-1.19
[-2, +2]	-2.60	-2.73***	48.24	-1.56*

Note: *** *p*<0.01, ** *p*<0.05, * *p*<0.1. The *p*-values are one-tailed. The sample size is 85.
[a] *p*-value is calculated based on Binomial sign test.

We find two interesting observations. First, 47.06% of the retailers have positive abnormal stock market returns around the excess inventory announcement. Hendricks and Singhal (2009) observed only 27% of firms to have positive stock market reaction. In other words, there is considerable heterogeneity in the market's response to the announcement. To explain this anomaly, we consider the information provided by the retailers for the reasons for excess inventory buildup and actions that they have taken or plan to take to handle the excess inventory, later in the section 4.5.4.

Second, comparing to Hendricks and Singhal (2009) that show the decline in the stock price by -6.79% to -6.93% due to excess inventory announcement, our finding is smaller in magnitude. This is because we focus on the retail sector whereas Hendricks and Singhal (2009) study all sectors in the U.S. It may indicate that the stock market perceives the excess inventory announcement more negatively when non-retailers announce it. In fact, Hendricks and Singhal (2009) show that if the excess inventory is with customers, the announcing firm has additional penalty of approximately 2.5% in the stock price. As retailers are the customer of other firms, but not the opposite, we expect to observe a smaller decline in the stock price when retailers announce excess inventory. Hence, our finding is consistent with Hendricks and Singhal (2009).

4.5.3.2 Results: Market reaction across the announcing firm's operational competence

Consistent with the approach used in many event studies (e.g., Brown and Warner 1985; Hendricks and Singhal 2009; Kalaignanam et al. 2013; Hendricks et al. 2014), we use a *CAR* measure in two-day event period as a dependent variable in our cross-sectional analysis. We estimate equation (20). In order to reduce the endogeneity issue, we ensure that all explanatory variables are in the most recent fiscal year completed before the date of the excess inventory announcement (i.e., one-year-lagged form). The results are presented in Table 4.6. We initially include TFP (i.e., $MeanAdj\_TFP_i$) and IT (i.e., $MeanAdj\_IT_i$) separately in columns (1) and (2), respectively, and then add both metrics in column (3). As a robustness check, we also report results with a three-day event window (i.e., $CAR_i[-1, 1]$) in columns (4)-(6).

Table 4.6: Regression Result for Productivity and Cumulative Abnormal Return

	$CAR_i[-1,0]$			$CAR_i[-1,1]$		
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
$MeanAdj\_TFP_i$	-29.35**		-29.98*	-39.12***		-39.03***
	(14.42)		(15.17)	(14.18)		(14.14)
$MeanAdj\_IT_i$		0.18	0.34		-0.26	-0.05
		(0.74)	(0.75)		(0.81)	(0.77)
$Sales\ Growth_i$	-8.98	-10.98	-8.27	-12.38	-16.01*	-12.48
	(7.65)	(7.86)	(7.79)	(7.68)	(8.38)	(8.02)
$lnSale_i$	-4.08***	-2.51**	-4.21***	-5.08***	-2.84**	-5.06***
	(1.28)	(1.17)	(1.33)	(1.56)	(1.41)	(1.56)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
# of observations	85	85	85	85	85	85
Adjusted R^2	0.198	0.119	0.187	0.180	0.079	0.165

Note: *** p<0.01, ** p<0.05, * p<0.1. Robust standard error is reported in the parenthesis.

Consider the result from columns (1) and (4). We find that one-year-lagged mean-adjusted TFP is significantly and negatively associated with the cumulative abnormal returns in a two-day event window (-29.35, p<0.05) and in a three-day event window (-39.12, p<0.01). This means that the stock market penalizes more severely when the announcing firm is indeed high TFP retailer. Ceteris paribus, an increase in one-year-lagged mean-adjusted TFP by one-standard-deviation (i.e., 0.141) is associated with -4.14% in the stock return over a two-day period (the day of the announcement and the day before the announcement) and -5.52% over a three-day period (from the day before the announcement to the day after the announcement). Hence, we find supporting evidence of Hypothesis 2A.

Consistent with Hendricks and Singhal (2009), we find that one-year-lagged mean-adjusted IT is not associated with the cumulative abnormal returns in two-day and three-day event windows (columns (2) and (5)). It can substantiate prior literature (e.g., Kesavan et al. 2010; Kesavan and Mani 2012) which finds that the stock market does not fully incorporate the information contained in inventory. When we use IT and TFP in the same model as shown in columns (3) and (6), we find that our results remain qualitatively the same, so we use it for further robustness checks.

We note that the coefficients' estimates of the control variables are in the expected direction. The natural log of sales, which is the proxy for firm size, is negatively associated with the abnormal return (-4.21, p<0.01). It implies that larger retailers experience more negative abnormal returns than smaller retailers. The sales growth, which is the proxy for firm's growth rate, is not associated with the abnormal return in our data. It may imply that the market does not differentiate its response to excess inventory announcement across the growth rate of announcing firm.

4.5.3.3 Robustness checks

In addition to using alternative dependent variable (i.e. $CAR_i[-1, 1]$) in columns (4)-(6) in Table 4.6, we perform following robustness checks for our main results, more negative reaction of the stock market to excess inventory announcement when the announcing firm is a high TFP retailer (Table 4.7): alternative adjustments for TFP and IT by using median-adjusted and raw variables shown in columns (1) and (2), respectively; alternative measure of TFP by using OP approach presented in column (3); alternative measure of inventory efficiency by using AIT and GMROI shown in columns (4) and (5), respectively; and using recession dummies instead of yearly dummies (omitted). The substantive conclusions remain unchanged throughout all different models, indicating that our main results are robust.

Table 4.7: (Robustness Check) Regression Result for Productivity and Cumulative Abnormal Return

DV: $\boldsymbol{CAR_i[-1,0]}$	**Alternative Adjustment**		**Alternative TFP**	**Alternative Inventory Efficiency**	
	(1)	**(2)**	**(3)**	**(4)**	**(5)**
	Median adjusted	**Raw**	**OP**	**AIT**	**GMROI**
$MeanAdj\_TFP_i$				-33.84*	-28.65*
				(16.94)	(12.49)
$MedianAdj\_TFP_i$	-28.55*				
	(15.87)				
TFP_i		-31.01**			
		(14.97)			
$MeanAdj\_TFP(OP)_i$			-33.25*		
			(17.01)		
$MeanAdj\_IT_i$			0.27		
			(0.73)		
$MeanAdj\_AIT_i$				1.47	
				(3.58)	
$MeanAdj\_GMROI_i$					0.54
					(0.99)
$MedianAdj\_IT_i$	0.39				
	(0.78)				
IT_i		0.40			
		(0.79)			
$Sales\ Growth_i$	-8.37	-8.13	-8.09	-11.80	-8.72
	(7.56)	(7.80)	(7.84)	(7.79)	(7.57)
$lnSale_i$	-4.12***	-4.25***	-3.64***	-4.78***	-3.94***
	(1.32)	(1.30)	(1.19)	(1.47)	(1.28)
Year fixed effects	Yes	Yes	Yes	Yes	Yes
Industry fixed effects	Yes	Yes	Yes	Yes	Yes
# of observations	85	85	85	80	85
Adjusted R^2	0.179	0.192	0.183	0.221	0.189

Note: *** p<0.01, ** p<0.05, * p<0.1. Robust standard error is reported in the parenthesis.

4.5.4 Value of information in excess inventory announcements

Another aspect to explore is whether our finding, the negative association between the announcing firm’s operational competence and abnormal returns due to excess inventory announcement, differs by the information contained in the announcement. Previous literature (Sorescu et al. 2007) shows the positive impact of the information offered in the announcement on short-term abnormal returns. They use price and time to introduction of new product as key information in the announcement as this information can reduce uncertainty on the future cash flow of the announcing firm.

In the excess inventory announcement, the actions that the firm has taken or plans to take to deal with excess inventory and the reasons of building excess inventory are potential information that investors can utilize. Providing information on follow-up actions to handle the buildup of excess inventory can reduce the investor's uncertainty on the future cash flow of the announcing firm. The stock market might perceive this information credible as the market can verify action information provided in the announcement. The announcing firm is also likely to fulfill its claim (i.e., action) because increasing reliability (i.e., the extent to which the firm has fulfilled claims it made) is a component of firm reputation (Sorescu et al. 2007). With an expectation that operationally competent retailers have high reputation compared to their less competent peers, competent retailers are more likely to keep their promise (i.e., action) as the costs of a loss of reputation are greater. Hence, the action information may moderate the negative link that we found in the previous section, so that the stock market may penalize less severely when operationally competent retailers announce excess inventory with follow-up action information compared to when they do not provide such action information.

On the contrary, providing information on why the announcing firm accumulated excess inventory might not help investors visualize the announcing firm's future cash flow. As the true reason of holding excess inventory is not observable to the stock market, the announcing firm is less likely to reveal it. Consistent with this argument, many retailers (35.42%) blame external factors such as sluggish sales as a main reason of building excess inventory whereas very few retailers (4.71%) mention internal factors such as internal inefficiency and poor execution as shown in section 4.2.1. Knowing this, the stock market might not perceive reason information credible. Hence, the reason information may not have a moderating effect.

To formally test this idea, we create two indicator variables: $Action_i$ and $Reason_i$. $Action_i$ ($Reason_i$) is defined as 1 if the announcement provides action (reason) information, and zero otherwise. We also create two interaction variables with TFP. Then we re-estimate the model in equation (4.20) with those created variables. Columns (1) – (3) in Table 4.8 are the main results and (4) – (7) are robustness checks, respectively. We find that providing action information in the announcement moderates the negative association between firm's operational competence and abnormal stock returns. An increase in one-year-lagged mean-adjusted TFP by one-standard-deviation (i.e., 0.141) is associated with -6.35% in the stock returns over a two-day period *without* action information whereas it is associated with -1.61% in the stock returns over a two-day period *with* action information as shown in column (1). In contrast, we do not find such moderating effect with reason information in column (2). These observations are consistent with our argument above.

Table 4.8: Regression Result for Moderator of Productivity and Cumulative Abnormal Return

DV: $CAR_i[-1,0]$	Main models			Alternative DV	Alternative Adjustment		Alternative TFP
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Action	Reason	Both	$CAR_i[-1,1]$	Median adjusted	Raw	OP
$MeanAdj\_TFP_i$	-45.05***	-20.81	-38.07	-58.11***			
	(13.24)	(24.18)	(26.81)	(11.82)			
$MedianAdj\_TFP_i$					-43.41***		
					(13.68)		
TFP_i						-43.54***	
						(12.49)	
$MeanAdj\_TFP(OP)_i$							-49.36***
							(15.50)
$MeanAdj\_IT_i$	0.43	0.59	0.65	0.06			0.45
	(0.72)	(0.75)	(0.72)	(0.74)			(0.72)
$MedianAdj\_IT_i$					0.46		
					(0.72)		
IT_i						0.53	
						(0.74)	
$Action_i$	2.14		1.82	0.82	1.29	-14.19*	1.73
	(4.81)		(5.30)	(4.57)	(4.75)	(7.49)	(4.55)
$Action_i \times MeanAdj\_TFP_i$	33.61*		29.35	37.89**			
	(17.06)		(19.33)	(16.52)			
$Action_i \times MedianAdj\_TFP_i$					33.31*		
					(17.04)		
$Action_i \times TFP_i$						31.73**	
						(13.98)	
$Action_i \times MeanAdj\_TFP(OP)_i$							37.46*
							(21.21)
$Reason_i$		-5.26*	-4.44				
		(2.99)	(3.06)				
$Reason_i \times MeanAdj\_TFP_i$		-8.65	-4.05				
		(25.32)	(25.68)				
$Sales\ Growth_i$	-10.67	-8.82	-10.78	-14.69*	-11.00	-11.09	-10.15
	(7.94)	(7.66)	(7.79)	(8.50)	(7.85)	(7.97)	(7.93)
$lnSale_i$	-3.51**	-3.71***	-3.16**	-4.42**	-3.45**	-3.73***	-3.17***
	(1.49)	(1.31)	(1.51)	(1.69)	(1.46)	(1.40)	(1.27)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# of observations	85	85	85	85	85	85	85
Adjusted R^2	0.199	0.203	0.204	0.173	0.189	0.209	0.188

Note: *** *p*<0.01, ** *p*<0.05, * *p*<0.1. Robust standard error is reported in the parenthesis.

Note that when we add both action and reason indicators with corresponding interactions in column (3), we have directionally the same result although main variables are insignificant due to multicollinearity and small sample. Hence, we use model in column (1) for robustness checks. We show the robustness of our result with following tests: alternative dependent variable by using $CAR_i[-1,1]$ (column (4)); alternative adjustment of TFP and IT by using median-adjusted (column (5)) and raw variables (column (6)); and alternative measure of TFP by using OP approach (column (7)). All results are consistent.

Our results show that when high TFP retailers announce the buildup of excess inventory, they can still be largely unscathed if they can explain how they plan to tackle the problem. In other words, when an operationally competent retailer announces excess inventory and explains the follow-up action, then the market is willing to minimize the penalty. When we divide 85 announcements into four groups (high vs. low TFP and with vs. without actions), we find that retailers in the top 50th percentile of TFP face a -3.78% (median) decline in stock returns when they announce excess inventory without providing follow-up actions but face a 1.74% (median) increase in stock returns when they provide follow-up actions. We conjecture that the market trusts the competent companies to turn-around their operations when provided with a definite plan-of-action. Hence, the worst case scenario is really when high competent firms announce excess inventory and do not provide follow-up actions. This seeds doubts about the company's competency and makes the stock price decline the most. In contrast, we do not find such difference in stock market reaction to whether retailers in the bottom 50th percentile of TFP provide follow-up actions or not to fixing the excess inventory problem (-0.75% vs. -1.86% in median).

4.6 Summary and Conclusion

Based on an analysis of a combined dataset of excess inventory announcements, annual financial statements, and daily stock prices of publicly traded retailers in the U.S. during 1990-2011, we document that excess inventory announcement is negatively affected by the announcing firm's operational competence, measured by TFP. We also show two findings from the stock market's response to excess inventory announcements. First, the market more severely penalizes the announcing firm's stock price when the firm is high TFP retailer. Ceteris paribus, an increase in one-year-lagged mean-adjusted TFP by one-standard-deviation (i.e., 0.141) is associated with -4.14% in the stock returns over a two-day period (the day of the announcement and the day before the announcement). Second, providing action information in the announcement can mitigate the negative abnormal returns when high operationally competent retailers announce excess inventory. An increase in one-year-lagged mean-adjusted TFP by one-standard-deviation is associated with -6.35% in the stock returns over a two-day period *without* action information whereas it is associated with -1.61% in the stock returns over a two-day period *with* action information.

The main results presented in this paper have a number of implications. First, by suggesting empirical evidence that firm's operational competence is negatively associated with the likelihood of announcing the buildup of excess inventory in the following year, we show that excess inventory is not a random phenomenon merely driven by demand uncertainty, which is typically harder for managers to control, but by management practices. Hence, excess inventory appears to be manageable through better operations.

Second, our result shows a potential value of using the firm level TFP as an important metric. By adding it to the current important metrics like IT, retailers may enjoy following benefits. Retailers will be able to predict the odd of announcing excess inventory as we find that TFP is a better predictor of excess inventory announcement than IT. In addition, retailers will be able to anticipate the stock market's response in the future when they announce excess inventory for a given level of their TFP. So, retailers can make a better plan for the future.

Third, our results provide a possible explanation for an observation made in operations management literature. Recent papers have shown that investments based on inventory turns yield higher abnormal stock market returns (Kesavan and Mani 2013; Alan et al. 2014). There are two possible expositions offered for this finding. One is *information*-based argument: the stock market might not be fully incorporating inventory information in pricing stocks (Kesavan et al. 2010; Kesavan and Mani 2013). The other is *risk*-based argument: high IT retailers could be riskier than low IT retailers as they have higher returns (Alan et al. 2014). Our study provides evidence for the former and against the latter. By contrasting with TFP, our study finds that stock market reaction differs across high TFP and low TFP retailers but there is no significant difference in market reaction to announcements from high IT and low IT retailers. This result suggests that the stock market may not be distinguishing between high IT and low IT retailers (i.e., information-based argument), leading to abnormal returns in the future. In addition, we show that the low IT retailers are potentially riskier than high IT retailers because they have a greater likelihood of announcing excess inventory compared to high IT retailers (i.e., against risk-based argument for IT). For the risk-based argument for TFP, consistent with Imrohoroglu and Tuzel (2014), we find that the low TFP retailers are riskier than high TFP retailers by providing evidence that low TFP retailers is more likely to announce excess inventory compared to high TFP retailers.

As with all studies, our work has limitations that bear noting and offer opportunities for future work. We find the heterogeneity of the firm level and industry level TFP in the U.S. retail sector (explained in details in the Appendix). This paper, however, does not describe the reasons why the firm level TFP varies across firms and across industries as that is not the main focus of this paper and hence beyond the scope of this work. By using firm specific and industry specific characteristics, the future research might be able to give rigorous expositions. Moreover, what causes differences in firm level TFP between online retailers and brick-and-mortar retailers would be fruitful avenue of future research.

Our study has a caveat on measurement that may be examined in future research. Our measure of excess inventory is a binary variable based on public announcement of excess inventory. We are unable to perform an analysis based on the magnitude of excess inventory since our data lack such information. For example, the available data only allow us to examine the probability of announcing excess inventory based on the operational competence in the previous year, but we do not know how much excess inventory the firm suffers from. If more sophisticate data are available in the future, one can extend our model to incorporate the magnitude of excess inventory into the relationship between firm's operational competence and excess inventory announcement.

Altogether this paper fills the gap suggested by Hendricks and Singhal (2009): "It could be useful to build an understanding of some of the ***underlying drivers of excess inventory*** and to find whether the negative effect of excess inventory ***varies by these drivers.***" We show that firm's operational competence, measured by TFP, is an underlying driver of excess inventory and the negative market reaction varies by it.

APPENDIX I: VALIDATING THE INVERTED U-SHAPED RELATIONSHIP AT RETAILER B

In this section, we present additional results for Chapter 2.

Table 5.1.1: Inverted-U Relationship with Hourly dummies (Retailer A)

Dependent Variable:	**Sales ($Sales_{th}$)**	
	(1) **interaction between hour-block dummies and day-of-the-week dummies**	**(2)** **hourly dummies**
$A\_Fit\_Traffic_{th}$	1.69***	1.51***
	(0.45)	(0.46)
$A\_Fit\_Traffic_{th}^2$	-0.04***	-0.03***
	(0.003)	(0.003)
$A\_Traffic_{th}$	13.71***	12.09***
	(0.29)	(0.34)
$A\_Traffic_{th}^2$	0.001	0.006***
	(0.001)	(0.001)
$Labor_{th}$	17.40***	14.38***
	(3.48)	(3.48)
$Promotion_t$	-106.81***	-80.23***
	(24.03)	(23.97)
Controls	Yes	Yes
Observations	5312	5312
Adjusted R^2	0.8204	0.8240

Note. *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Controls include interactions between hour-block dummies and day-of-the-week dummies, and monthly dummies in column (1) and hourly dummies, day-of-the-week dummies, and monthly dummies in column (2), respectively. In column (2), store traffic and its square have VIFs beyond ten. It indicates multicollinearity issue, making harder to interpret coefficient estimates for these two variables.

To see whether our result is robust to other retailers, we further test the inverted U-shaped relationship between fitting room traffic and store sales at another retailer (retailer B), where the fitting room layout is completely different from that of retailer A. We have only 24 hourly data points (6 hours per day for four days) that we collected by observing each customer in the period of a field study, as retailer B had not installed technology to obtain traffic data.

Table 5.1.2-a shows the inverted U-shaped relationship between fitting room traffic and sales at the nearest checkout counter from the studied fitting room area. Model (1) does not control for any time effect as we have limited sample size. Model (2) includes hour dummies, Model (3) includes a Saturday indicator, and Model (4) includes both. Throughout different model specifications, even though the sample size is small, we find that coefficient of $A\_Fit\_Traffic_{th}$ is positive and statistically significant (*p*<0.1) and $A_Fit_Traffic_{th}^2$is negative and statistically significant (*p*<0.1) and the peak point is within the data range (0, 35); again, this supports Hypothesis 1b. We also conduct several robustness tests for the inverted U-shape. For example, Fieller's interval of [10.26, 34.27] for the peak point is within the data. The spline regressions in Table 5.1.2-b shows that the coefficient of the first spline is positive and significant (*p*<0.05), whereas the second is negative and insignificant (*p*=0.103), partially supporting the inverted U-shaped relationship between fitting room traffic and sales. The conclusions are similar when we consider the case with three knots, as shown in column (2). Since we have very limited sample size, coefficients are not significant, but they are all on the right direction.

In conclusion, we find the inverted U-shaped relationship between fitting room traffic and store sales for both retailers A and B. Since these two retailers are different in terms of location, carried products, gross margin, and notably fitting room layouts, this finding could be robust to any retailer that uses fitting rooms.

Table 5.1.2-a: Inverted-U Relationship (Retailer B)

Dependent Variable: $\boldsymbol{Sales_{th}}$	**Model (1)**	**Model (2)**	**Model (3)**	**Model (4)**
$A\_Fit\_Traffic_{th}$	128.06***	121.82***	96.44**	84.69*
	(31.43)	(33.63)	(40.87)	(43.05)
$A\_Fit\_Traffic_{th}^2$	-4.54***	-4.27***	-3.49**	-3.03*
	(1.31)	(1.44)	(1.57)	(1.68)
Saturday dummy	No	No	Yes	Yes
Hour dummies	No	Yes	No	Yes
Observations	24	24	24	24
Adjusted R^2	0.4417	0.4193	0.4527	0.4462

Note. *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 5.1.2-b: Robustness Checks Using Spline Regressions (Retailer B)

Dependent Variable: $\boldsymbol{Sales_{th}}$	**(1) One knot**	**(2) Two knots**
$A\_Fit\_Traffic_{th}$ *1*	72.64**	54.42
	(25.82)	(38.33)
$A\_Fit\_Traffic_{th}$ *2*	-26.90	32.47
	(15.75)	(29.56)
$A\_Fit\_Traffic_{th}$ *3*		-48.30
		(42.18)
Saturday dummy	Yes	Yes
Observations	24	24
Adjusted R^2	0.4965	0.3829

Note. *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 5.1.3: Phantom Stockouts (Details on day 1, Sunday)

Time (each visit)	Total items brought by associate (each visit)	# items available in the store					Phantom Stockouts (each visit)
		1	2	3	4	Over 5	
11:45	4	0	2	0	0	2	0.00
12:15	18	5	9	2	0	2	0.28
12:40	1	0	1	0	0	0	0.00
12:48	1	0	1	0	0	0	0.00
13:02	2	1	0	1	0	0	0.50
13:10	4	2	2	0	0	0	0.50
13:13	1	1	0	0	0	0	1.00
13:25	5	1	4	0	0	0	0.20
13:35	7	1	4	0	2	0	0.14
13:50	5	0	2	1	2	0	0.00
13:56	12	7	4	1	0	0	0.58
14:24	3	2	1	0	0	0	0.67
14:40	11	6	1	3	0	1	0.55
14:52	3	2	1	0	0	0	0.67
14:55	6	1	4	1	0	0	0.17
15:24	11	4	6	1	0	0	0.36
15:35	18	7	6	1	3	1	0.39
15:50	24	8	12	1	1	2	0.33
16:00	8	3	4	0	1	0	0.38
16:18	24	7	6	6	3	2	0.29
16:30	13	7	4	2	0	0	0.54
16:48	4	2	2	0	0	0	0.50
17:05	1	1	0	0	0	0	1.00
17:20	5	1	1	0	0	3	0.20
17:25	14	5	6	1	2	0	0.36
17:40	13	3	3	3	2	2	0.23
17:58	5	1	2	1	0	1	0.20
Total	223	78	88	25	16	16	0.37

Note. Phantom stockouts (each visit) is defined as "#items with availability of 1 (each visit) / Total items brought by associate (each visit)," where the total items brought by associate is a part of items left behind in the fitting room area.

Table 5.1.4: Phantom Stockouts with Price, Size, and Category (Details on day 2, Saturday)

			# items available in the store				
		Total	**1**	**2**	**3**	**4**	**Over 5**
	# items	336	138	106	55	22	15
	(%)	1.00	0.41	0.32	0.16	0.07	0.04
	Petite	45	23	14	8	0	0
	(%)	0.13	0.17	0.13	0.15	0.00	0.00
	Clearance	18	10	6	1	0	0
	(%)	0.05	0.07	0.06	0.02	0.00	0.00
	Swimming	53	21	20	8	2	2
	(%)	0.16	0.15	0.19	0.15	0.09	0.13
Price	mean	53.77	61.53	53.29	45.62	37.27	39.73
	s.d.	28.52	34.37	26.05	13.33	16.96	13.87
	median	49	50	48.5	49	40	42
	min	9.97	11.97	9.97	9.99	11.97	11.97
	max	160	160	151.95	69	69	58
Size	2	2	1	1	0	0	0
	4	1	0	1	0	0	0
	6	9	6	1	2	0	0
	8	12	5	6	0	1	0
	10	7	3	4	0	0	0
	12	17	10	3	3	1	0
	14	23	4	10	7	1	1
	16	14	9	3	0	0	2
	18	1	1	0	0	0	0
	XS	6	6	0	0	0	0
	S	36	17	11	6	1	1
	M	46	19	14	4	7	2
	L	72	15	28	18	5	6
	XL	45	19	10	7	6	3
	4P	1	1	0	0	0	0
	6P	1	1	0	0	0	0
	8P	3	2	0	1	0	0
	10P	2	1	1	0	0	0
	12P	1	0	0	1	0	0
	PP	1	1	0	0	0	0
	PS	13	4	6	3	0	0
	PM	12	4	5	3	0	0
	PL	8	7	1	0	0	0
	PXL	3	2	1	0	0	0

Note. Average price of 3200 SKUs in this section, the immediate sales floor from the studied fitting room area, is $46.

APPENDIX II: ALTERNATIVE VIEW OF THE INCENTIVE CHANGE

In this section, we provide additional results for Chapter 3.

An alternative view of the incentive change is to increase the relative importance of dependent component to the independent component in the bonus scheme. While the way to calculate the amount of eligible bonus based on the store performance (i.e., independent component) is unchanged, managers obtain bonus only when entire retail chain has a positive profit (i.e., dependent component). This transition might boost cooperation among stores as the bonus is strongly linked with how other stores perform. In general, the role of cooperative actions such as helping and knowledge sharing in improving organizational performance has been widely acknowledged (Cummings 2004). Hence, it is possible that the change in incentive improves the overall performance of the retail chain.

An opposite viewpoint is also possible for the following reasons. First, performance improvement due to cooperation among stores such as helping and knowledge sharing requires stores to be linked with each other. Siemsen (2007) identified three distinct types of linkages: (1) outcome linkages, (2) help linkages, and (3) knowledge linkages. In our setting, none of them is strong enough to influence each other. Store outcomes such as sales are not dependent upon other stores' outcomes. Store managers in one location cannot directly help the other managers in a different location. Knowledge sharing might be possible, but it would be limited due to heterogeneity across stores with local knowledge. Second, the profit for this retailer had significantly dropped in the previous 5 quarters and became negative in the previous quarter. This trend might make store managers pessimistic about getting bonus. As a result, it is possible that the new incentive considerably reduced the motivation for store managers, resulting in lower performance of the retail chain.

We test these arguments using similar model specification proposed in the main analysis. We use sales growth, defined as difference between sales in 2014 and that in 2013 in percentages of sales in 2013 (i.e., $Sales\ growth_{it} = \frac{Sales\ 2014_{it} - Sales\ 2013_{it}}{Sales\ 2013_{it}}$), to capture overall store sales performance. We use traffic growth, defined as difference between average daily traffic in 2014 and that in 2013 in percentages of the average daily traffic in 2013 (i.e., $Traffic\ growth_{it} = \frac{Avg\ daily\ traffic\ 2014_{it} - Avg\ daily\ traffic\ 2013_{it}}{Avg\ daily\ traffic\ 2013_{it}}$), to capture the unexpected demand shock. Table 5.2.1 shows the results. We do not have controls in column (1) and add store fixed effect in column (2). We further add traffic growth to account for the time-related demand shock in column (3). Finally, we rule out selection mechanism in column (4) by considering 33 stores with no manager change during study period. Throughout all columns, we find that overall store sales performance is smaller under the post-incentive change period compared to the pre-incentive change period. This result indicates that we can see the change in incentive as a transition from the strong incentive to the weak one. We use this interpretation in Chapter 3.

Table 5.2.1: Overall Store Performance is Smaller in the Post-Incentive Change Period

Dependent variable:	$\boldsymbol{Sales\ growth_{it}}$			
	(1) **Unconditional**	**(2)** **Store heterogeneity**	**(3)** **Unobservable surprise**	**(4)** **No selection**
$Weak\ incentive_{it}$	-0.027***	-0.027***	-0.025***	-0.016*
	(0.005)	(0.005)	(0.005)	(0.008)
$Traffic\ growth_{it}$			0.51***	0.56***
			(0.09)	(0.07)
Store effect	No	Yes	Yes	Yes
Observations	3248	3248	2922	1197
Adjusted R^2	0.0110	0.0925	0.3410	0.3646

Note. *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Standard error, clustered by store, is reported in parenthesis.

APPENDIX III: HETEROGENEITY OF FIRM LEVEL AND INDUSTRY LEVEL TFP IN THE U.S. RETAIL SECTOR

In this section, we provide additional results for Chapter 4.

We find that the overall retail productivity has been increased in the U.S. during 1963 to 2011 (Figure 5.3.1). However, the different industry shows different tendency during the recent five decades. Even in the same industry, each firm has different TFP over time. Figure 5.3.2 and 5.3.3 depict the heterogeneity of retail productivity across industries and years during 1963-2011, respectively. For example, while the retail giant Wal-Mart (SIC 53) shows a decreasing trend of productivity, the online retailer Amazon.com (SIC 59) has an increasing trend of productivity. In the same industry (SIC 52), Home Depot and Lowe's show different tendency though both firms have a decreasing trend of productivity.

Figure 5.3.1: Mean Retail Productivity across Years (1963~2011)

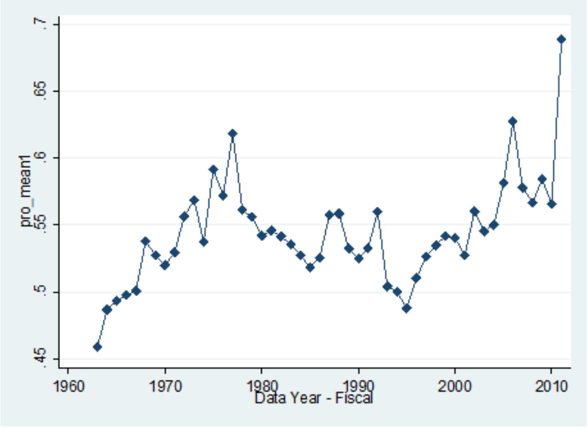
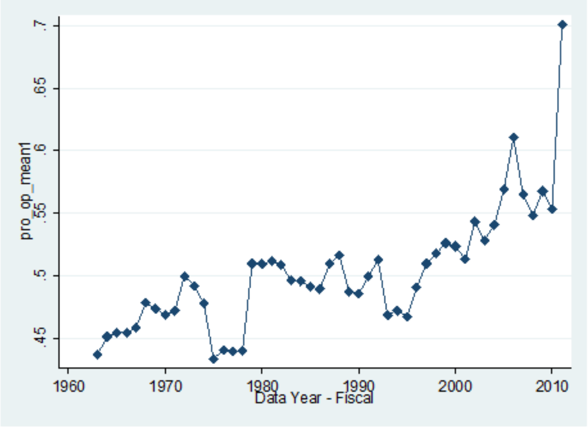
Levinsohn & Petrin (2003)	Olley & Pakes (1996)
	

Figure 5.3.2: Heterogeneity of Retail Productivity across Industries (SIC2)

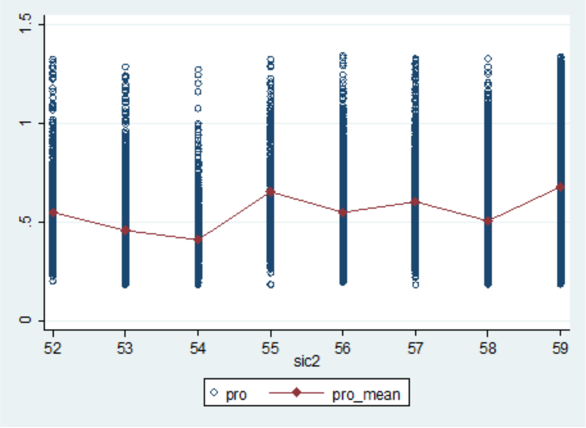
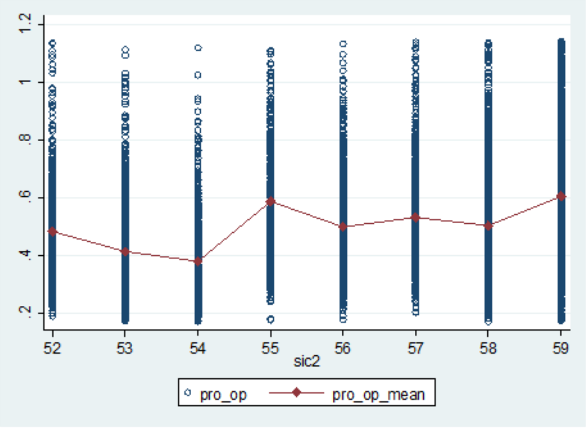
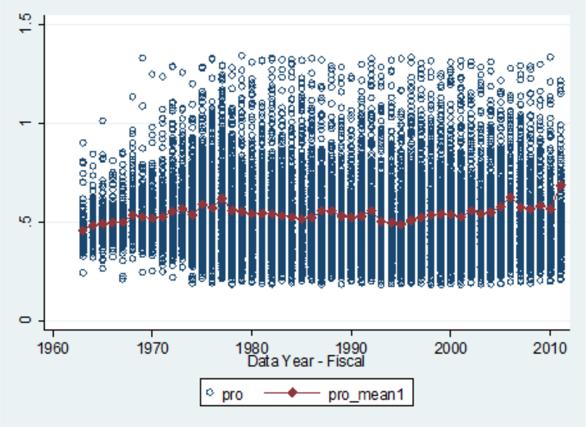
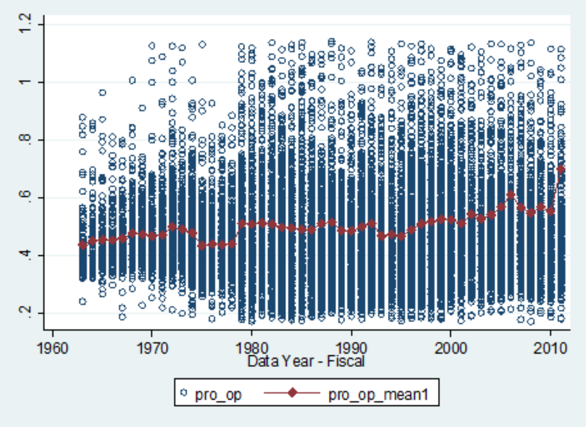
Levinsohn & Petrin (2003)	Olley & Pakes (1996)
	

Figure 5.3.3: Heterogeneity of Retail Productivity across Years (1963~2011)

Levinsohn & Petrin (2003)	Olley & Pakes (1996)
	

The two main robust findings in the productivity literature also show heterogeneity across industries in the U.S. retail sector. The first finding, which is the ubiquity of the estimated productivity gap, is supported by all retail industries, but the magnitude differs by industry. Table 5.3.1 illustrates the 90-10 TFP ratio across two-digit SIC industries. The overall 90-10 TFP ratio in the U.S. retail sector is 2.54 and the standard deviation is 0.352 while Syverson (2004) finds the average 90-10 TFP ratio of 1.92 and standard deviation of 0.173 in the U.S.

manufacturing sector. It indicates that the U.S. retail sector has a larger productivity gap than U.S. manufacturing sector and the former has larger difference between industries than the latter. For instance, the SIC 55, Automotive dealers & service stations, shows the largest productivity gap between the firms in 90th and 10th productivity distribution (2.919) followed by SIC 59, Miscellaneous retail, while the SIC 54, Food stores, has the least productivity gap (2.055). We note that SIC 59 includes online retailers like Amazon.com.

Table 5.3.1: 90-10 TFP Ratio across Industries (SIC2)

sic2	52	53	54	55	56	57	58	59	Overall	Stand.Dev.
LP	2.726	2.479	2.055	2.919	2.139	2.450	2.078	2.874	2.540	0.352
OP	2.055	1.828	1.618	2.387	1.786	1.931	1.738	2.405	2.035	0.294

The second finding, stickiness of tremendous and persistent productivity gap, is also supported by all retail industries, but again the magnitude varies across industries. Table 5.3.2-a presents the overall autoregressive coefficients in the retail sector and Table 5.3.2-b and 5.3.2-c describe the autoregressive coefficients across industries in LP and in OP, respectively. We find the autoregressive coefficient of 0.64 in the retail sector. It is robust with productivity literature, which ranges autoregressive coefficients between 0.6 and 0.8 (e.g., Abraham and White 2006; Foster et al. 2008). Interestingly, SIC 55, Automotive dealers & service stations, has the highest autoregressive coefficients of 0.862 while SIC 53, General merchandise, shows the lowest autoregressive coefficients of 0.292. It implies that SIC 55 shows very sticky productivity over year while SIC 53 has less sticky productivity across years.

Table 5.3.2-a: Autoregressive Coefficients

$Productivity_{i,t}$	**LP**	**OP**
$Productivity_{i,t-1}$	**0.644**	**0.730**
	(0.004)	(0.004)
# of observations	16387	15136
R^2	0.585	0.637

Note: All bold coefficients have p<0.01.

Table 5.3.2-b: Autoregressive Coefficients across Industries (LP)

$Productivity_{i,t}$	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**
$Productivity_{i,t-}$	**0.834**	**0.292**	**0.787**	**0.862**	**0.614**	**0.527**	**0.794**	**0.641**
	(0.019)	(0.009)	(0.009)	(0.019)	(0.019)	(0.016)	(0.008)	(0.009)
# of observations	680	2302	2364	709	2006	1127	3533	3666
R^2	0.749	0.295	0.778	0.800	0.335	0.478	0.7265	0.582

Note: All bold coefficients have p<0.01.

Table 5.3.2-c: Autoregressive Coefficients across Industries (OP)

$Productivity_{i,t}$	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**
$Productivity_{i,t-}$	**0.805**	**0.265**	**0.796**	**0.851**	**0.761**	**0.676**	**0.835**	**0.758**
	(0.023)	(0.010)	(0.009)	(0.018)	(0.25)	(0.018)	(0.008)	(0.009)
# of observations	680	2302	2364	709	2006	1127	3533	3666
R^2	0.653	0.243	0.755	0.767	0.320	0.562	0.740	0.672

Note: All bold coefficients have p<0.01.

BIBLIOGRAPHY

Abowd J (1990) Does performance-based managerial compensation affect corporate performance? *Industrial and Labor Relations Review* **43**(3): 52-74.

Abraham A, White K (2006) The dynamics of plant-level productivity in U.S. manufacturing. *Center for Economic Studies Working Paper* 06-20.

Aiken LS, West SG (1991) *Multiple Regression: Testing and Interpreting Interactions.* Sage, Newbury Park, CA.

Aksin Z, Ata B, Emadi S, Su CL (2013) Structural estimation of callers' delay sensitivity in call centers. *Management Sci.* **59**(12): 2727-2746.

Alan Y, Gao GP, Gaur V (2014) Does inventory productivity predict future stock returns? A retailing industry perspective. *Management Science* **60**(10): 2416-2434.

Allon G, Hanany E (2012) Cutting the line: Social norms in queues. *Management Sci.* **58**(3): 493-506.

Al-Ubaydli O, List JA (2015) Do natural field experiments afford researchers more or less control than laboratory experiments? A simple model. *Working paper*.

Amemiya Y (1985) Instrumental variable estimator for the nonlinear errors-in-variables model. *J. Econometrics* **28**(3): 273-289.

Anand KS, Pac MF, Veeraraghavan S (2011) Quality-speed conundrum: Trade-offs in customer-intensive services. *Management Sci.* **59**(1): 157-171.

Anderson S, Kekker H, Sedatole K (2010) An empirical examination of goals and performance-to-goal following the introduction of an incentive bonus plan with participative goal setting. *Management Science* **56**(1):90-109.

Arellano M, Honore B (2001) Panel data models: Some recent developments. Leamer E and Heckman J, eds., *Handbook of Econometrics*, Volume 5, Amsterdam, North Holland, 3229-3296.

Bandiera O, Barankay I, Rasul I (2005) Social preferences and the response to incentives: Evidence from personnel data. *Quarterly Journal of Economics* **120**(3): 917-962.

Banker RD, Lee S-Y, Potter G (1996) A field study of the impact of a performance-based incentive plan. *Journal of Accounting and Economics* **21**: 195-226.

Banker RD, Lee S-Y, Potter G, Srinivasan D (2000) An empirical analysis of continuing improvements following the implementation of a performance-based compensation plan. *Academy of Management Journal* **30**(3): 315-350.

Bartelsman JE, Mark D (2000) Understanding productivity: Lessons from longitudinal microdata. *Journal of Economic Literature* **38**(3): 569-594.

Barua A, Kribel C, Mukhopadhyay T (1995) Information technology and business value: An analytic and empirical evaluation. *Information Systems Research* **7**(4): 409-428.

Basu A, Lal R, Srinivasan V, Staelin R (1985) Salesforce compensation plans: An agency theoretic perspective. *Marketing Science* **4**(Fall): 267-291

Batt RJ, Terwiesch C (2015) Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Sci.* **61**(1): 39-59.

Batt RJ, Terwiesch C (2016) Early task initiation and other load-adaptive mechanisms in the emergency department. *Management Sci. forthcoming*.

Benabou R, Tirole J (2003) Intrinsic and extrinsic motivation. *Review of Economic Studies* **70**: 489-520.

Bernstein E, Kesavan S, Staats B, Hassall L (2014) Towards exceptional scheduling. *Harvard Business School Case* 415-023.

Bertrand M, Schoar A (2003) Managing with style: The effect of managers on firm policies. *Quarterly Journal of Economics* **118**(4): 1169-1208.

Bertrand M, Deflo E, Mullainathan S (2004) How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics* **119**(1): 249-275.

Bertschek I, Kaiser U (2004) Productivity effects of organizational change: Microeconometric evidence. *Management Science* **50**(3): 394-404.

Bidwell M, Briscoe F, Fernandez-Mateo I, Sterling A (2013) The employment relationship and inequality: How and why changes in employment practices are reshaping rewards in organizations. *Academy of Management Annals* **7**(1): 61-121.

Billington C, Johnson B, Triantis A (2002) A real options perspective on supply chain in management of technology. *Journal of Applied Corporate Finance* **15**(Summer): 32-43.

Bitran GR, Ferrer J, Oliveira PR (2008) Managing customer experiences: Perspectives on the temporal aspects of service encounters. *Manufacturing Service Oper. Management* **10**(1): 61-83.

Bloom N, Van Reenen J (2007) Measuring and explaining management practices across firms and countries. *Quart. J. Econom.* **122**(4): 1351-1408.

Bonner S, Sprinkle G (2002) The effects of monetary incentives on effort and task performance: Theories, evidence, and a framework for research. *Accounting, Organizations and Society* **27**: 303-345.

Brooks R, May D, Mishra C (2001) The performance of firms before and after they adopt accounting-based performance plans. *Quarterly Review of Economics and Finance* **41**(2): 205-222.

Brown SJ, Warner JB (1985) Using daily stock returns: The case of event studies. *Journal of Financial Economics* **14**(1): 3-31.

Brynjolfsson E, Hitt LM (1996) Paradox lost? Firm-level evidence on the returns to information systems spending. *Management Science* **42**(4): 541-559.

Brynjolfsson E, Hitt LM (2003) Computing productivity: Firm-level evidence. *Review of Economics and Statistics* **85**: 793-808.

Buell WR, Campbell D, Frei FX (2010) Are self-service customers satisfied or stuck? *Production Oper. Management* **19**(6): 679-697.

Buell WR, Norton MI (2011) The labor illusion: How operational transparency increases perceived value. *Management Sci.* **57**(9): 1564-1579.

Cachon G (2003) Sypply chain coordination with contracts. Graves A, de Kok SC, eds. *Handbooks in Operations Research and Management Science: Supply Chain Management*. North Holland, Armsterdam, chapter 6, 227-339.

Cadsby C, Song F, Tapon F (2007) Sorting and incentive effects of pay for performance: An experimental investigation. *Academy of Management Journal* **50**(2): 387-405.

Camerer C, Hogarth R (1999) The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty* **19**(1): 7-42.

Campbell D (1988) Task complexity: A review and analysis. *Academy of Management Review* **13**: 40-52.

Carhart MM (1997) On persistence in mutual fund performance. *The Journal of Finance* **52**(1): 57-82.

Chan CW, Yom-Tov G, Escobar G (2014) When to use speedup: An examination of service systems with returns. *Oper. Res.* **62**(2): 462-482.

Chan T, de Vericourt F, Besbes O (2014) Contracting in medical equipment maintenance services: An empirical investigation. *Working paper.*

Chen H, Frank M, Wu O (2005) What actually happened in the inventories of American companies between 1981 and 2000? *Management Science* **51**(7): 1015-1031.

Chopra S, Sodhi MS (2004) Managing risk to avoid supply-chain breakdown. *Sloan Management Review* **45**(Fall): 53-61.

Cummings JN (2004) Work groups, structural diversity, and knowledge sharing in a global organization. *Management Science* **50**(3): 352-364.

Deci E, Koestner R, Ryan R (1999) A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin* **125**(6): 627-668.

DeHoratius N, Raman A (2008) Inventory record inaccuracy: An empirical analysis. *Management Sci.* **54**(4): 627-641.

DeHoratius N, Ton Z (2015) The role of execution in managing product availability. Agrawal N, Smith S, eds. Retail Supply Chain Management. Springer, 53-77.

Deo S, Gurvich I (2011) Centralized vs. decentralized ambulance diversion: A network perspective. *Management Sci.* **57**(7): 1300-1319.

Dewan S, Kraemer KL (2000) Information technology and productivity: Evidence from country-level data. *Management Science* **46**(4): 548-563.

Dion D (1999) A theoretical and empirical study of retail crowding. *Eur. Advances Consumer Res.* **4**: 51-57.

Dong J, Feldman P, Yom-Tov GB (2015) Service systems with slowdowns: Potential failures and proposed solutions. *Oper. Res.* **63**(2): 305-324.

Eroglu SA, Machleit KA (1990) An empirical study of retail crowding: Antecedents and consequences. *J. Retailing* **66**(2): 201-221.

Fama EF, French K (1992) The cross section of expected stock returns. *The Journal of Finance* **47**(2): 427-465.

Fama EF, French K (1993) Common risk factors in the returns on stocks and bonds. *The Journal of Financial Economics* **33**(1): 3-56.

Feldman Z, Mandelbaum A, Massey WA, Whitt W (2008) Staffing of time-varying queues to achieve time-stable performance. *Management Science* **54**(2): 324-338.

Fieller EC (1954) Some problems in interval estimation. *J. Roy. Statist. Soc. Ser. B.* 175-185.

Fisher ML, Raman A, McClelland A (2000) Rocket science retailing is almost here. Are you ready? *Harvard Business Review* **78**(March-April): 115-124.

Fisher ML, Krishnan J, Netessine S (2007) Retail store execution: An empirical study. *Working paper*.

Fisher ML, Raman A (2010) *The New Science of Retailing*. Harvard Business Press, Boston.

Foster L, Haltiwagner J, Syverson C (2008) Reallocation, firm turnover and efficiency: Selection on productivity or profitability. *American Economic Review* **98**: 394-425.

Frei FX (2006) Breaking the trade-off between efficiency and service. *Harvard Bus. Rev.* **84**(11).

Frei FX (2008) The four things a service business must get right. *Harvard Bus. Rev.* **86**(4).

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review and research prospects. *Manufacturing Service Oper. Management* **5**(2): 79-141.

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

Gavett G (2015) How self-service kiosks are changing customer behavior. *Harvard Bus. Rev.* March 11.

Gielens K, Van De Gucht LM, Steenkamp JBEM, Dekimpe MG (2008) Dancing with a giant: The effect of Wal-Mart’s entry in the United Kingdom on the performance of European retailers. *Journal of Marketing Research* **45**(5): 519-534.

Gneezy U, Rustichini A (2001) Pay enough or don’t pay at all. *The Quarterly Journal of Economics* **115**(3): 791-810.

Green LV (2004) Capacity planning and Management in Hospitals. Brandeau ML, Sainfort F, Pierskalla WP, eds. *Operations Research and Health Care: A Handbook of Methods and Applications.* KIluwer Academic Publishers, Norwell, MA, 15-42.

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**: 98-119.

Guajardo JA, Cohen MA, Kim SH, Netessine S (2012) Impact of performance-based contracting on product reliability: An empirical analysis. *Management Science* **58**(5): 961-979.

Hahn J (2001) The information bound of a dynamic panel logit model with fixed effects. *Econometric Theory* **17**: 913-932.

Harrison GW, List JA (2004) Field Experiments. *J. Econ. Lit.* **42**(4): 1009-1055

Hassin R, Haviv M (2003) *To Queue or Not To Queue: Equilibrium Behavior in Queueing Systems*, Boston: Kluwer Academic Publishers.

Hendricks KB, Singhal VR (2009) Demand-supply mismatches and stock market reaction: Evidence from excess inventory announcements. *Manufacturing Service Oper. Management* **11**(3): 509-524.

Hendricks KB, Hora M, Singhal VR (2014) An empirical investigation on the appointments of supply chain and operations management executives. *Management Science* **61**(7): 1562-1583.

Hennessy DA, Wiesenthal DL (1997) The relationship between traffic congestion, driver stress and direct versus indirect coping behaviours. *Ergonomics* **40**(3): 348-361.

Hu K, Allon G, Bassamboo A (2016) Understanding customers retrial in call centers: Preferences for service quality and service speed. *Working paper*.

Ichniowski C, Shaw KL (2009) Insider econometrics: Empirical studies of how managerment matters. *Working paper*.

Ilmakunnas P, Maliranta M, Vainiomaki J (2004) The roles of employer and employee characteristics for plant productivity. *Journal of Productivity Analysis* **21**(3): 249-276.

Imrohoroglu A, Tuzel S (2014) Firm level productivity, risk, and return. *Management Science* **60**(8): 2073-2090.

Kalaignanam K, Kushwaha T, Steenkamp J-B, Tuli KR (2013) The effects of CRM outsourcing on shareholder value: A contingency perspective. *Management Science* **59**(3): 748-769.

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

Kc DS (2013) Does multitasking improve performance? Evidence from the emergency department. *Working Paper*.

Kesavan S, Gaur V, Raman A (2010) Do inventory and gross margin data improve sales forecasts for U.S. public retailers? *Management Science* **56**(9): 1519-1533.

Kesavan S, Mani V (2013) The relationship between abnormal inventory growth and furture earnings for U.S. public retailers. *Manufacturing Service Oper. Management* **15**(1): 6-23.

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

Kesavan S, Mani V (2015) An overview of industry practice and empirical research in retail workforce management. Agrawal N, Smith S, eds. *Retail Supply Chain Management.* Springer Science+Business Media, New York, 113-145.

Kesavan S, Kushwaha T, Gaur V (2016) Do high and low inventory turnover retailers respond differently to demand shocks? *Manufacturing Service Oper. Management*, forthcoming.

Kulkarni VG (2009) *Modeling and Analysis of Stochastic Systems*, 2nd ed. Chapman & Hall, London, UK.

Laisney F, Lechner M (2003) Almost consistent estimation of panel probit models with “small” fixed effects. *Econometric Reviews* **22**(1): 1-28.

Lam SY, Vandenbosch M, Hulland J, Pearce M (2001) Evaluating promotions in shopping environments: Decomposing sales response into attraction, conversion, and spending effects. *Marketing Sci.* **20**(2): 194-215.

Lambert S, Henly J (2010) *Work scheduling study: Managers’ strategies for balancing business requirements with employee needs: Manager survey results*, University of Chicago SSA, Chicago, IL.

Langer EJ, Saegert S (1977) Crowding and cognitive control. *J. Personality Soc. Psych.* **35**(3) 175-182.

Lazear EP (2000) Performance pay and productivity. *The American Economic Review* **90**(5): 1346-1361.

Lee HS, Kesavan S, Deshpande V (2017) Understanding and managing customer-induced negative externalities in congested self-service environments. *Working paper*.

Lee H-H, Pinker EJ, Shumsky RA (2012) Outsourcing a two-level service process. *Management Sci.* **58**(8): 1569-1584.

Levinsohn J, Petrin A (2003) Estimating production functions using inputs to control for unobservables. *Review of Economic Studies* **70**(2): 317-341.

Li J, Granados N, Netessine S (2014) Are consumers strategic? Structural estimation from the air-travel industry. *Management Sci.* **60**(9): 2114-2137.

Libby R, Lipe M (1992) Incentives, effort, and the cognitive processes involved in accounting-related judgements. *Journal of Accounting Research* **30**(2): 249-273.

Lieberman MB, Demeester L (1999) Inventory reduction and productivity growth: Linkages in the Japanese automotive industry. *Management Science* **45**(4) 466-485.

Lind JT, Mehlum H (2010) With or without U? The appropriate test for a U-shaped relationship. *Oxford Bull. Econom. Statist.* **72**(1): 109-118.

Linebaugh K (2009) Corporate news: Toyota reduces production. *The Wall Street Journal*, January 16, 2009, B2.

Lo D, Ghosh M, LaFontaine F (2011) The incentive and selection roles of sales force compensation contracts. *Journal of Marketing Research* **48**(4): 781-798.

Lu Y, Olivares M, Musalem A, Schilkurt A (2013) Measuring the effect of queues on customer purchases. *Management Sci.* **59**(8): 1743-1763.

Mani V, Kesavan S, Swaminathan JM (2015) Estimating the impact of understaffing on sales and profitability in retail stores. *Production Oper. Management* **24**(2): 201-218.

Marschak J, Andrews WH (1944) Random simultaneous equations and the theory of production. *Econometrica* **12**: 143-205.

McWilliams G, Dodes R (2007) Fashion faux pas hurts Wal-Mart. *The Wall Street Journal*, May 21, 2007, A8.

Milgrom P, Roberts J (1992) *Economics, Organization, and Management*, Prentice Hall, Englewood Cliffs, NJ.

Moon Y, Frei F (2000) Exploding the self-service myth. *Harvard Bus. Rev.* **78**(3): 26-27.

Morgan A, Poulsen A (2001) Linking pay to performance-compensation proposals in the S&P 500. *Journal of Financial Economics* **62**(3): 489-523.

Mukhopadhyay T, Kekre S, Kalathur S (1995) Business value of information technology: A study of electronic data interchange. *MIS Quarterly* **19**(2): 137-156.

Musalem A, Olivares M, Schilkrut A (2016) Retail in high definition: Monitoring customer assistance through video analytics. *Working paper*.

Narayanan VG, Raman A (2004) Aligning incentives in supply chains. *Harvard Business Review* **82**(November): 105-116.

Natvig B (1975) On a queueing model where potential customers are discouraged by queue length. *Scandinavian Journal of Statistics* **2**: 34-42.

Netessine S, Fisher ML, Krishnan J (2010) Labor planning, execution, and store performance: An exploratory investigation. *Working paper*.

Nevo A, Wolfram C (2002) Why do manufacturers issue coupons? An empirical analysis of breakfast cereals. *RAND J. Econom*. **33**(2): 319-339.

Oettinger G (2001) Do piece rates influence effort choices? Evidence from stadium vendors. *Economics Letters* **73**:117-123.

Olley S, Pakes A (1996) The dynamics of productivity in the telecommunications equipment industry. *Econometrica* **64**(6): 1263-1297.

Paarsch H, Shearer B (2000) Piece rates, fixed wages, and incentive effects: Statistical evidence from payroll records. *International Economic Review* **41**(1):59-91.

Pelham B, Neter E (1995) The effect of motivation of jedgment depends on the difficulty of the judgment. *Journal of Personality and Social Psychology* **68**(4):581-594.

Perdikaki O, Kesavan S, Swaminathan JM (2012) Effect of traffic on sales and conversion rates of retail stores. *Manufacturing Service Oper. Management* **14**(1):145-162.

Prendergast C (1999) The provision of incentives in firms. *Journal of Economic Literature* **37**(1):7-63.

Prendergast C (2011) What have we learnt about pay for performance? Geary lecture winter 2010. *The Economic and Social Review* **42**(2):113-134

Quan V (2004) Retail labor scheduling. *OR/MS Today*.

Rafaeli A, Barron G, Haber K (2002) The effects of queue structure on attitudes. *J. Service Res.* **5**(2): 125-139.

Raman A, DeHoratius N, Ton Z (2001) Execution: The missing link in retail operations. *California Management Rev*. **43**(3): 136-152.

Remus W, O’Connor M, Griggs K (1998) The impact of incentivees on the accuracy of subjects in judgemental forecasting experiments. *International Journal of Forecasting* **14**: 514-522.

Rumyantsev S, Netessine S (2007). Should inventory policy be lean or responsive? Evidence for US public companies. INSEAD, *Working paper*.

Sasser W (1976) Match suppy and demand in service industries. *Harvard Bus. Rev.* **54**(6): 133-140.

Shende S (2015) Global self service technologies market – opportunities and forecasts, 2015-2020. *Allied Market Research Report*, May 2015.

Siebert WS, Zubanov N (2010) Management economics in a large retail company. *Management Sci*. **56**(8): 1398-1414.

Siemsen E, Balasubramanian S, Roth AV (2007) Incentives that induce task-related effort, helping, and knowledge sharing in workgroups. *Management Science* **53**(10): 1533-1550.

Simonsohn U (2016) Twi-lines: The first valid test of U-shaped relationships. *Working paper*.

Sorescu A, Shankar V, Kushwaha T (2007). New product preannouncements and shareholder value: Don't make promises you can't keep. *Journal of Marketing Research* **44**(3): 468-489.

Staiger D, Stock JH (1997) Instrumental variables regression with weak instruments. *Econometrica* **65**(3): 557-586.

Staiger D, Stock JH, Giland WG (1997) The NAIRU, unemployement and monetary policy. *J. Econom. Perspect.* **11**(1): 33-49.

Syverson C (2004) Product substitutability and productivity dispersion. *Review of Economics and Statistics* **86**(2): 534-550.

Syverson C (2011) What determines productivity? *Journal of Economic Literature* **49**(2): 326-365.

Tan TF, Netessine S (2014) When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Sci.* **60**(6): 1574-1593.

Tang CS (2006) Robust strategies for mitigating supply chain disruptions. *International Journal of Logistics: Research and Applications* **9**(5): 33-45.

Thirumalai S, Singha KK (2011) Product recalls in the medical device industry: An empirical exploration of the sources and financial consequences. *Management Science* **57**(2): 376-392.

Thomas JK, Zhang H (2002) Inventory changes and future returns. *Review of Accounting Studies* **7**: 163-187.

Ton Z, Huckman RS (2008) Managing the impact of employee turnover on performance: The role of process conformance. *Organ. Sci.* **19**(1): 56-68.

Ton Z, Raman A (2010) The effect of product variety and inventory levels on retail store sales: A longitudianl study. *Production Oper. Management* **19**(5): 546-560.

Villas-Boas JM, Winer RS (1999) Endogeneity in brand choice models. *Management Sci.* **45**(10): 1324-1338.

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

Wood RE (1986) Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes* **37**: 60-82.