International Series in
Operations Research & Management Science

Narendra Agrawal
Stephen A. Smith *Editors*

Retail Supply Chain Management

Quantitative Models and Empirical Studies

Second Edition

Springer

International Series in Operations Research & Management Science

Volume 223

More information about this series at http://www.springer.com/series/6161

Narendra Agrawal • Stephen A. Smith
Editors

Retail Supply Chain Management

Quantitative Models and Empirical Studies

Second Edition

Springer

Editors
Narendra Agrawal
Department of Operations Management
and Information Systems
Leavey School of Business
Santa Clara University
Santa Clara, CA, USA

Stephen A. Smith
Department of Operations Management
and Information Systems
Leavey School of Business
Santa Clara University
Santa Clara, CA, USA

ISSN 0884-8289 ISSN 2214-7934 (electronic)
International Series in Operations Research & Management Science
ISBN 978-1-4899-7561-4 ISBN 978-1-4899-7562-1 (eBook)
DOI 10.1007/978-1-4899-7562-1

Library of Congress Control Number: 2015934150

Springer New York Heidelberg Dordrecht London

Printed on acid-free paper

Springer Science+Business Media LLC New York is part of Springer Science+Business Media (www.springer.com)

Foreword

It is with great pleasure that I can write a foreword for the second edition of the book *Retail Supply Chain Management*: *Quantitative Models and Empirical Studies*. I want to congratulate the editors, Narendra Agrawal and Stephen Smith, for compiling this impressive volume. Like its first edition, this volume continues to be a book that provides a solid reference on research on retail supply chains and inspires new research on this subject.

Retailing forms the part of the supply chain that interfaces between the ultimate consumers and the rest of the supply chain. As such, it is often viewed as the part of the supply chain where the real demands of the consumers first show up. Whether we are talking about a physical retail store or a virtual store, the consumer demands that occur here drive the demands in the rest of the supply chain. So in that sense, it is like the frontier of all supply chains.

It is therefore gratifying to see Naren and Steve focusing their volume on retail supply chains. The innovations, lessons in practice, and new technological solutions in managing retail supply chains are not just important in retailing but crucial in the ultimate effective management of the complete supply chain.

There are two distinguishing features in the research of retail supply chains, which the current volume captures well. First, retail supply chains are loaded with a lot of empirical data. This is an area that has traditionally been rich in data, which provides fertile grounds for us to pursue empirical research. Second, research on retail supply chains naturally intersects with research in marketing in two ways—category management and pricing. Of course, category management and pricing have traditionally been key areas in the marketing literature. But what the current volume has added is the dimension of supply chain management to these marketing approaches. Integrating category management with inventory planning and coordinating price optimization with supply chain management are unique dimensions that distinguish this book.

The second edition expanded on the distinguishing features of the previous one with new analytics on data accuracy and visibility, retail workforce management, and business models of fast fashion. These are topics that are both timely and critical to successes in retailing.

I am sure that the readers will share my great enthusiasm for this book as a wonderful addition to the emerging literature of retail supply chain management.

Thoma Professor of Operations, Information and Technology, Graduate School of Business, Stanford University
Stanford, CA, USA
Hau L. Lee

Preface

We began working in retail supply chain management through the retail research program of the Retail Management Institute (RMI) at Santa Clara University. RMI was founded in 1980 by its current Executive Director, Dale Achabal, who is the L.J. Skaggs Distinguished Professor of Marketing at Santa Clara University. Research at RMI has focused on marketing and supply chain decisions in department store chains and specialty retailers. Over 30 major retail chains have participated in our research by providing data and problem descriptions and by sponsoring projects. The goal of our research has always been to develop new analytical tools for supporting the operational and planning decisions that retailers face. The sponsoring organizations saw the potential benefit from developing new analytical methodologies that could take advantage of the capabilities offered by emerging information technologies in retailing. Consequently, a number of the decision support prototypes developed at RMI were later converted into operational software systems by consulting organizations and application software products by independent vendors. In this sense, the research done at RMI, as well as the research by other authors of chapters in this volume, has led to an array of retailing applications that constitute a great success story for management science and for supply chain management in particular.

We are very pleased to present the second edition of our book following the tremendous success of the first one. This has provided the authors an opportunity to update their contributions to include the most recent developments in our field since 2009 when the first edition was published. We have also added three new chapters on recent topics which reflect areas of great interest and relevance to the academic and professional communities alike. These topics are fast fashion retail strategies, decision making in the presence of inventory record inaccuracies, and retail workforce scheduling. We hope that the new edition will serve as a useful resource for academic researchers and practitioners who are looking for the state of the art on studies on the topic of retail supply chain management.

We are grateful to all authors who have contributed their research to this endeavor and thank them for their patience as we went through multiple rounds of the review process for their submissions. We are indebted to our colleagues who painstakingly reviewed the various revisions of the submissions, adhering to standards typical of professional journals. These reviewers include Goker Aydin (Indiana University), Gerard Cachon (University of Pennsylvania), Nicole DeHoratius (University of Chicago), Vishal Gaur (Cornell University), Warren Hausman (Stanford University), Kirthi Kalyanam (Santa Clara University), Gürhan Kök (Koç University), Steven Nahmias (Santa Clara University), Marcelo Olivares (Columbia University), Andy Tsay (Santa Clara University), and Jin Whang (Stanford University). Finally, we would like to thank Gary Folven, our original editor with Kluwer and later with Springer Publishing, who encouraged us to undertake this project and supported our efforts. Matthew Amboy, the current Editor (Business & Economics: OR & MS) at Springer, has also been extremely supportive of our work and patient with our schedules.

We wish to thank our colleagues from the Marketing Department, Dale Achabal, Shelby McIntyre, and Kirthi Kalyanam, for collaborating with us on a wide range of projects. Many retail executives from sponsoring companies have contributed immensely to our research. There are simply too many for us to acknowledge individually but we are very grateful for their continued support. And we are especially grateful to our wives, Niti Agrawal and Karen Graul, and our children, Nishant and Nihar Agrawal, and Greg and Daniel Smith, for graciously supporting us during the time it took to complete this volume.

Santa Clara, CA, USA

Narendra Agrawal
Stephen A. Smith

Contents

1 **Overview of Chapters** 1
Narendra Agrawal and Stephen A. Smith

2 **Supply Chain Planning Processes for Two Major Retailers** 11
Narendra Agrawal and Stephen A. Smith

3 **The Effects of Firm Size and Sales Growth Rate on Inventory Turnover Performance in the U.S. Retail Sector** 25
Vishal Gaur and Saravanan Kesavan

4 **The Role of Execution in Managing Product Availability** 53
Nicole DeHoratius and Zeynep Ton

5 **Analytics for Operational Visibility in the Retail Store: The Cases of Censored Demand and Inventory Record Inaccuracy** 79
Li Chen and Adam J. Mersereau

6 **An Overview of Industry Practice and Empirical Research in Retail Workforce Management** 113
Saravanan Kesavan and Vidya Mani

7 **Category Captainship Practices in the Retail Industry** 147
Mümin Kurtuluş and L. Beril Toktay

8 **Assortment Planning: Review of Literature and Industry Practice** 175
A. Gürhan Kök, Marshall L. Fisher, and Ramnath Vaidyanathan

9 **Fast Fashion: Business Model Overview and Research Opportunities** 237
Felipe Caro and Victor Martínez-de-Albéniz

10 Managing Variety on the Retail Shelf: Using Household Scanner Panel Data to Rationalize Assortments 265
Ravi Anupindi, Sachin Gupta, and M.A. Venkataramanan

11 Optimizing Retail Assortments for Diverse Customer Preferences 293
Stephen A. Smith

12 Multi-location Inventory Models for Retail Supply Chain Management 319
Narendra Agrawal and Stephen A. Smith

13 Manufacturer-to-Retailer Versus Manufacturer-to-Consumer Rebates in a Supply Chain 349
Goker Aydin and Evan L. Porteus

14 Clearance Pricing in Retail Chains 387
Stephen A. Smith

15 Markdown Competition 409
Seungjin Whang

Index 425

Contributors

Narendra Agrawal Department of Operations Management and Information Systems, Leavey School of Business, Santa Clara University, Santa Clara, CA, USA

Victor Martínez-de-Albéniz IESE Business School, University of Navarra, Barcelona, Spain

Ravi Anupindi Technology and Operations, Stephen M. Ross School of Business, University of Michigan, Ann Arbor, MI, USA

Goker Aydin Kelley School of Business, Indiana University, Bloomington, IN, USA

Felipe Caro UCLA Anderson School of Management, Los Angeles, CA, USA

Li Chen Fuqua School of Business, Duke University, Durham, NC, USA

Nicole DeHoratius Booth School of Business, University of Chicago, Chicago, IL, USA

Marshall L. Fisher The Wharton School, University of Pennsylvania, Philadelphia, PA, USA

Vishal Gaur Johnson Graduate School of Management, Cornell University, Ithaca, NY, USA

Sachin Gupta Johnson Graduate School of Management, Cornell University, Ithaca, NY, USA

Saravanan Kesavan Kenan-Flagler Business School, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

A. Gürhan Kök College of Administrative Sciences and Economics, Koç University, Istanbul, Turkey

Mümin Kurtuluş Owen Graduate School of Management, Vanderbilt University, Nashville, TN, USA

Vidya Mani Supply Chain and Information Systems, Penn State Smeal College of Business, University Park, PA, USA

Adam J. Mersereau Kenan-Flagler Business School, University of North Carolina, Chapell Hill, NC, USA

Evan L. Porteus Graduate School of Business, Stanford University, Stanford, CA, USA

Stephen A. Smith Department of Operations Management and Information Systems, Leavey School of Business, Santa Clara University, Santa Clara, CA, USA

L. Beril Toktay College of Management, Georgia Institute of Technology, Atlanta, GA, USA

Zeynep Ton MIT Sloan School of Management, Cambridge, MA, USA

Ramnath Vaidyanathan Desautels Faculty of Management, McGill University, Montréal, QC, Canada

M.A. Venkataramanan School of Business, Indiana University, Bloomington, IN, USA

Seungjin Whang Graduate School of Business, Stanford University, Stanford, CA, USA

Editor's Biography

Narendra Agrawal is the Benjamin and Mae Swig Professor in the Department of Operations Management & Information Systems and has served on the faculty since 1992. He currently serves as the Associate Dean of Faculty. He holds an undergraduate degree in Mechanical Engineering from the Indian Institute of Technology, B.H.U., India; M.S. in Management Science from the University of Texas at Dallas; and an M.A. and Ph. D. in Operations and Information Management from The Wharton School of Business, University of Pennsylvania.

Naren's research is in the areas of supply chain management, service supply chain management, and manufacturing competitiveness. He has published his research in journals such as *Harvard Business Review*, *IIE Transactions*, *Interfaces*, *Journal of Retailing*, *Manufacturing & Service Operations Management*, *Naval Research Logistics*, *Operations Research*, *and Production and Operations Management* and has contributed chapters in a number of books. He is as an Associate Editor for *Manufacturing and Service Operations Management*, serves on the editorial review board of *Production and Operations Management*, and has been an Associate Editor for *Management Science*. Naren has received a number of teaching awards including the *Dean's Award for Teaching Excellence* (at Santa Clara University) every year since 1996 and the *MBA Core Curriculum Teaching Award* at The Wharton School. He has conducted numerous management development seminars internationally and consulted with companies in the retail and high-technology industries, including AAFES, Adaptec, Barco, The Gap, Hewlett Packard, IBM, KLA-Tencor, ONGC (India), Overstock.com, Pemex (Mexico), Schlumberger, and Silicon Image. He has been an advisor to several Silicon Valley start-ups and is a trustee of Give2Asia, a nonprofit organization that promotes philanthropy to Asia.

Stephen A. Smith is Professor of Operations Management and Information Systems in the Leavey School of Business at Santa Clara University, where he served as the Director of Research for the Retail Workbench and Education Center. He received a Ph.D. in Engineering-Economic Systems from Stanford University

and Bachelor and Master of Science Degrees in Mathematics. Before joining Santa Clara, Professor Smith was a Research Scientist at the Xerox Palo Alto Research Center and was previously a Principal of Pricing Strategy Associates, a consulting partnership. He has received the University Award for Sustained Excellence in Scholarship and was also awarded a Faculty Senate Professorship. He has served in various editorial positions for *Operations Research*, *Management Science*, *Manufacturing and Service Operations Management*, *Industrial Engineering Transactions*, and *International Commerce Review*.

Professor Smith's current research focuses on inventory and pricing decisions in retail supply chains. He is an author of over 60 research publications, which have appeared in a variety of journals including *Management Science*, *Operations Research*, *Marketing Science*, *Journal of Retailing*, *Journal of Marketing Research*, *Econometrica*, and *Journal of Economic Theory*, and two books: *Service Opportunities for Electric Utilities* and *Retail Supply Chain Management*, which was first published in 2009. He has consulted for a variety of retailers including Target, The Gap, and Levi Strauss and served as a technical advisor for three software start-up companies that developed decision support systems for retailers.

Chapter 1
Overview of Chapters

Narendra Agrawal and Stephen A. Smith

1 Background

The retail industry has emerged as a fascinating choice for researchers in the field of supply chain management. It presents a vast array of stimulating challenges that have long provided the context of much of the research in the area of operations research and inventory management. However, in recent years, advances in computing capabilities and information technologies, hyper-competition in the retail industry, emergence of multiple retail formats and distribution channels, an ever increasing trend towards a globally dispersed retail network, and a better understanding of the importance of collaboration in the extended supply chain have led to a surge in academic research on topics in retail supply chain management. Many supply chain innovations (e.g., vendor managed inventory) were first conceived and successfully validated in this industry, and have since been adopted in others. Conversely, many retailers have been quick to adopt cutting edge practices that first originated in other industries.

However, for every example of leading edge progressive thinking among retailers, there are numerous examples of archaic systems and planning processes. Moreover, there continue to be a host of open problems facing practitioners and academics. All of this is, of course, good news for academics engaged in research in retail supply chain management. The recent past has witnessed exciting new research—theoretical as well as applied—aimed at addressing some of the retail industry's many pressing challenges. This book is an attempt to summarize some of this research, as well as a perspective on what new applications may lie ahead.

N. Agrawal (✉) • S.A. Smith
Department of Operations Management and Information Systems, Leavey School of Business, Santa Clara University, 500 El Camino Real, Santa Clara, CA 95053, USA
e-mail: nagrawal@scu.edu

N. Agrawal, S.A. Smith (eds.), *Retail Supply Chain Management*,
International Series in Operations Research & Management Science 223,
DOI 10.1007/978-1-4899-7562-1_1

The past 20 years have seen a revolution in retailer's computing capabilities. Circa 1990, retailers' information systems tracked and stored dollar receipts for their merchandise, but often retained only cumulative sales data, as opposed to the selling patterns for individual SKUs by time period. Merchandise planners had access to various kinds of product level financial and inventory count information through computer terminals connected to the corporate data base systems. But there was no computing technology capable of applying quantitative forecasting and inventory management methods to evaluate alternative strategies, analyze market sensitivity to assumptions or optimize buying, promotions and clearance markdown decisions.

Since that time, the technology required to implement these methodologies has become widely available to buyers and inventory control analysts, as retailers have greatly expanded the information captured in their data bases and have distributed networked PCs to their professional employees. Retailers today can choose from a variety of commercial products that perform sales forecasting, pricing and inventory management functions, integrated as modules in their corporate information systems. Networked personal computers allow access to detailed sales and financial information, as well as offering localized processing power to analyze certain types of decisions. While the analytical methods imbedded in today's commercial offerings may appear to be fairly simple by academic standards, retailers' increasing investment and reliance on these systems indicates that they are providing value to retail supply chain operations today.

There is a natural development path for academic research in supply chain management to find its way into general use by major retailers. A number of the authors of the chapters in this volume have been instrumental in the successful implementation of methodologies for retailers. For purposes of illustration, let us consider the typical steps leading the implementation of a new methodology developed at the Retail Workbench at Santa Clara University. First, working with a sponsoring retailer, a decision support prototype is designed and developed for testing by buyers or other analysts in the merchandise planning cycle. Successful decision support prototypes were then adapted into an operational system by a third party software company or consulting organization, that works in cooperation with the sponsoring retailer. Finally, if market demand is perceived to be large enough, the one of a kind operational system is transformed into a commercial software product to be sold by an independent software vendor. It is hoped that many of the methodologies presented in this volume will find their way into mainstream retail practice through such a process as well.

2 The Focus of Academic Research in this Volume

Despite the advances in analytical applications discussed in the preceding section, retailers today face many important unsolved problems in supply chain management. The chapters in this book focus on three crucial areas of retail supply chain management in which academic researchers have been very active recently: (1) empirical studies of retail supply chain practices, (2) assortment and inventory

planning and (3) integrating price optimization into retail supply chain decisions. There are clearly other important research areas related to retail supply chain management, but in these three areas, recent research has successfully addressed some problems, while significant challenges remain.

2.1 Empirical Studies of Retail Supply Chain Practices

Chapter 2 (Agrawal and Smith), begins with a description of supply chain practices and processes observed at two retailers in the home furnishing sector. Because of the large number of stock-keeping-units (SKUs), the inter-relationships among the SKUs, as well as use of multiple store formats and multiple marketing channels targeted to different customer segments, home furnishings is one of the most complex retail sectors. In addition to documenting the complex flows of materials and information in such multi-channel environments, we present details of key supply chain planning processes: product design and assortment planning, sourcing and vendor selection, logistics planning, distribution planning and inventory management, clearance and markdown optimization, and cross-channel optimization.

Due to its complexity, we believe that the assortment selection and supply chain management decisions for this sector pose many challenging problems, whose solutions extend beyond the current state of the art. At the same time, the challenges in this sector are relevant to many other retail sectors as well. Thus, we hope that documenting the practices for these supply chains will provide a foundation for future methodological research, some of which are identified in the chapter.

Product level inventory management has been the subject of numerous papers in the area of supply chain management. More recently, researchers have begun to evaluate empirical evidence regarding the relationship between inventory management and overall firm performance. Some past research shows that inventory turnover varies substantially across firms as well as over time. Gaur et al. (2005) demonstrate that a significant portion of this variation can be explained by gross margin, capital intensity, and sales surprise (the ratio of actual sales to expected sales for the year). Using additional data, in Chap. 3, Gaur and Kesavan confirm these previously published results. Extending the findings of Gaur et al. (2005), they investigate the effects of firm size and sales growth rate on inventory turnover using data for 353 public listed US retailers for the period 1985–2003. With respect to size, they find strong evidence of diminishing returns to size: inventory turnover increases with size at a slower rate for large firms than for small firms. With respect to sales growth rate, they find that inventory turnover increases with sales growth rate, but its rate of increase depends on firm size and on whether sales growth rate is positive or negative. Their results are useful in (1) helping managers make aggregate-level inventory decisions by showing how inventory turnover changes with size and sales growth, (2) employing inventory

turnover in performance analysis, benchmarking and working capital management, and (3) identifying the causes of performance differences among firms and over time.

In Chap. 4, de Horatius and Ton direct attention to store level performance. In order to ensure product availability in retail settings, most existing research in this area has focused on two factors—poor assortment and poor inventory planning. The authors' research with several retailers during the last few years highlights a third factor, poor execution, or the failure to carry-out an operational plan. Poor store execution leads to stock outs and distorts sales and inventory data that are important inputs to assortment and inventory planning.

In this chapter they focus on two common execution problems—inventory record inaccuracy and misplaced products. Drawing on well-researched case studies, they describe the magnitude and root causes of these problems. They also describe the findings of empirical studies that have identified factors that exacerbate the occurrence of these problems. These factors include product variety, inventory levels, employee turnover and training, employee workload and employee effort. They describe the effect of inventory record inaccuracy and misplaced products on inventory planning and summarize how researchers have incorporated these problems into existing inventory models. They also discuss future research opportunities for studying the impact of store execution on product availability, in particular, and on retail supply chains, in general.

In Chap. 5, Chen and Mersereau take a detailed look into the literature on two established streams of OM research that try to overcome one of the key shortcomings noted by de Horatius and Ton—lack of visibility into operational data. The first is demand estimation and inventory optimization in the presence of data censoring, where imperfect data may cause significant estimation biases and inventory cost inefficiencies. The second is inventory record inaccuracy, where intelligent replenishment and inspection policies may be able to reduce inventory management costs even without real-time tracking technologies like radio frequency identification (RFID). Common themes of these literatures are that lack of visibility can be costly if not properly accounted for, that intelligent analytical approaches can potentially substitute for visibility provided by technology, and that understanding the best possible policy without visibility is needed to properly evaluate visibility technologies. The authors include a survey of modern and emerging visibility technologies and a discussion of several new avenues for analytical research.

In recent years, the focus of retail operations and supply chain management literature has also begun to include practices and decision making tools that focus on a key element of any brick-and-mortar retail store—store associates. Retail store associates are frontline employees of retail organizations and are responsible for delivering superior in-store experience to its customers. Store associates provide customer service through direct interaction with customers as well as through indirect means such as maintaining a clean store and ensuring that the shelves are fully stocked. While labor is critical to drive store sales, it needs to be planned for carefully as it is one of the largest expenses for retailers. Therefore, retailers deploy workforce management solutions to balance their need for labor to drive sales against their need

to control store expenses to improve profitability. While labor planning is not a new decision for retailers, there continue to be considerable differences in the way labor planning is performed in the retail industry. These approaches differ in the level of sophistication used to manage the payroll and the degree to which different departments within a retail organization are involved in labor planning. In Chap. 6, Saravanan and Mani provide an overview of the literature on workforce management in the retail industry and survey the empirical research on this topic. They discuss some of the new technologies that have the potential to shape this aspect of the retail landscape, and conclude with directions for future research.

In addition to scientific inventory management and keen attention to execution of operational policies, leading edge retailers are resorting to other innovative management practices. In Chap. 7, Kurtulus and Toktay discuss one interesting example from the consumer goods sector, called category captainship. It is a form of manufacturer-retailer collaboration in which retailers rely on a leading manufacturer for management of items in a given category. There are reported success stories about category captainship, but also a growing debate about its potential for creating anti-competitive practices by category captains. The goal of this chapter is to provide an overview of the existing research on category captainship.

Despite a decade of implementation, there is limited academic research concerning category captainship. The existing research on captainship can be grouped into four broad categories that aim to answer the following questions: (1) What are the consequences of the retailer delegating the pricing decision to a category captain? (2) What are the consequences of the retailer delegating the assortment selection decision to a category captain? (3) When will category captainship emerge? What are the category characteristics that facilitate the emergence of category captainship? (4) What are the antitrust concerns that may arise as a result of using category captains for category management? What can be done to mitigate these antitrust concerns? The limited research in this field is due to challenges arising from the broad scope of implementation of category captainship programs. This chapter reviews the current research on category captainship and proposes some avenues for future research that could potentially overcome these challenges and improve our understanding of category captainship practices. The chapter also sheds light on how category captainship practices could potentially change the nature of the manufacturer-retailer relationships and the landscape in the retail industry.

2.2 Assortment and Inventory Planning

The assortment a retailer carries has a significant impact on sales, margins and customer traffic. Therefore, assortment planning has received high priority from retailers, consultants and software providers. The academic literature on assortment planning from an operations perspective is relatively new, but quickly growing. The basic assortment planning problem focuses on choosing the optimal set of products to be carried and the inventory level of each product. Decisions for products are

interdependent and complex, due to considerations such as shelf space availability, substitutability between products, and brand management by vendors.

An in depth review of the research on this topic is presented by Kok, Fisher and Ramnath in Chap. 8. This chapter is composed of four main parts. In the first part they discuss empirical results on consumer substitution behavior and present three demand models used in assortment planning: the multinomial logit, exogenous demand and locational choice models. In the second part, they describe optimization based assortment planning research. In the third part, they discuss demand and substitution estimation methodologies. In the fourth part, they present industry approaches to assortment planning by describing the assortment planning process at four prominent retailers. The authors conclude by providing a critical comparison of the academic and industry approaches and identifying research opportunities to bridge the gap between the two approaches.

One of the most fascinating recent developments in the apparel retail industry is the emergence of *fast fashion* retailers—companies such as H&M, Zara, Uniqlo, Mango, etc. These companies have thrived by offering fashionable designs at affordable prices, frequent assortment changes, and quick response to changes in their markets. In an industry that is often characterized by extremely long and inefficient supply chains—product concept to end of sales cycles of up to 18 months is not uncommon—these companies can introduce new products in a matter of days. What makes it even more impressive is the fact that these new designs are created in response to observed consumer choices.

Caro and Martinez-de-Albeniz examine the underlying business model of such companies from an operations perspective in Chap. 9. In particular, they describe the key operational competencies that such firms must develop, and present a survey of the literature on methodologies for making several important supply chain planning and control decisions. The paper also points to several open questions that are interesting opportunities for future research.

In Chap. 10, Anupindi, Gupta and Venkatraman present a specific optimization methodology for the rationalization of retail assortment and stocking decisions for retail category management. They assume that consumers are heterogeneous in their intrinsic preferences for items and are willing to substitute less preferred items to a limited extent if their preferred items are not available. The authors propose an objective function for a far-sighted retailer that includes not only short-term profits but also a penalty for disutility incurred by consumers who do not find their preferred items in the available assortment. The retailer problem is formulated as a constrained integer programming problem. They demonstrate an empirical application of their proposed model using household scanner panel data for eight items in the canned tuna category. Their results indicate that the inclusion of the penalty for disutility in the retailer's objective function is informative in terms of choosing an assortment to carry. They find that customer disutility can be significantly reduced at the cost of a small reduction in short term profits. They also find that the optimal assortment behaves non-monotonically as the weight on customer disutility in the retailer's objective function is increased.

Smith, in Chap. 11, considers an assortment planning model for retailers who sell multi-featured products such as consumer electronics and must tailor their

assortments to appeal to a diverse set of customer tastes. The assortment decision affects both the probability that customers choose a particular retailer and the demands for the various products in the retailer's assortment. By explicitly including diverse customer segments, this paper develops an operational methodology for optimizing retail assortments for heterogeneous product preferences. A multinomial logit model is used for computing customers' joint probabilities of retailer choice and product choice. An optimization problem is then formulated for determining the assortment that maximizes the retailer's expected profit. The relationship between the optimal assortment and the retailer's competitive strength is also analyzed. Limiting properties of the relationship are derived for the special cases of a monopoly retailer and perfect competition among retailers. A commercial data base of consumer preferences for DVD players is used to illustrate the assortment optimization methodology and the sensitivity to various input assumptions. It was found that including customer heterogeneity in the choice model had a significant impact on expected profits for this data set.

The assortment planning decision is tightly connected to the inventory planning decision, about which there is extensive literature in the field of operations management. However, much of this literature assumes that the assortment has already been specified, and focuses solely on the inventory management decision. In Chap. 12, Agrawal and Smith provide a review of some recent research on multi-location inventory that is related to retail supply chain management.

In order for the review to be meaningful, it is restricted in scope in a number of ways. First, the focus is on papers that model multi-level inventory systems, since virtually all retail supply chains are multi-level. Second, attention is restricted to papers after 1993, and the reader is referred to the reviews in other papers for articles prior to 1993. For example, Axsater (1993), Federgruen (1993), and Nahmias and Smith (1993) contain excellent reviews of the work up to that point. Third, certain model formulations that are not typical of retail inventory management are also excluded, such as serial systems, since they are not representative of typical retail chains, and are a special case of general multi-location multi-echelon systems. Also excluded are papers that assume deterministic demand, since demand uncertainty is a key aspect of most retail systems.

Finally, the primary focus is on periodic review systems. Most retail chains today employ technologies such as point-of-sale (POS) scanner systems that provide real time access to sales and inventory data. Consequently, in principle, continuous review models could be an appropriate construct for these retail systems. However, two issues limit the practical applicability of this assumption. First, due to contracts with vendors and shipping companies, shipments occur primarily on a pre-specified schedule, and often a variety of items are delivered simultaneously. Second, despite the real time access to sales information, the ERP databases and inventory allocation algorithms are typically updated periodically. Thus, strictly speaking, inventory decisions must be made by planners according to predefined cycles. Thus, periodic review systems are a better representation of the inventory management systems used by most retailers. They conclude with suggestions for future research in this area.

2.3 Integrating Price Optimization into Retail Supply Chain Decisions

In addition to more efficient operational decisions, recent research has shown that better designed incentive systems can also be very effective in improving the operational and financial performance of supply chains. These incentive systems are captured in the supply chain contracts that define the relationship between buyers and suppliers. Reviews of some of the supply chain literature that focuses on the design of these contracts are contained in Tsay et al. (1999) and Cachon (2003).

In Chap. 13, Aydin and Porteus study the effect of the type of rebate offered to customers on the performance of the supply chain, and on the preference of the manufacturer and the retailer for such rebates. Starting with a newsvendor model (single-product, single-period, stochastic demand), they build a single-retailer, single-manufacturer supply chain with endogenous manufacturer rebates and retail pricing. The demand uncertainty is multiplicative, and the expected demand depends on the effective (retail) price of the product. A retailer rebate goes from the manufacturer to the retailer for each unit it sells. A consumer rebate goes from the manufacturer to the consumers for each unit they buy. Each consumer's response to consumer rebates is characterized by two exogenous parameters: α, the effective fraction of the consumer rebate that the consumer values, leading to the lower effective retail price perceived by the consumer, and, β, the probability that a consumer rebate will be redeemed. The type(s) of rebate(s) allowed and the unit wholesale price are given exogenously. Simultaneously, the manufacturer sets the size of the rebate(s) and the retailer sets the retail price. The retailer then decides how many units of the product to stock and the manufacturer delivers that amount by the beginning of the selling season. Compared to no rebates, an equilibrium retailer rebate leads to a lower effective price (hence, higher sales volume) and higher profits for both the supply chain and the retailer. An equilibrium consumer rebate also leads to a lower effective price and higher profits for the retailer, but not necessarily for the chain. Under their assumptions, such a consumer rebate (with or without a retailer rebate) allocates a fixed fraction of the (expected) supply chain profits to each player: The retailer gets $\alpha/(\alpha+\beta)$ and the manufacturer gets the rest, leading to interesting consequences. However, both firms prefer a higher α and a lower β, even though the manufacturer gets a smaller share of the chain profits, the total amount received is higher. Neither the retailer nor the manufacturer always prefers one particular kind of rebate to the other. In addition, contrary to popular belief, it is possible for both firms to prefer consumer rebates even when all such rebates are redeemed.

Another important aspect of pricing that has received some attention in the operations management literature is markdown planning, i.e., the price charged by the retailer at the end of the season to clear leftover inventory. This is important financially for retailers, since studies by the National Retail Federation have found that over one third of merchandise is sold on markdowns in some retail chains. Clearance markdowns are the focus of Chap. 14 by Smith. In the basic newsvendor model, the salvage value (which is related to the markdown price) is assumed to be

fixed, but, in practice, this will depend upon the retailer's markdown pricing strategy. As the season draws to a close, sales rates depend upon price, seasonal effects and the remaining assortment of items available to customers. There is little time to react to observed sales, and pricing errors result in either loss of potential revenue or excess inventory to be liquidated. This chapter develops optimal clearance price trajectories and inventory management policies that take into account the impact of reduced assortment and seasonal changes on sales rates. Versions of these policies have been implemented and tested at a number of major retail chains and these results are summarized and discussed.

Finally, in Chap. 15, Whang extends the markdown strategy discussion by including the element of retailer competition, using a stylized model of markdown competition. He considers two retailers who compete in a market with a fixed level of initial inventory. The initial inventory level is known to one retailer, but not to the other. To maximize the profit, each retailer marks down at a time of his individual choice. The model assumes deterministic demands, a single chance of price change, and a prefixed set of prices. He considers a two-parameter strategy set where a retailer chooses the timing of markdown as a function of the current time, his inventory level and the other retailer's actions so far. The paper characterizes the equilibrium of the game and derives managerial insights.

Retail supply chain management is a relatively new but very exciting field of research. Fortunately, there is a substantial body of research in the areas of traditional inventory management, multi-echelon systems, channel coordination and pricing that the can be applied in the field of retailing. The challenge, of course, is to develop and adapt methodologies that most accurately reflect the realities and constraints faced by retailers. As the practice of retailing evolves at increasing speed because of changes in the global competitive landscape, technology, and consumer expectations, we expect the array of research challenges facing academics and practitioners to expand as well. We hope that this book will serve as a useful reference for these researchers, and look ahead to the evolution of this field with much anticipation.

References

Axsater, S. (1993). Continuous review policies for multi-level inventory systems with stochastic demand. In S. C. Graves, A. H. G. Rinnooy Kan, & P. H. Zipkin (Eds.), *Logistics of production and inventory* (Handbooks in operations research and management science, Vol. 4, pp. 175–197). Amsterdam, The Netherlands: Elsevier Science Publishing Company.

Cachon, G. (2003). Supply chain coordination with contracts. In S. C. Graves & A. G. De Kok (Eds.), *Handbooks in operations research and management science: supply chain management* (pp. 229–340). The Netherlands: Kluwer Academic Publishers.

Federgruen, A. (1993). Centralized planning models for multi-echelon inventory systems under uncertainty. In S. C. Graves, A. H. G. Rinnooy Kan, & P. H. Zipkin (Eds.), *Logistics of production and inventory* (Handbooks in operations research and management science, Vol. 4, pp. 133–173). Amsterdam: Elsevier. Ch. 3.

Gaur, V., Fisher, M. L., & Raman, A. (2005). An econometric analysis of inventory turnover performance in retail services. *Management Science, 51*(2), 181–194.

Nahmias, S., & Smith, S. A. (1993). Mathematical models of retailer inventory systems: a review. In R. K. Sarin (Ed.), *Perspectives in operations management* (pp. 249–278). Norwell, MA: Kluwer Academic Publishers.

Tsay, A. A., Nahmias, S., & Agrawal, N. (1999). Modeling supply chain contracts: a review. In S. Tayur, R. Ganeshan, & M. Magazine (Eds.), *Quantitative models for supply chain management* (pp. 299–336). Norwell, MA: Kluwer Academic Publishers.

Chapter 2
Supply Chain Planning Processes for Two Major Retailers

Narendra Agrawal and Stephen A. Smith

1 Introduction

This chapter provides descriptions of the supply chain structures and planning processes of two major retailers in the home furnishings sector. These descriptions are based on a series of interviews with senior executives at these two retailers. Our objective is not to provide a comprehensive survey of such retail firms, but rather to describe the structures and planning processes commonly found in this sector and the corresponding implications for supply chain management based on these two case studies.

Home furnishings is one of the most complex areas in retailing, because of the large number of stock-keeping-units (SKUs), the inter-relationships among the SKUs, as well as use of multiple brands and multiple marketing channels targeted to different customer segments. Due to its complexity, we believe that the assortment selection and supply chain management decisions for this sector pose many challenging problems, whose solutions extend beyond the current state of the art. Thus, we hope that documenting the practices for these supply chains will provide a foundation for future methodological research.

Since both companies requested that we not reveal their identities, we will refer to them as Companies A and B. A number of our observations about planning processes were similar at the two retailers. Also, as described later, Company A has a more complex supply chain because it is a multi-channel retailer. Thus, its structure and planning process are more general than Company B. Therefore, rather than presenting two separate case studies, we will discuss them simultaneously,

N. Agrawal (✉) • S.A. Smith
Department of Operations Management and Information Systems,
Leavey School of Business, Santa Clara University, Santa Clara, CA 95053, USA
e-mail: nagrawal@scu.edu

N. Agrawal, S.A. Smith (eds.), *Retail Supply Chain Management*,
International Series in Operations Research & Management Science 223,
DOI 10.1007/978-1-4899-7562-1_2

focusing primarily on Company A, while highlighting the differences at Company B.

Company A, with revenues of about $3.5 Billion per year, consists of six different retail brands or "concepts," with a total of the nearly 600 stores in over 40 states in the US. Each brand sells products through its own distinct set of retail stores. For example, while one brand focuses on casual home furnishings, another focuses on cookware essentials, and a third focuses on children's furnishings. In addition, Company A also operates direct-to-consumer channels, with eight different brands of catalogs and six different web sites. A true multi-channel retailer, this firm generates nearly 40 % of its revenues from its direct-to-consumer marketing channels.

Company B has yearly revenues of approximately $1 Billion, and operates roughly 300 stores, selling products in the casual home furnishings, housewares, gifts, decorative accessories categories. In contrast to Company A, this retailer is primarily a single channel retailer, selling mostly through stores. Its Internet channel was initiated very recently, and it does not have a catalog channel. Also, the great majority of its products are branded merchandise. Therefore, its supply chain structure is much simpler than Company A's. However, Company B generates a significant fraction of its revenue from foods and beverages, which present special challenges due to the perishable nature of these products.

The number of different SKUs is quite large for both retailers. Within their largest brand, Company A offers roughly 70,000 different SKUs at a given point in time. Company B operates smaller stores (about 18,000 square feet), with approximately 36,000 SKUs at each store. The SKUs are partitioned into categories, such as furniture, home accessories, table top accessories, food and decorative accessories. Within a category, strong demand interactions across SKUs could be expected to occur, e.g., many SKUs may complement or substitute for each other. SKUs across different categories would have weaker and less specific demand interactions. The products vary significantly in their prices, physical characteristics, prices, perishability, seasonality, procurement lead times and country of origin.

The assortment must address two key marketing objectives (1) providing customers with as complete an assortment as possible and (2) providing an assortment that creates attractive **presentations**. Since stores carry manufacturers' name brands, it is important to provide a comprehensive selection of related items within a given brand, e.g., Sheffield cutlery. Both retailers emphasized that "presentation drives demand" in each of the channels. Therefore, products are often displayed as they might actually appear in a customer's home for maximum advertising impact. In fact, some customers will purchase an entire room as displayed in the store, or will purchase the complete set of items in a tabletop display. In addition, the best types of items to feature in the catalog or Internet presentations may differ from those in the ideal store presentation. For example, a completely furnished room works well in a store, but would be difficult to capture photographically for a catalog. A large assortment of wall hangings shows well in a catalog, but would require too much wall space in a store.

The merchandise featured in each channel's presentation is, of course, only a small subset of the available merchandise. Store and catalog presentations are

modified as frequently as every 30 days depending on the seasons of the year. The products offered in the assortments change much less frequently than the presentations, with the majority of the SKUs continuing for at least 6 months or more. One rapidly changing type of SKU, known as "ornamentation," is seasonal and fashion driven, and thus the ornamentation assortment tends to change with the presentation. Also, some products may be discontinued in their original sales channel, but still continue to be offered through the outlet stores or Internet and catalog channels. Therefore, the presentation requirements lead to additional constraints on both the assortment planning process and the management of the supply chain.

Neither retailer optimizes supply chain costs as part of the product design and assortment selection process. Instead sourcing costs and financial outcomes are viewed as constraints, rather than primary objectives. Supply chain decisions are handled by a sourcing team, which is separate from the design and assortment selection team. In general, the sourcing team is responsible for managing the supply chain as effectively as possible for whatever assortment is chosen. If problems arise, the sourcing team does have some power to initiate assortment modifications later in the planning process, as we discuss in the next section. It is generally recognized that this partitioning of responsibilities is suboptimal, but the problem persists because of the complexity of the decisions.

We note that some of these characteristics of home furnishings supply chains are common to retailers in other areas, which indicates that the structures described here have broader significance. For example, The Gap, similar to Company A, sells its apparel and accessories through a number of different store concepts that include The Gap stores (including Gap Kids, Baby Gap, Gap Outlet and Gap Body), Old Navy, Banana Republic and Piper Lime. While The Gap focuses on casual and fashion apparel and accessories for men and women, Old Navy is positioned for the more value conscious consumer, and Banana Republic is positioned at price points that are higher than The Gap channel. Products are sold through retail stores and the Internet channel for each concept. Similarly, Target operates Target Stores, Mervyns and Dayton Hudson stores, which carry both private label brands and branded merchandise. Internet channels are also associated with each store concept at Target.

The objective of "presenting an attractive assortment" to the consumer is equally important to these retailers as well. For example, it is common practice to display complete apparel and accessory outfits from a given manufacturer, e.g., Ralph Lauren, both in stores and in the Internet channels. It is common knowledge across the retail industry that matching assortments that are displayed on the covers of catalogs, or displayed prominently in stores, generate a significantly larger level of sales than products stocked on shelves or racks. Thus the assortment selection and presentation design decisions are closely linked across many retail categories.

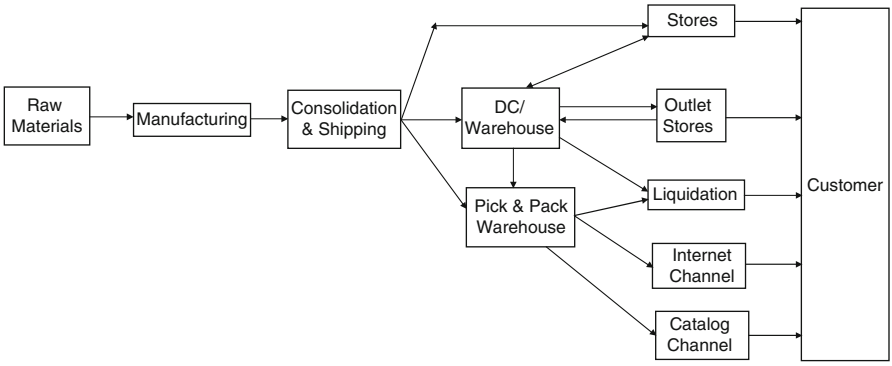


Fig. 2.1 Retail supply chain for Company A

2 Supply Chain Description

Company A's supply chain is illustrated in Fig. 2.1. While the supply chain varies somewhat across brands, this figure illustrates the most general case. The overlap across supply chains for the various brands is minimal and limited to sharing of warehouse space and merchandise handling capabilities at the distribution center (DC).

Since Company B is primarily a single channel retailer, its supply chain lacks the pick-&-pack warehouse, outlet stores, and the Internet and Catalog channels in the figure above.

Company A's products are sourced from both domestic and foreign suppliers. The foreign suppliers are located in 35 different countries, and are responsible for nearly two-thirds of the total merchandise purchased. A particular brand or concept that offers 60,000–70,000 stock-keeping-units SKUs may be sourced from as many as 1,000 different vendors. Nearly 60 % of the products are basics, which continue for at least two selling seasons. The planning calendar consists of four seasons, with the Fall season responsible for the majority of annual sales. Stores may carry both nationally known brands of products as well as private label products. Company B sources its products primarily from foreign vendors. It utilizes about 30 agents to obtain 36,000 SKUs from about 1,000 active vendors. 65–70 % of its furnishing products and almost 90 % of its food products are basic (its core products can have a selling season that is 2–10 years long). It too plans for four separate seasons over the year.

Shipping from foreign sources is primarily by boat, in large metal shipping containers. Containers destined for multiple stores need to be sent to a DC and unpacked. Company A, with the more complex supply chain, operates three such DCs. The largest facility, with nearly 6 million square feet of space, is located in Memphis. It provides replenishments for all the stores, as well the sourcing for the direct-to-consumer shipments for the Internet and catalog channels for all products other than furniture. Furniture, given its physical size, is distributed through two

separate distribution centers, one on the East coast and one on the West coast. Store-bound merchandise is then transferred to trucks for delivery. Direct-to-consumer shipments are handled by two independent shipping companies. Company B operates two DCs, one on each coast. Demand fulfillment for their Internet business, when it is ready, will occur from a separate, outsourced DC on the east coast.

Merchandise can also follow a variety of paths during the selling process. Store customers usually pick up items at the store. But bulky items such as furniture are displayed in the store, while deliveries take place directly from a DC/ warehouse to the customer. In order to combine customer orders and reduce trucking costs, customer delivery time may require a lead time of several weeks. Items that are direct shipped are handled by third party logistics (TPLs) companies and delivered to the customer. Similarly, non-conveyable items that are purchased through the Internet or Catalog channels may ship directly to the customer from the DC/ warehouse. Thus, multiple items that the customer purchases at the same time may be delivered in different ways and at different times. The same customer may also shop in different channels at various times. Thus, the customers' level of satisfaction with their overall shopping experience in one channel will influence their future purchases in other channels. This cross channel interaction is not currently considered in selecting inventory service levels.

Certain items in any channel may not sell as well as originally anticipated. Slow sellers or discontinued items in the stores are often sent to one of the retailer's outlet stores, and offered at a reduced price. The outlet channel may also be used for returned merchandise that the retailer does not wish to offer in the regular stores. Merchandise from the regular stores destined for the outlet stores is typically moved first to the DC, where it is consolidated and then allocated to the outlet stores based on their anticipated demands. In order to maintain an attractive presentation and selection in the outlets, about 30–40 % of the outlet merchandise for Company A is sourced specifically for outlets, and consists of items that are not offered in regular stores. Some items that are no longer carried in stores may continue to be offered through the Internet or Catalog channels. Since customers can retain catalogs for some time, orders will sometimes be filled for items that are no longer carried in the most recent catalog.

3 Supply Chain Planning Processes

Let us now turn our attention to the various planning processes in these supply chains. We begin by describing a typical planning calendar (Fig. 2.2), which can be 12–16 months long, and is implemented in a rolling horizon basis. Our description of this calendar is primarily based on our discussions with Company B, although the process is very similar at Company A.

While the details of these steps are presented subsequently, we note that the first key interaction between the merchandising team and supply chain planning team

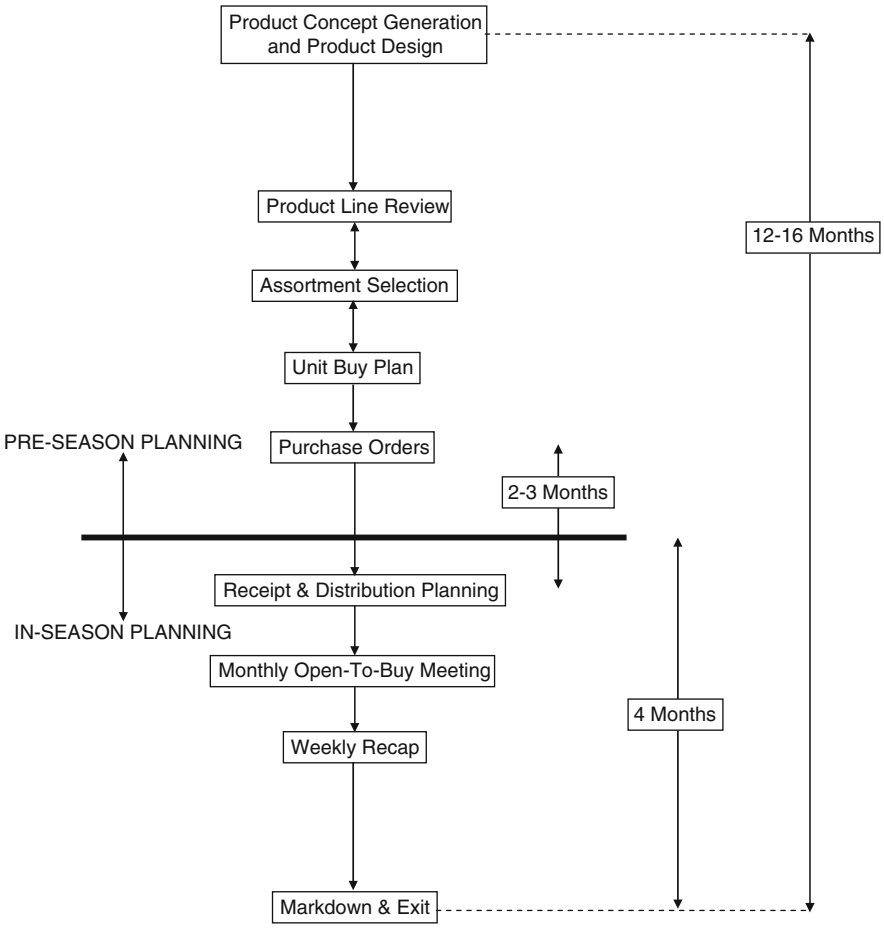


Fig. 2.2 Retail Master Calendar

occurs during the Assortment Selection step. As part of this step, not only do the teams formalize the assortment, they also perform financial analyses to determine whether the sales targets, specified in the company's financial plans, will be met. This is based on a top-down analysis of the sales forecasts. Following this, when the unit buy plans are created, the teams forecast unit sales at the different stores for given pricing policies. This is a bottom-up analysis. A very important step at this point is the reconciliation between the top-down and bottom-up predictions. This may lead to a revision of the company targets for sales and margins and/or modifications in the assortment. These targets are further reviewed at the monthly review meetings, and may be revised, along with targets for initial markup, inventory turns and markdowns.

Decision making in this process tends to consist of a series of "what if" analyses, with little reliance on analytical optimization. Moreover, the process of revising company targets involves addressing a number of tradeoffs, which is often done in a subjective manner. These decisions may be greatly influenced by personal incentives. For instance, if the unit buy plan turns out to exceed the financial targets, the teams would typically simply promise to meet the target, i.e., they would much rather perform better than predicted than to show a shortfall.

3.1 Product Design and Assortment Planning

Retailer A has a highly "vertical structure" with respect to its planning processes. The planners assigned to the various processes tend to be specific to each brand, with minimal overlapping responsibilities across brands. The percentage of private label merchandise is small in the retailer's flagship brand, while it is quite high in its other brands. Each brand has its own product design team. As a specific example, in one brand, 40 *product designers* search the world for new product designs and material concepts. Merchandise is divided into a number of different categories, each with its own design team and buyers. The designers present their ideas to the *merchants* and *sourcing* specialists during a product line review process, where they evaluate sketches and samples of products, and consider pricing decisions. Upon approval, these specs are then given to independent *sourcing agents*, spread across the world, who seek out the appropriate *vendors* for product prototypes.

Upon receipt of these prototypes, the merchants consider how the assortment as a whole will be presented to the consumer, and suggest appropriate modifications. This is a very important step in the process, since individual product decisions must be made subject to the constraints and limitations imposed by the assortment presentation. The assortment is also reviewed by the *visual and marketing* group, which specializes in creating store presentations. Finally, the products are adopted and handed over to the sourcing and *inventory teams*. The inventory team is responsible for producing high level forecasts, and determining if the product line can deliver its sales and revenue targets. Typically, the elapsed lead time from a new product's concept stage to delivery into the stores is about 12 months.

In this planning process, the central role in assortment decisions is played by merchants. The process architecture is illustrated in Fig. 2.3 below, where the merchants are at the hub. Product design groups within a brand tend to work all year round, since about a third of the SKUs tend to be new at Company A each year. The in-store presentation changes frequently, giving consumers the impression of a rapidly changing assortment. Catalogs are also shipped to consumers frequently with different assortments of featured merchandise, corresponding to the season of the year. As noted previously, the total assortment of products in each of these channels turns over much less frequently than the presentations. Finally, the product lines in the three marketing channels overlap somewhat, but each line also contains many unique products.

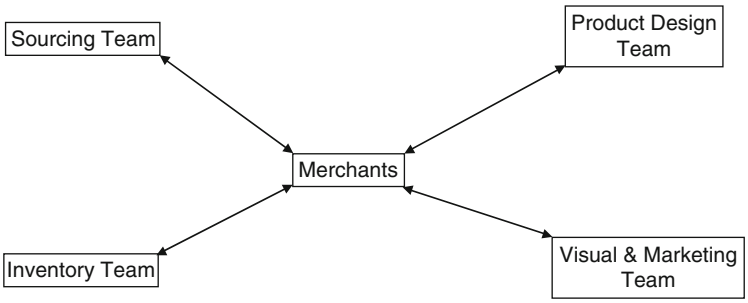


Fig. 2.3 Product design process architecture

3.2 Sourcing and Vendor Selection

As mentioned earlier, due to the retailer's vertical structure, each brand tends to have its own sourcing teams. This is a recognized weakness with regard to sourcing, since it does not exploit the potential synergies due to consolidation of buying across brands. As is typical of most retailers, Company A does not manufacture its own products. In fact, Company A manages most of its vendor interactions through independent agents, who are domain experts. These agents identify the vendors, ensure the ability of these vendors to execute purchase orders in a timely and financially sound manner, implement quality protocols in-line and for final goods, verify packaging, and determine the vendors' social compliance. Ensuring social compliance by vendors is becoming increasingly important for US based retailers, and continues to be a very difficult challenge.

The retailer evaluates the vendors primarily based on their past performance. Vendor evaluation score cards are selected for ongoing vendors, but no such metric is used to evaluate new vendors during the selection process. This retailer's sourcing organizations are generally not involved in the vendors' actual production planning process beyond shared forecasts and receiving purchase orders. This is in contrast to what we have observed at some other retailers who actively engage in the vendor's capacity planning (Agrawal et al. 2002) Also, because of capacity limitations, multiple vendors may sometimes be used for the same product.

The manufacturing process itself may take as long as 3–5 months to complete. But the manufacturing lead time can be as short as 30 days for products that consist primarily of upholstery or fabrics. The total order quantity for the merchandise is manufactured over a period of time and the goods are typically flowed to the retailer in multiple lots. For core products that are carried over multiple seasons, contracts often allow for modifications in order quantities within certain ranges, depending on the observed demand for the product.

3.3 Logistics Planning

As mentioned earlier, shipments from foreign sources are primarily in large metal shipping containers. The shipping time for a container, including delivery to the DC or directly to a store, is from 30 to 40 days, depending on the final destination. Ideally, a shipment to the retailer fills one or more containers exactly. This objective may influence lot size selection for both shipping and manufacturing. The allocation of merchandise across shipping containers can be quite complex. For example, it is highly desirable to have a dedicated shipping container that can be transported directly to a store. At the same time, stores have limited space and holding too much merchandise in the store at one time is not acceptable.

Planning the shipping needs for retailers is a complex but critical activity. For example, at Company B, logistics planning begins right after the merchandising plans are set for the following year. Unfortunately, merchandising plans do not specify how the percentage of imports relative to total purchases will change in the upcoming year. Nor do they specify how product inflows from particular countries may change. This information is important for logistics planning since securing shipping container capacity on specific freight lanes in a timely manner is critical to ensuring delivery reliability. This decision problem is dimensionally complex—Company B utilizes five different steamship lines and fills about 7,000 40-ft containers annually. In the absence of the detailed capacity requirements, retailers use rudimentary forecasting methods for planning purposes.

Based on these rough forecasts, retailers negotiate rates for shipments with shipping companies. Rate negotiations typically happen in February and March for shipments starting in May through the following April. Contracts typically specify the total number of containers that will be used, with guaranteed minimums, but not the actual timing of the shipments. Rates have been hard to predict in the recent past due to significant uncertainty in the cost of fuel. The average cost of shipping a full container to the US is $3,200, and partial container shipments incur roughly a 33 % cost premium.

Containers destined for multiple stores need to be sent to the DCs to be unpacked. The merchandise is then transferred to trucks for delivery to the stores, which also adds to the shipping time. Retailers typically set aggressive targets for transfer time in the DC, e.g., less than 24 h turnaround time. Depending on their country of origin and the quantity of items, some merchandise shipments do not fill a whole shipping container. In this case, the shipment is handled by local freight consolidators who pool shipments from multiple retailers. For these items, the retailer also needs to make arrangements for where the container will be unpacked and how the merchandise will be transported to its final destination. In order to facilitate shipping, the container requirements could thus potentially influence the retailer's choice of sourcing location or manufacturer. While the sourcing team at this retailer tries to deal with this problem subjectively, they do not consider the joint optimization of shipping and sourcing decisions in a systematic way.

The retailer also operates a "Pick and Pack" warehouse, where merchandise is "direct shipped" to customers from the Internet and Catalog channels. This requires special packaging that can be done at the manufacturing site. In some cases, the direct ship merchandise comes in larger packages that require additional set up for the automated pick and pack process at the warehouse. An important distinction is made between items that are "conveyable," i.e., can be put on a conveyer belt. Those that are not conveyable (items with very large dimensions or irregular shapes) cannot be handled by automated pick and pack equipment. Again, items shipped from the vendor to the pick and pack facility may not always fill a whole shipping container. In this case, they are combined with other retailers' merchandise by a consolidator, and later separated and trucked to the pick and pack warehouse.

Shipments from the DCs to stores are primarily by truck. This shipping time was as high as 10 days, but has now shrunk to 2–3 days because of the use of TPLs like UPS. Oversized packages that are not handled by UPS are sent via other independent shippers.

Interestingly, we learned at Company B that domestic shipping can be more onerous than international shipping because the trucking industry capacity in the US is unpredictable. We were told that from the retailers' perspective, the performance of the trucking industry seems to be negatively correlated with the state of the construction industry, because the better the construction industry does, the fewer drivers are available for the trucking industry. Reliability of truck drivers and availability of equipment (trucks) capacity is a constant challenge. Finally, since shipments by trucks often require multiple handoffs due to the hub-and-spoke system used by shippers, numerous errors in shipping information and damages to products are often introduced.

Appropriate packaging design is a very important issue for two reasons. First, it affects the probability of damage, which continues to present a significant challenge, especially for bulky items. For some items, the probability of damage was reported to be as high as 1/3 for each loading and unloading cycle. Packaging also affects the handling time and storage space required per item, and the need for repackaging at the DC. In order to minimize the complexity and cost associated with different packaging requirements across the channels, packaging tends to be designed for the most demanding channel (often the catalog/Internet channel). This can increase the product costs in other channels. Some retailers, such as Walmart, have achieved significant cost savings by redesigning their product packaging to facilitate shipping (Plambeck and Denend 2007).

3.4 Distribution Planning and Inventory Management

Company A operates in a centralized planning environment. Store managers do not place merchandise orders, but rely instead on decisions made by central planners. Nearly 50 % of goods are on auto-replenishment programs, where replenishments

come from the DC/ warehouse. Some branded merchandise can be replenished directly from the vendor. The systems in place for communication between stores and DCs are viewed as satisfactory, but they are still in the process of rolling out EDI linkages with vendors.

The frequency of shipping to stores presents an interesting challenge. Shipping less frequently reduces shipping costs, but increases the size of the shipments. Large shipments can generally be received by stores only before they open for business, which presents considerable staffing challenges. Consequently, smaller and more frequent shipments tend to be preferred, since they can be received by the store during normal working hours. Stores generally maintain only small back-rooms for stocking inventory, and may occasionally also rent off-site lockers for additional storage needs.

Scientific inventory management and demand forecasting is an acknowledged shortcoming of the present system at both of these retailers. Inventory management decisions are often made in an ad hoc manner, using rule of thumb weeks-of-supply (WOS) targets for merchandise at stores and in the DC/ warehouse, without a clear understanding the cost implications of over- or under-stocking. The result tends to be higher than optimal levels of inventory and an annual inventory turnover of less than 2.0 for Company A, which is well below that of some other home furnishings retailers. However, this retailer's strategy focuses on carrying the latest trends in home furnishings together with a fairly high markup. This has produced satisfactory results from a profitability standpoint, but the logistics planners believe that there are significant opportunities for cost reductions.

3.5 Clearance and Markdown Optimization

As mentioned earlier, unsold or slow-moving items are sent to one of the retailer's outlet stores, or sold through the Internet channel. It is important at some point to clear the discontinued items to make room for new merchandise. One option is to take markdowns at stores, but deeper price markdowns generally occur in outlet stores or on the Internet. A second liquidation option is to sell discontinued merchandise to a discounter, after removing labels that identify its origin. Some items may be donated to charitable organizations, which creates a tax deduction. Still others may simply be discarded.

The logistics planners that we spoke with felt that markdown planning and pricing decisions are not made in a scientific manner by this retailer. Often, the merchandise planners wait too long before implementing markdowns or liquidating products. This is also recognized as an opportunity for improving profits (see Chaps. 13–15 for further discussion of pricing and markdown issues).

3.6 Cross-Channel Optimization

While Company A has done little to integrate many of the supply chain processes across the various brands, they do make use of cross-channel marketing (Kalyanam and Achabal 2005). For instance, their advertising expense in the traditional print and mass media is minimal. In fact, their catalogs are used as the primary advertising mechanism, with about 400 million catalogs shipped annually. Many of their catalogs are shipped to areas where stores already exist, and this serves as an instrument to drive store traffic. To compensate the catalog channel for this service, which is significantly cheaper than actual advertising, they receive a fixed percentage of store revenue as a fee. Aggregate information about consumers and their buying behavior in the catalog channel is also used in making decisions about store location and for assessing the market potential of new products. This could likely produce additional benefits if cross channel supply chain interactions were included in the decision making process.

4 Conclusion

These discussions of the supply chain operations at two home furnishings retailers highlight a wide variety of unsolved analytical problems. One specific problem that is analytically challenging is the optimal use of containers to transport the flow of various quantities of merchandise from different supplier locations to the retailer's DC and stores, subject to delivery scheduling constraints. While some models exist in the literature for optimal container packing (Martello et al. 2000), the more general problem of optimally using of an integer number of containers to deliver a flow of merchandise over time appears to be unsolved. For example, it may be advantageous, based on inventory versus shipping cost tradeoffs, to deliver some merchandise ahead of schedule and store it, in order to achieve the objective of exactly filling a container. A complete container that can be shipped to the retailer's DC avoids the additional expense of consolidation with another retailer's merchandise. A further objective is to ship a complete container directly to a store, if possible.

Chapter 8 in this volume discusses a number of papers that deal with the combined problems of assortment selection and inventory management. But modeling the life cycle costs associated with flowing the merchandise in the assortment through the retailer's complete supply chain is beyond the scope of the currently available methods. For example, how does the assortment selection affect the shipping container and inventory cost tradeoffs discussed above?

Additional aspects of assortment planning and inventory management are the presentation requirements for merchandise in stores and catalogs. Chapter 14 in this volume discusses several papers that have studied the impacts of inventory level on sales. But these models do not address the requirement to feature a combination of

items that creates an attractive display. That is, assortment optimization models should somehow include these presentation effects.

The sequential nature of the retailer's decision making process is also an interesting variation on what existing supply chain models tend to assume. That is, assortment decisions are made first, followed by sourcing decisions, inventory ordering decisions, and finally shipping decisions. The timing for these retailer decisions is largely determined by the different lead times associated with each decision. That is, the two retailers described here have elected to postpone each separate decision as long as possible, rather than making them jointly. Conceptually, the overall problem could be modeled as one gigantic dynamic programming problem, but it would clearly be completely intractable. Models that capture the timing of these decisions in a way that includes sequentially updated states of information about demand could potentially be quite useful.

Finally, cross channel optimization clearly offers a number of opportunities for improving supply chain performance at both of these retailers. There are economies of scale across the channels in sourcing, in optimizing shipping containers, and in the use of trucks to deliver shipments to stores, which are currently not being exploited. In many cases, this is because retailers do not have methodologies that can capture these tradeoffs. Cross channel pricing tradeoffs are also important, in particular when a different channel is used to clear the excess merchandise from the original sales channel. There are also cross channel impacts of promotions, some of which are discussed in Kalyanam and Achabal 2005.

In summary, these two case studies illustrate the complexity of retailers' supply chain decisions in practice, and the gaps that exist between the currently available methodologies and the actual decision making environment. We hope that these discussions, as well as the methods and empirical studies presented in this volume, will provide the foundation for future research that will advance the state of the art in retail supply chain management and provide significant additional value for retailers' supply chain operations.

References

Agrawal, N., Smith, S. A., & Tsay, A. A. (2002). Multi-vendor sourcing in a retail supply chain. *Production and Operations Management, 11*(2), 157–182.

Kalyanam, K., Achabal, D.D. (2005). Cross-channel optimization: a strategic roadmap for multichannel retailers. Working paper, Retail Management Institute, Santa Clara University, CA.

Martello, S., Pisinger, D., & Vigo, D. (2000). The three dimensional bin packing problem. *Operations Research, 48*(2), 256–267.

Plambeck, E., Denend, L. (2007). Walmart's sustainability strategy. Stanford Graduate School of Business Case, OIT-71.

Chapter 3
The Effects of Firm Size and Sales Growth Rate on Inventory Turnover Performance in the U.S. Retail Sector

Vishal Gaur and Saravanan Kesavan

1 Introduction

Inventory management is critical to the success of a retailer, whether brick-and-mortar or online, for several reasons. First, inventory constitutes a significant fraction of the assets of a retail firm. Specifically, it is the largest asset on the balance sheet for 57 % of publicly traded retailers in our dataset.[1] The ratio of inventory to total assets averages 35.1 % with buildings, property, and equipment (net) constituting the next largest asset at 31 %. Moreover, the ratio of inventory to current assets averages 58.4 %. Second, inventory, being a current asset, is typically the largest use of working capital of a retailer. Therefore, inventory management is an important determinant of liquidity risk of a retailer. Third, inventory is not only large in dollar value but also critical to the performance of retailers because a retailer cannot sell what it doesn't have. For example, according to Standard & Poor's industry survey on general retailing (Sack 2000), "Merchandise inventories are a retailer's most important asset, even though buildings, property and equipment usually exceed inventory value in dollar terms." Thus, the importance of improving inventory management in retail trade cannot be overemphasized.

[1] The data set consists of a large cross-section of US public listed retailers for the time-period 1985–2003. The data set is summarized in Sect. 3.

V. Gaur (✉)
Johnson Graduate School of Management, Cornell University, Sage Hall,
Ithaca, NY 14853, USA
e-mail: vg77@cornell.edu

S. Kesavan
Kenan-Flagler Business School, University of North Carolina at Chapel Hill,
Chapel Hill, NC 27599, USA
e-mail: skesavan@unc.edu

N. Agrawal, S.A. Smith (eds.), *Retail Supply Chain Management*,
International Series in Operations Research & Management Science 223,
DOI 10.1007/978-1-4899-7562-1_3

At the firm level, managers and analysts commonly use either inventory turnover (defined as the ratio of cost of goods sold to average inventory) or its reciprocal—days of inventory, as a measure to assess how well a retailer is managing its inventory. The statistics for inventory turnover are publicly available from the financial statements of those retailers that are listed on the stock exchange (NYSE, AMEX or NASDAQ), making it an attractive metric for retailers as well as analysts.

However, Gaur, Fisher and Raman (2005), henceforth referred to as GFR, show that inventory turnover varies widely not only across firms but also within firms over time. They further show that a large fraction of the variation in inventory turnover can be explained by three performance variables obtained from public financial data: gross margin (the ratio of gross profit net of markdowns to net sales), capital intensity (the ratio of average fixed assets to average total assets), and sales surprise (the ratio of actual sales to expected sales for the year). They use the estimation results to propose a metric, adjusted inventory turnover, for benchmarking inventory productivity of retail firms.

In this paper, we extend the model of GFR to investigate the effects of firm size and sales growth rate on inventory turnover performance of U.S. retailers. The EOQ and newsvendor models, commonly used in theoretical operations management, show that inventory turnover should increase with the size of a firm due to economies of scale and scope. Several factors contributing to economies of scale and scope have been studied in the operations management literature, including statistical economies of scale (Eppen 1979, Eppen and Schrage 1981), fixed costs in inventory and transportation models, and demand pooling effects in product variety. However, to our knowledge, there are no research papers using real data to estimate the effect of size on inventory turnover. Our results provide such estimates for retailers.

The relationship between sales growth rate and inventory turnover, while not directly studied in the academic literature, is commonly tracked by managers and analysts. For example, the aforementioned industry survey on general retailing by Standard & Poor's (Sack 2000) states that year-over-year growth in inventory should be in line with sales growth rate; if inventory growth exceeds sales growth rate, then it may be a warning that stores are over-stocked and vulnerable to markdowns. Raman et al. (2005) present a case study of a hedge fund investor who uses the ratio of sales growth rate to inventory growth rate as one of the metrics in making investment decisions on retail stock. The case presents several examples from financial performance of firms to illustrate this metric. It also makes a separate point that this relationship is ignored by financial investors. In this paper, we focus on examining evidence for the relationship of sales growth rate with inventory turnover, but do not assess its use by investors. We motivate this relationship using the operations management literature by using an instance of the newsboy model. For our analysis, we do not directly work with sales growth rate because we use a logarithmic regression model which precludes negative values of sales growth rate. Instead, we conduct our analysis using sales ratio, which we define as the ratio of sales in the current year to sales in the previous year.

The main results of our paper are as follows. First, we find that inventory turnover is positively correlated with firm size where size is defined as annual firm sales in the previous year. On average, in our data set, a 1 % increase in firm size is associated with a 0.035 % increase in inventory turnover (statistically significant at $p < 0.0001$). We find evidence of diminishing returns to size: inventory turnover increases with size at a slower rate for large firms than for small firms. These results present evidence in support of the existence of economies of scale and scope in a retail setting.

Next, we find that inventory turnover is positively correlated with sales ratio. A 1 % increase in this ratio is associated with a 0.38 % increase in inventory turnover in our data set. We also find that inventory turnover is more sensitive to sales ratio when a firm is experiencing sales decline than when a firm is experiencing sales growth. A 1 % increase in sales ratio is associated with 0.67 % increase in inventory turnover when sales are declining and with 0.19 % increase in inventory turnover when sales are increasing. Our results suggest that firms would find it harder to improve inventory turnover performance during periods of sales decline than during periods of sales growth. Thus, firms should use their forecast of future sales ratio to determine the amount of attention to give to inventory management.

The third main result of this paper is achieved through re-testing the hypotheses in GFR regarding gross margin, capital intensity and sales surprise on our data set. We test these hypotheses again because we use a larger and more recent data set than GFR. Our results for these tests are consistent with those obtained by GFR. We find that inventory turnover is negatively correlated with gross margin and positively correlated with capital intensity and sales surprise.

Our paper contributes to the academic literature by extending the methodology in GFR for empirical research on inventory productivity in retailing. We find that a significant fraction of the variation in inventory turnover for retailers can be explained by the selected performance variables. The models used in this paper and in GFR are useful to retail managers for comparing inventory turnover performance across firms and for a firm over time. They are also useful in helping retailers estimate inventory turnover as a function of their future growth, profit margin, and capital investment projections. With respect to the effects of firm size and sales ratio on inventory turnover, we describe several factors, based on the literature, which would imply either positive or negative correlations between size and inventory turns as well as between sales ratio and inventory turns. Thus, we set up competing hypotheses, and our tests enable us to state which of these effects will dominate. We believe that there is considerable scope for future research on these topics, and our results represent a first step.

The rest of this paper is organized as follows. Section 2 reviews the relevant literature; Sect. 3 describes our data set; and Sect. 4 summarizes the empirical model and findings from GFR that are useful in this paper. Section 5 presents our hypotheses, followed by the estimation model in Sect. 6, and the estimation results in Sect. 7. Finally, Sect. 8 discusses the limitations of our analysis and directions for future research.

2 Literature Review

The recent years have seen the emergence of a rich literature on econometrics-based research in inventories within Operations Management. Research papers in this area have targeted three types of applications:

1. *Performance benchmarking*: This involves developing methods for benchmarking inventory-related performance in a cross-section or time-series of data.
2. *Generation of descriptive insights*: Researchers have tested hypotheses from inventory theory and investigated the effects of characteristics such as capital intensity, demand uncertainty, and gross margin on inventory data. Recent papers have also developed methods to impute inventory-related costs from structural models of optimal inventory decisions.
3. *Prediction of future performance*: While the above applications treat inventory as the dependent variable, some research papers have treated inventory as a lagged explanatory variable and investigated its information content for predicting future sales, earnings, or stock returns.

The data used in this area of research are typically at an aggregated level, either the firm-level or the industry-level, with a few exceptions. The usage of such aggregated data has been common in economics to study business cycles and production smoothing. In operations, it contrasts with item-level models that have been the subject of much research in inventory theory. However, aggregate-level models are nevertheless valuable in many ways:

1. Firms make many decisions at the aggregate level, such as the fraction of the budget to be set aside for inventory in a given quarter, the bonus to be given to logistics managers based on the performance of a group of products, or whether to discontinue a product line or close a store or a warehouse. Some of these decisions are required in the Sales and Operations Planning (S&OP) processes in firms. Aggregate-level econometric models are useful for making such decisions.
2. Aggregate firm-level data are typically the only kind of inventory data available to analysts, investors, and lenders. Aggregate-level models are useful to such stakeholders.
3. A firm, while possessing detailed internal data for its own products, has access to only aggregated data for other firms in its marketplace. Therefore, it must use an aggregate-level model to utilize information from a panel of other firms in its own operations.

Our paper focuses on performance benchmarking using firm-level data. We review the relevant literature in this section, first discussing descriptive models, then summarizing predictive models of inventory.

Cachon et al. (2007) examine evidence for the occurrence of the bullwhip effect using industry-level data from the U.S. Census Bureau. They find that wholesale

trade industries exhibit the bullwhip effect, whereas retail trade and most manufacturing industries do not. They show that seasonality of demand mediates this result—industries smooth seasonally unadjusted data but amplify the volatility of deseasonalized data. Rajagopalan and Malhotra (2001) use industry-level time-series data from the U.S. Census Bureau for 20 industrial sectors for the period 1961–1994 to investigate whether inventory turns for manufacturers have decreased with time due to the adoption of JIT principles. They find that raw material and work-in-process inventories decreased in a majority of industry sectors, but do not find any overall trends in finished good inventories.

Chen at al. (2005) use firm-level inventory data from publicly traded manufacturing firms for the period 1981–2000 to study trends in inventory levels for each of raw material inventory, work-in-process inventory and finished-good inventory. They find that raw-material and work-in-process inventories have declined significantly while finished-goods inventory remained steady during this period. These results are consistent with Rajagopalan and Malhotra (2001) although, notably, the two papers use data with different granularity.

Gaur et al. (2005) find wide variation in within-firm inventory turnover of U.S. public-listed retailers, and argue that changes in inventory turnover cannot be directly interpreted as performance improvement or deterioration because they may be caused by changes in product portfolio, pricing, demand uncertainty, and many other firm-specific and environmental characteristics. They propose a benchmarking methodology that combines inventory turnover, gross margin, capital intensity and sales surprise to provide a metric of inventory productivity, which they term as adjusted inventory turnover.

Rumyantsev and Netessine (2007) use quarterly data from over 700 public US companies to test some of the theoretical insights derived from classical inventory models developed at the SKU level. They use proxies for demand uncertainty and lead time, and conduct a longitudinal study to show that inventory levels are positively correlated with demand uncertainty, lead times, and gross margins. The authors also find evidence for economies of scale as larger firms carry relatively lower levels of inventory compared to smaller firms.

Olivares and Cachon (2009) and Cachon and Olivares (2010) study finished goods inventory productivity in the automotive supply chain by using stock data at the dealership level. The first paper examines the effect of local competition among dealerships on inventory holdings by using instrument variables to disentangle two effects—a sales effect of the entry or exit of a competitor, and a service-level effect due to a change in the optimal service level for a dealer due to competitive changes. The second paper compares the level of finished goods inventory across automotive firms and traces their differences to the number of dealerships in the network and production flexibility.

While the above papers develop increasingly sophisticated single-equation panel data models, Olivares et al. (2008) propose a method to conduct a structural estimation of unobserved cost parameters of a newsvendor model from observed data on inventory levels and sales, assuming that the decision-maker optimizes inventory. Bray and Mendelson (2012) conduct the structural estimation of a

multi-period inventory model with time-varying demand with the object of determining the information lead time of inventory procurement decisions. Applying this model to quarterly firm-level data for U.S. public-listed firms, they assess the occurrence of the bullwhip effect and decompose it into information transmission leadtime components. Kesavan et al. (2010) present a simultaneous equations model of inventory, sales, and gross margin to represent contemporaneous relationships among these three variables. That is, increase in inventory fuels an increase in sales and a decrease in margin; an increase in sales leads to larger investment in inventory and an increase in margins; finally, an increase in margins leads to a decrease in sales and an increase in inventory. Kesavan et al. (2010) test this model on data for U.S. public listed retailers. Jain et al. (2013) extend this simultaneous equations model to examine the effect of outsourcing on inventory levels. They merge financial data for public-listed firms with international trade transaction data from the U.S. Customs Department and examine the effect of location of sourcing and use of multiple suppliers on the inventory levels of firms.

The above series of papers have led to an evolution of increasingly sophisticated descriptive models of inventory. The interaction of inventory with sales and gross margin suggests that inventory data may contain unique information predictive of future financial performance of firms. Indeed, such a hypothesis is suggested by the case study Raman et al. (2005), which examines the usefulness of inventory data for forecasting future stock returns of firms. Investigating this hypothesis, Kesavan et al. (2010) augment time-series sales forecasting methods with inventory data and show that the resulting 1-year-ahead sales forecasts improve upon benchmark sell-side equity analysts. They further show that lagged inventory data are predictive of bias in those analysts' forecasts. Kesavan and Mani (2013) build on this result and show that lagged inventory is predictive of 1-year-ahead future earnings of U.S. retail firms.

Researchers have also related inventory turnover performance with stock returns in both contemporaneous and predictive models. Gaur et al. (1999) conduct a long-term contemporaneous analysis, and show that for time periods varying in length from 5 to 20 years, the cross-section of average stock returns is significantly positively correlated with average annual inventory turnover over the same period (controlling for gross margin). Chen et al. (2005, 2007) investigate whether abnormal inventory predicts future stock returns. Using the three-factor time-series regression model of stock returns (Fama and French 1993), they find that abnormally high and abnormally low inventories in the manufacturing sector are associated with abnormally poor long-term stock returns. The results for wholesale and retail trade sectors, however, differ from the manufacturing sector. Alan et al. (2014) build on this research and investigate whether inventory productivity is predictive of future stock returns for U.S. public-listed retailers using different measures of inventory productivity and a non-parametric portfolio formation method. They find that inventory turnover and adjusted inventory turnover is strongly predictive of future stock returns using both level- and change-based metrics.

Several researchers have studied the effects of specific operational decisions on firm performance. For example, Balakrishnan et al. (1996) study the effect of adoption of just-in-time (JIT) processes on return on assets (ROA). They compare a sample of 46 firms that adopted JIT processes against a matched sample of 46 control firms that did not. They do not find any significant ROA response to JIT adoption. Billesbach and Hayen (1994), Chang and Lee (1995), and Huson and Nanda (1995) study the impact of adopting JIT processes on inventory turns. Lieberman and Demeester (1999) study the impact of JIT processes on manufacturing productivity in the Japanese automotive industry. Their study suggests that reduction in inventory brought about by JIT practices enabled the firms to improve their productivity.

Our paper contributes to this research stream by extending Gaur et al. (2005) and Rumyantsev and Netessine (2007). We discuss various factors that could cause positive or negative correlations of size and sales growth rate with inventory turnover, and provide evidence regarding the existence of economies of scale and scope in retailing as well as the effect of growth rate of firms on their inventory turnover performance. Our results are useful to retailers to assess their performance changes over time.

3 Data Description

We use financial data for all publicly listed U.S. retailers for the 19-year period 1985–2003 drawn from their annual income statements and quarterly and annual balance sheets. These data are obtained from Standard & Poor's Compustat database using the Wharton Research Data Services (WRDS).

The U.S. Department of Commerce assigns a four-digit Standard Industry Classification (SIC) code to each firm according to its primary industry segment. For example, the SIC code 5611 is assigned to the category "Men's and Boys' Clothing and Accessory Stores", 5621 is assigned to "Women's Clothing Stores", 5632 to "Women's Accessory and Specialty Stores", etc. We group together firms in similar product groups to form ten segments in the retailing industry. For example, all firms with SIC codes between 5600 and 5699 are collected in a single segment called apparel and accessories. Table 3.1 lists all the segments, the corresponding SIC codes, and examples of firms in each segment.

Figure 3.1 presents a simplified view of an income statement and balance sheet that emphasizes the principal variables of interest in this paper. From Compustat annual data for firm i in segment s in year t, let S_{sit} denote the sales net of markdowns in dollars (Compustat annual field Data12), CGS_{sit} denote the corresponding cost of goods sold (Data41), and $LIFO_{sit}$ be the LIFO reserve (Data240). From Compustat quarterly data for firm i in segment s at the end of quarter q in year t, let GFA_{sitq} denote the gross fixed assets, comprised of buildings, property, and equipment (Compustat quarterly field Data118), and Inv_{sitq} denote the

Table 3.1 Classification of data into retail segments using SIC codes

Retail industry segment	SIC codes	Number of firms	Number of observations	Examples of firms
Apparel and accessory stores	5600–5699	75	944	Ann Taylor, Filenes Basement, Gap, Limited
Catalog, mail-order houses	5961	51	540	Amazon.com, Lands end, QVC, Spiegel
Department stores	5311	26	374	Dillard's, Federated, J. C. Penney, Macy's, Sears
Drug and proprietary stores	5912	23	254	CVS, Eckerd, Rite Aid, Walgreen
Food stores	5400, 5411	62	756	Albertsons, Hannaford Brothers, Kroger, Safeway
Hobby, Toy, and game shops	5945	11	118	Toys R Us
Home furniture and equip stores	5700, 5712	24	260	Bed Bath & Beyond, Linens N' Things
Jewelry stores	5944	17	210	Tiffany, Zale
Radio, TV, consumer electronics stores	5731, 5734	20	276	Best Buy, Circuit City, Radio Shack, CompUSA
Variety stores	5331, 5399	44	514	K-Mart, Target, Wal-Mart, Warehouse Club
Aggregate statistics		353	4246	

inventory valued at cost (Data38). From these data, we compute the following performance variables:

Inventory turnover (also called inventory turns), $IT_{sit} = \frac{CGS_{sit}}{\left(\frac{1}{4}\sum_{q=1}^{4} Inv_{sitq}\right) + LIFO_{sit}}$,

Gross margin, $GM_{sit} = \frac{S_{sit} - CGS_{sit}}{S_{sit}}$,

Capital intensity, $CI_{sit} = \frac{\sum_{q=1}^{4} GFA_{sitq}}{\sum_{q=1}^{4} Inv_{sitq} + 4 \cdot LIFO_{sit} + \sum_{q=1}^{4} GFA_{sitq}}$, and

Sales ratio, $g_{sit} = \frac{S_{sit}}{S_{si,t-1}}$.

It is useful to note the following aspects of the measurement of these variables.

1. The Compustat database identifies ten methods for inventory valuation. Four of these are commonly used by retailers: FIFO (first in first out), LIFO (last in first out), average cost method, and retail method. The LIFO reserves of a firm vary depending on the method of valuation used, and adding back the LIFO reserves provides us a FIFO valuation of inventory.

a Income Statement

	Notation	*Amount ($)*
Sales (net of markdowns)	S	100
Cost of Goods Sold (includes Occupancy and Distribution Costs)	CGS	(60)
Gross Profit		40
Selling, General & Administrative Expenses	SGA	(20)
Operating Profit	EBITDA	20
Depreciation & Amortization Expenses		(5)
Interest Costs		(6)
Profit Before Tax	PBT	9
Taxes		(4)
Net Profit	PAT	5

b Balance Sheet

Assets			*Liabilities*		
Fixed Assets (includes Owned Property and Capitalized Leases)	FA	30	Owner's Equity (includes Retained Earnings)	OE	40
Cash		15			
Inventory	Inv	45	Long-term Debt	LTD	20
Accounts Receivable		10	Accounts Payable		40
Total Assets	TA	100	Total Liabilities		100

Fig. 3.1 Simplified view of income statement and balance sheet of a retail firm. (**a**) Income statement. (**b**) Balance sheet

2. The cost of goods sold line on the income statement comprises a number of expenses other than the purchase cost of merchandise. Costs of warehousing, distribution, freight, occupancy, and insurance can all be included in CGS_{sit}. Further, the components of CGS_{sit} may vary from company to company. Most commonly, occupancy costs may be a separate line item on the income statement rather than being included in CGS_{sit}. This lack of uniformity in reporting reduces the comparability of results among retailers. Thus, we restrict our analysis to comparisons within firm. Compustat indicates whether a firm changed its accounting policies with respect to a particular variable during a year; it provides footnotes to variables containing this information. We use these footnotes to identify firms that underwent accounting policy changes, and exclude them from our sample.
3. In the computation of inventory turns and capital intensity, we calculate average inventory and average gross fixed assets using quarterly closing values in order to control for systematic seasonal changes in these variables during the year. LIFO reserves are reported annually. We add the annual LIFO reserves to the average quarterly inventory to compute average inventory.

After computing all the variables, we omit from our data set those firms that have less than five consecutive years of data available for any sub-period during 1985–2003; there are too few observations for these firms to conduct time-series analysis. These missing data are caused by new firms entering the industry during the period of the data set, and by existing firms getting de-listed due to mergers, acquisitions, liquidations, etc. Further, we omit firms that had missing data or accounting changes other than at the beginning or the end of the measurement period. These missing data are caused by bankruptcy filings and subsequent emergence from bankruptcy, leading to fresh-start accounting.

Our final data set contains 4,246 observations across 353 firms, an average of 12.03 years of data per firm. Table 3.2 presents summary statistics by retailing segment for the performance variables used in our study. It lists the mean, median and standard deviation by segment for each variable. Observe that food retailers have the highest median inventory turns of 10.0 and the lowest median gross margin of 0.26. On the other hand, jewelry retailers have the lowest median inventory turns of 1.54 and the highest median gross margin of 0.46. Also note that the coefficient of variation of inventory turnover (the ratio of standard deviation of IT_{sit} to mean IT_{sit}) is quite high: it is larger than 50 % for six out of ten retail segments and its average value across all segments is 74 %. This statistic shows that inventory turnover has a large variation even within each retail segment. Table 3.3 shows the Pearson correlation coefficients for $(\log IT_{sit} - \log IT_{si})$, $(\log GM_{sit} - \log GM_{si})$, $(\log CI_{sit} - \log CI_{si})$, $(\log S_{si,t-1} - \log S_{si})$ and $(\log g_{sit} - \log g_{si})$ for our data set. Here, we use log-values of all variables because we shall construct a multiplicative regression model in the rest of this paper. We compute the correlation coefficients for mean-centered log-values of variables because our model seeks to explain intra-firm variation in inventory turns. Mean centering is done by subtracting out the mean for each variable for each firm from the data columns; for example, $\log IT_{si}$ denotes the average of $\log IT_{sit}$ for firm i in segment s. Notice that $(\log IT_{sit} - \log IT_{si})$ is negatively correlated with $(\log GM_{sit} - \log GM_{si})$ and $(\log S_{si,t-1} - \log S_{si})$, and positively correlated with $(\log CI_{sit} - \log CI_{si})$ and $(\log g_{sit} - \log g_{si})$. Testing hypotheses on these correlations will require a multivariate model which is discussed in subsequent sections.

4 Adjusted Inventory Turnover

GFR study the correlation of inventory turnover with gross margin, capital intensity and sales surprise using data for 311 publicly listed U.S. retailers for the period 1985–2000. In their paper, gross margin, and capital intensity are defined as shown in Sect. 3. Sales surprise, denoted SS_{sit}, is defined as the ratio of current year sales to the forecast of current year sales, where the forecast is computed by GFR using a time-series forecasting method. GFR hypothesize that inventory turnover is negatively correlated with gross margin, and positively correlated with capital intensity and sales surprise.

Table 3.2 Summary statistics of the variables for each retailing segment

Retail industry segment (1)	Number of firms (2)	Number of annual observations (3)	Average annual sales ($ million) (4)	Average inventory turnover (5)	Average gross margin (6)	Average capital intensity (7)	Average sales ratio (8)	Median inventory turnover (9)	Median gross margin (10)	Median capital intensity (11)	Median sales ratio (12)
Apparel and accessory stores	75	944	1201.8	4.60	0.36	0.60	1.14	4.20	0.35	0.62	1.10
				2.15	0.08	0.14	0.38				
Catalog, mail-order houses	51	540	489.3	8.63	0.38	0.50	1.62	5.39	0.38	0.50	1.18
				9.10	0.17	0.18	2.83				
Department stores	26	374	7068.6	4.61	0.34	0.65	1.08	3.55	0.35	0.66	1.05
				3.93	0.07	0.12	0.20				
Drug and proprietary stores	23	254	2327.6	5.26	0.29	0.48	1.17	4.37	0.29	0.50	1.11
				2.91	0.07	0.12	0.28				
Food stores	62	756	5518.4	10.81	0.26	0.76	1.08	9.98	0.26	0.77	1.05
				4.41	0.06	0.08	0.20				
Hobby, toy, and game shops	11	118	1638.1	3.16	0.34	0.47	1.18	2.75	0.35	0.45	1.14
				1.31	0.07	0.14	0.30				
Home furniture and equip stores	24	260	391.1	4.76	0.42	0.58	1.23	3.11	0.43	0.57	1.13
				7.40	0.09	0.15	0.61				
Jewelry stores	17	210	485.0	3.18	0.41	0.39	1.12	1.54	0.46	0.37	1.10
				7.04	0.15	0.14	0.36				

(continued)

Table 3.2 (continued)

Retail industry segment (1)	Number of firms (2)	Number of annual observations (3)	Average annual sales ($ million) (4)	Average inventory turnover (5)	Average gross margin (6)	Average capital intensity (7)	Average sales ratio (8)	Median inventory turnover (9)	Median gross margin (10)	Median capital intensity (11)	Median sales ratio (12)
Radio, TV, cons electr stores	20	276	1779.1	4.27	0.32	0.45	1.17	3.97	0.30	0.46	1.14
				1.71	0.12	0.10	0.24				
Variety stores	44	514	7763.8	4.63	0.29	0.52	1.12	3.72	0.28	0.52	1.10
				3.00	0.09	0.15	0.22				
Aggregate statistics	353	4246	3222.8	6.14	0.33	0.58	1.19	4.36	0.32	0.59	1.09
				5.53	0.11	0.17	1.05				

Note: the values given in columns (5)–(8) are the mean and standard deviation of each variable for the respective segment. The "Aggregate statistics" row refers to the complete data set

Table 3.3 Pearson correlation coefficients for all mean-centered variables

	$\log GM_{sit} - \log GM_{si}$	$\log CI_{sit} - \log CI_{si}$	$\log S_{si,t-1} - \log S_{si}$	$\log g_{sit} - \log g_{si}$
$\log IT_{sit} - \log IT_{si}$	−0.2747	0.1762	−0.04269	0.2651
	<.0001	<.0001	0.0081	<.0001
$\log GM_{sit} - \log GM_{si}$		0.0514	−0.0102	0.0509
		0.0014	0.5265	0.0016
$\log CI_{sit} - \log CI_{si}$			0.2501	−0.1830
			<.0001	<.0001
$\log S_{si,t-1} - \log S_{si}$				−0.4838
				<.0001

Note: for every pair of variables, the table provides the Pearson's correlation coefficient and its p-value for the hypothesis H_1: $|\rho| \neq 0$

GFR use the following empirical model to test their hypotheses:

$$\log IT_{sit} = F_i + c_t + b_s^1 \log GM_{sit} + b_s^2 \log CI_{sit} + b_s^3 \log SS_{sit} + \varepsilon_{sit}. \quad (3.1)$$

Here, F_i is the time-invariant firm-specific fixed effect for firm i, c_t is the year-specific fixed effect for year t, b^1_s, b^2_s, b^3_s are the coefficients of log GM_{sit}, log CI_{sit}, and log SS_{sit}, respectively, for segment s, and ε_{sit} denotes the error term for the observation for year t for firm i in segment s. The hypotheses of GFR imply that, for each segment s, b^1_s must be less than zero, and b^2_s and b^3_s must be greater than zero. The main features of this model are as follows:

1. The model has a log-linear specification. Thus, it is assumed that a multiplicative model is suitable to represent the relationship between inventory turns, gross margin, capital intensity and sales surprise. This assumption is supported in GFR with simulation analysis.
2. The model includes an intercept for each firm in order to control for differences across firms. Note from the discussion in Sect. 3 that inventory turnover may not be comparable across firms due to differences in accounting policies for cost of goods sold. Other factors that can confound comparisons across firms include differences in managerial efficiency, marketing, real estate strategy, etc. Since data on these factors are omitted in GFR, attention is focused on year-to-year variations within a firm only. We call such a model an intra-firm model.

GFR find strong support for all three hypotheses in their data set. Based on these results, they propose a tradeoff curve that computes the expected inventory turnover of a firm for given values of gross margin, capital intensity, and sales surprise. They term the distance of the firm from its tradeoff curve as its A*djusted Inventory Turnover*, denoted AIT, and use it as a metric for benchmarking inventory productivity of retailers by controlling for differences in gross margin, capital intensity, and sales surprise. The value of AIT for firm *i* in segment *s* in year *t* is computed as

$$\log AIT_{sit} = \log IT_{sit} - b^1 \log GM_{sit} - b^2 \log CI_{sit} - b^3 \log SS_{sit} \tag{3.2}$$

or, equivalently, as

$$AIT_{sit} = IT_{sit}(GM_{sit})^{-b^1}(CI_{sit})^{-b^2}(SS_{sit})^{-b^3} \tag{3.3}$$

Note that log AIT_{sit} is equal to the sum of the fixed effects terms, F_i and c_t, and the residual error, ε_{sit}, in Eq. (3.1). Thus, it captures the amount of variation in log IT_{sit} that is not explained by the regressors in Eq. (3.1). According to these results, managers of firms with low AIT should investigate whether their firms are less efficient than their peers, and identify steps they might take in order to improve their inventory productivity.

We employ the methodology from GFR in this paper. In particular, we use an intra-firm model with a log-linear specification. We use log GM_{sit} and log CI_{sit} as control variables for testing our hypotheses because GFR found them to be correlated with log IT_{sit} and they may further be correlated with firm size and sales ratio. We, however, do not use sales surprise in our model because data on managements' forecasts of sales are not available to us. If we were to estimate sales forecasts using our own time-series forecasting methods, then log SS_{sit} and log g_{sit} would be highly correlated and cause collinearity in the model. Hence, in the model in this paper, we replace log SS_{sit} by log g_{sit}.

5 Hypotheses

In this section, we discuss various reasons why inventory turnover can be correlated with firm size and sales ratio. We find that there are arguments in favor of both positive and negative correlation between inventory turns and size as well as between inventory turns and sales ratio. We also find that the effects of size and sales ratio on inventory turnover can vary across firms depending on their supply chain characteristics, business environment and growth strategy. Thus, we identify the mediating variables that are expected to cause size and sales ratio to be correlated with inventory turnover. Since we do not have data on the mediating variables, our hypotheses are limited to testing which effects dominate, positive or negative. We set up competing hypotheses to test these effects. The task of identifying the causes of these correlations is deferred to future research.

5.1 Effect of Firm Size on Inventory Turnover

We explain arguments for inventory turnover to be positively correlated with size using the effects of economies of scale and scope. We also discuss hindrances to economies of scale and scope that may reduce their effect or cause a negative

correlation. Subsequently, we frame competing hypotheses to test the sign of correlation between inventory turnover and size. We measure size by the mean annual sales of the retailer lagged by 1 year, i.e., $S_{si,t-1}$ is the measure of size for year t for firm i in segment s.

Economies of scale and scope can manifest themselves for each item, or in a growth of number of stores, or in a growth of number of items at each retail location. In all three cases, we would expect inventory to increase less than linearly in sales, so that size and inventory turnover would be positively correlated. In the first case, if the mean demand for items at a retail location increases and the retailer maintains a fixed service level, then its safety stock requirement at the location increases less than proportionately because standard deviation of demand typically increases in the square root of mean demand. This relationship is precise when demand follows a Poisson distribution. For other distributions, this relationship has been tested by estimating the first two moments of the distribution. For example, Silver et al. (1998: p.126, 342) estimate the standard deviation of demand as $\sigma = a \cdot (\text{mean})^b$. They state that $0.5 < b < 1$ is typical and "this relationship has been observed to give a reasonable fit for many organizations."[2] As another example, Gaur et al. (2005) estimate the relationship among analysts' forecasts of total sales of firms, actual sales realizations and standard deviation of total sales. Their results are consistent with Silver et al. (1998), with the average estimated value of b across several data sets being 0.71. Therefore, if safety stock increases less rapidly than cycle stock as sales increase, then inventory turnover should increase with the size of each location due to economies of scale.

Second, inventory turnover should increase with sales when a retailer expands its geographical market by opening new retail locations which are served by existing warehouses or distribution centers. Eppen (1979) and Eppen and Schrage (1981) showed how pooling inventory in a centralized location can lead to a reduction in safety stock due to risk pooling. In their models, safety stock grows as $\sqrt{n}$ in the number of locations n if inventory is pooled at a central location rather than distributed across the n locations. Thus, as a firm adds new retail locations, it can achieve a more than proportionate reduction in its inventory level, and a corresponding increase in inventory turnover due to economies of scale in its distribution network.

Third, as a retailer grows in size, it is able to provide more frequent shipments to its stores due to economies of scale and/or economies of scope in fixed replenishment costs as explained by the EOQ model. For example, such economies of scale and scope can be realized in transportation costs through better utilization of labor and transportation capacity. They would result in an increase in inventory turnover with the size of the firm.

[2] This section of Silver et al. (1998) focuses on estimation of demand uncertainty. It does not refer to this relationship as economies of scale.

The above three contributing factors may exist for different firms in different years in varying measures depending on the actions taken by the firms. For example, suppose that a firm increases size in a particular year by adding more products to its assortment without affecting the demand for existing products. For this action, the third argument would contribute to economies of scope, but the first and second arguments would not apply. Our hypotheses do not specify the above three effects separately, but instead specify the average tendency across the cross-section of retail firms for the years included in our data set. This implies that any differences in economies of scale and scope across firms or over time will contribute to the residuals in our model.

Apart from differences across firms, there could be hindrances to economies of scale and scope that may result in a negative correlation between size and inventory turns. First, economies of scale and scope require that a retailer's supply chain infrastructure have excess capacity. For example, distribution centers should be able to meet the requirements of new stores being added, and transportation logistics should be able to handle increase in volume of shipments. If a retailer does not have excess capacity in its supply chain infrastructure, it may need to add new capacity in order to grow. Such hindrances may create diseconomies of scale, implying that size and inventory turnover may be negatively correlated with each other. Second, it is often harder to manage a large firm than a small firm because their operations are more complex. Thus, firms may be unable to exploit operational synergies as they grow in size.[3]

Thus, the above discussion shows that a number of hypotheses can be formulated to estimate different drivers of economies of scale and scope effects among retailers. As a first step, we test the following hypotheses.

Hypothesis 1(a). *Inventory turnover of a firm is positively correlated with changes in its size*.

Hypothesis 1(b). *Inventory turnover of a firm is negatively correlated with changes in its size*.

Here, we use the retailer's sales lagged by 1 year as a measure of size. Our hypotheses may also be set up using relative sales, i.e., the ratio of sales lagged by 1 year to sales at the beginning of the time horizon for the firm. Since we use an intra-firm model, these two measures of size are equivalent.

[3] A counter argument is that as a retailer increases in size, it might have better forecasting tools and thus, might be better able to get the right product to the right place (and therefore, increase turns). Retailers' ability to forecast may even vary non-linearly in size: they may be really good at forecasting when they are very small (not listed publicly, and hence, omitted from our data set), have difficulty as they grow and until they have reached a size such that they have good systems in place and are incorporating sophisticated decision support tools. We incorporate such differences in systems in our model by using capital intensity as a control variable.

5.2 *Effect of Sales Ratio on Inventory Turnover*

We identify reasons why sales ratio can be either positively or negatively correlated with inventory turnover. We construct both arguments using the newsboy model.

First consider the arguments for a positive correlation between sales ratio and inventory turnover. Consider a given retailer with known sales in period $t-1$ making inventory decisions for the next period, t. The retailer first determines the inventory level, q, for an item and then fulfills random demand over one period. Given the value of q, as realized demand increases, sales increase, and thus, sales ratio increases. Further, as realized demand increases, the retailer's average inventory over the period declines. Thus, its inventory turns increase. This implies a positive correlation between sales ratio and inventory turnover. We call this reasoning the *positive effect* of sales ratio on inventory turnover.

Now suppose that the retailer increases q in order to target a higher sales growth rate. As q increases, expected sales increase, and thus, expected sales ratio increases. However, it can also be shown that as q increases, average inventory increases more than proportionately than sales, and expected inventory turnover declines. Alternatively, a retailer may reduce q in order to improve its cash flows. In such a case, the retailer would find its expected inventory turns increasing, but expected sales and expected sales ratio decreasing. This implies a negative correlation between sales ratio and inventory turnover. We call this reasoning the *negative effect* of sales ratio on inventory turnover.

We now try to characterize the situations in which one or the other of these two effects will dominate. Changes in the inventory level or the service level of a retailer can be driven by a number of factors. There is extensive literature on how firms forecast sales growth. Makridakis et al. (1998) state that organizations need to consider several factors such as overall economy, their customers, distributors, competitors, etc. Further from an operations standpoint the firm needs to take into account its inventory levels, capacity constraints, ability to procure inventory from its suppliers, etc. before forecasting sales growth. Once a sales growth rate has been forecasted for the firm it plans to meet this target. The firm has competing objectives in setting its sales growth rate. Some of the common goals are profits, return on investment, market share, product leadership, etc. Hence, it is possible that the overall strategy of the firm may dictate growth while maintaining or improving inventory turnover or it may require the firm to pursue growth at the cost of excess inventory in the short-term.

For example, suppose that a retailer has a large untapped market potential. This is not an uncommon situation because a retailer cannot realize its full market potential overnight. Instead, its growth rate is limited by its capacity to hire and train employees, add new stores, and expand various functions of its organization such as distribution logistics, merchandising, accounting, information systems, etc. Thus, the growth rate of such a retailer can be restricted by its capacity and budget constraints. We expect that for such a retailer, sales could exceed inventory hence

the *positive effect* will dominate so that there will be a positive correlation between sales ratio and inventory turnover.

Alternatively, consider a retailer that is close to saturating its market and has a small untapped market potential. Such a retailer may try to increase its sales growth rate by pushing more inventory to its stores. For example, it may increase service levels of existing products in order to stimulate demand. Or it may open new stores or expand its product line. As the retailer saturates its market, it realizes diminishing sales growth from each new store, store expansion, or new product line. However, all these activities require a fixed inventory outlay to stock the shelves. Therefore, we expect that for such a retailer, the *negative effect* will dominate so that there will be a negative correlation between sales ratio and inventory turnover.

In practice, it is difficult to estimate the market potentials of retailers and classify them into one type or the other. Therefore, we shall estimate the relationship between sales ratio and inventory turnover pooled across all retailers. We set up Hypotheses 2(a)–(b) to test whether positive correlation dominates of negative correlation dominates in our data set.

We also expect that retailers who experience sales decline will find it harder to manage inventory than retailers who experience sales growth because retailers who experience sales decline have to additionally find ways to dispose off excess inventory. Thus, we divide sales ratio into two regions: the *sales expansion region* where $g_{sit} \geq 1$, and the *sales contraction region* where $0 < g_{sit} \leq 1$. We set up Hypothesis 3 comparing these two regions in order to test whether inventory turnover is more sensitive to decline in sales or to increase in sales. Figure 3.2 depicts the relationship proposed in Hypothesis 3.

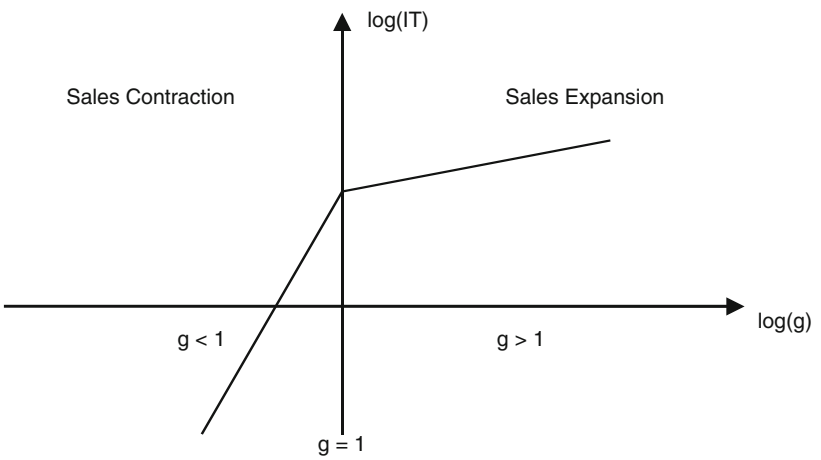


Fig. 3.2 Illustration of Hypothesis 3. *Note*: This figure depicts a piecewise linear fit between the logarithm of inventory turnover, log(IT), and the logarithm of sales ratio, log(g), because we use a log–log model to test our hypotheses

Hypothesis 2(a). *Inventory turnover of a firm is positively correlated with changes in its sales ratio in the sales expansion region as well as the sales contraction region.*

Hypothesis 2(b). *Inventory turnover of a firm is negatively correlated with changes in its sales ratio in the sales expansion region as well as the sales contraction region.*

Hypothesis 3. *Inventory turnover of a firm is more sensitive to sales ratio in the sales contraction region than in the sales expansion region.*

6 Model

We first estimate model (3.1) to re-test the hypotheses in GFR with our data set. Then, we modify the model in GFR to test our hypotheses. The model is specified as follows:

$$\begin{aligned}\log IT_{sit} = F_i + c_t + b^1 \log GM_{sit} + b^2 \log CI_{sit} + b^4 \log S_{si,t-1} \\ + b^5 \log g_{sit} + b^6 \max[0, \log g_{sit}] + \varepsilon_{sit}.\end{aligned} \tag{3.4}$$

Here, F_i is the time-invariant firm-specific fixed effect for firm i; c_t is the year-specific fixed effect for year t; b^1, b^2, b^4, b^5, and b^6 are the coefficients of log GM_{sit}, log CI_{sit}, log $S_{si,t-1}$, log g_{sit}, and log[max(0, g_{sit})], respectively; and ε_{sit} denotes the error term for the observation for year t for firm i in segment s. Hypothesis 1 (a) implies that $b^4 > 0$, Hypothesis 2(a) implies that $b^5 > 0$ and $b^5 + b^6 > 0$, and Hypothesis 3 implies that $b^6 < 0$. The main features of this model are as discussed in Sect. 4.

We estimate several variations of Eq. (3.4) to test our hypotheses. For example, we add the quadratic term, $[\log S_{si,t-1}]^2$, to test whether the effect of firm size on inventory turnover shows decreasing or increasing economies of scale. We also partition our data by firm size in order to study whether sales ratio has different effects on inventory turns for large and small firms. In another modification, we estimate the coefficients of the explanatory variables separately for each segment to test if the results are consistent across all segments or are driven by only a few of the segments in the data set. We use ordinary least squares estimation for simplicity. The estimators thus obtained are consistent in the presence of heteroscedasticity.

7 Results

Table 3.4 shows the results for model (3.1). The three hypotheses in GFR are supported for our larger and more recent data set. The coefficient of gross margin is −0.287, the coefficient of capital intensity is 0.633, and the coefficient of sales surprise if 0.034. All three coefficients are statistically significant at $p < 0.0001$.

Table 3.4 Re-test of the hypotheses in Gaur et al. (2005)

	Estimate	Std. error
R^2 (%)	93.86	
log GM_{sit}	-0.287^{***}	0.024
log CI_{sit}	0.633^{***}	0.037
log SS_{sit}	0.034^{***}	0.008

Statistically significant at ***$p < 0.0001$

Table 3.5 OLS regression estimates for model (3.4)

	Model (3.1) without quadratic size term			Model (3.1) with quadratic size term		
(1)	Estimate (2)	Std. error (3)	Std. Coeff. estimate (4)	Estimate (5)	Std. error (6)	Std. Coeff. estimate (7)
R^2 (%)	94.06			94.09		
log GM_{sit}	-0.364^{***}	0.047	-0.302^{***}	-0.347^{***}	0.023	-0.302^{***}
log CI_{sit}	0.687^{***}	0.036	0.271^{***}	0.712^{***}	0.037	0.279^{***}
log $S_{si,t-1}$	0.035^{***}	0.011	0.078^{***}	0.105^{***}	0.023	0.165^{***}
$[\log S_{si,t-1}]^2$				-0.006^{***}	0.001	-0.092^{***}
log g_{sit}	0.670^{***}	0.050	0.691^{***}	0.669^{***}	0.048	0.694^{***}
max{0, log g_{sit}}	-0.480^{**}	0.061	-0.388^{**}	-0.454^{**}	0.061	-0.375^{**}

Statistically significant at ***$p < 0.0001$

Table 3.5 shows the fit statistics and coefficients' estimates for model (3.4) in columns (2)–(4). The F-statistic for the model is significant at $p < 0.0001$, and the R^2 value is 92.5 %. The rest of this section discusses the support for hypotheses regarding size and sales ratio.

First, consider the test of Hypotheses 1(a)–(b). We find that inventory turns are positively correlated with size, supporting Hypothesis 1(a). A 1 % increase in the size of a firm leads to a 0.035 % increase in inventory turns ($p < 0.0001$).[4] Note that the effect of size on inventory turns appears to be small compared to other explanatory variables. This may be so because log $S_{si,t-1}$ has a higher standard deviation than the other explanatory variables. In order to control for this difference, we compute the standardized coefficient estimates as shown in column (4) of the Table 3.5 (see Schroeder et al. (1986, p. 31–32) for a description of standardized coefficients). The standardized coefficient of log $S_{si,t-1}$ is 0.078; thus, size still has a smaller effect on inventory turns compared to other variables in our model.

We now investigate whether the coefficient of log $S_{si,t-1}$ differs across firms and across model specifications. The object of this analysis is to characterize how the effects of economies of scale and scope vary across our data set. We first investigate the presence of diminishing economies. Since we have so far shown a linear relationship between log IT_{sit} and log $S_{si,t-1}$, the coefficient of log $S_{si,t-1}$ in this model can be biased downwards if there are diminishing economies of scale and

[4] Relative size, Sales(i,t − 1)/Sales(i,0), yields identical results in an intra-firm model.

scope. To address this possibility, we add a quadratic term, $[\log S_{si,t-1}]^2$, to model (3.4). Columns (5)–(7) in Table 3.5 show the estimation results for this model. We find that the coefficient of $\log S_{si,t-1}$ increases from 0.035 to 0.105, and the coefficient of $[\log S_{si,t-1}]^2$ is -0.006 ($p < 0.01$). Thus, we see that the quadratic model supports the hypothesis that there are diminishing returns to scale as firm size increases.

Another way to identify diminishing economies of scale is to perform the regression separately for small and large firms. We classify firms as small or large using the following approach. We compute the median of $\log S_{si,t-1}$ for every firm, and then use these values to compute the 25th percentile and the median of $\log S_{si,t-1}$ for each segment. In the first regression, firms whose median value of $\log S_{sit}$ falls below the 25th percentile are classified as small firms and the remaining as large firms. In the second regression, the cut-off point is set at the median. Table 3.6 shows the results for the first regression in columns (2)–(5) and for the second regression in columns (6)–(9). We see that in the first regression, the coefficient of $\log S_{si,t-1}$ is 0.11 ($p < 0.0001$) for small firms, and is not statistically significant for large firms. In the second regression, the coefficient of $\log S_{si,t-1}$ is 0.06 ($p < 0.0001$) for small firms, and is again not significant for large firms. The comparison of estimates between small and large firms is consistent with the results from the quadratic model, and provides strong support for the hypothesis that there are diminishing economies of scale as firm size increases. Note that the decrease in the coefficient estimate for small firms from 0.11 to 0.06 when we increase the set of small firms from the first quartile to the first two quartiles of size distribution is also consistent with the diminishing economies to scale argument.

The coefficient of $\log S_{si,t-1}$ may also differ across retail segments. To investigate this possibility, we estimate the coefficients of the model separately for each retail segment. Table 3.7 shows the results obtained. We find that four of the ten segments have positive and statistically significant ($p < 0.01$) coefficient estimates, one segment has negative and statistically significant ($p < 0.01$) coefficient estimate, and the remaining five segments do not show any statistical relationship. Where positive, the coefficient estimate ranges between 0.06 and 0.16. Jewelry stores have a negative and statistically coefficient estimate of -0.223. We find that the result for jewelry stores is not caused by the presence of any outliers, rather it holds consistently across firms. This suggests that the arguments for economies of scale and scope may not apply to jewelry products because the costs of distribution and logistics that these arguments are based on may not be critical to jewelry retailers.

In summary, we have shown two important relationships between firm size and inventory turnover. The first relationship supports the hypothesis that inventory turnover increases with size. The second relationship relates to diminishing returns to scale.

We now consider the tests of Hypotheses 2(a)–(b) and 3. The results in columns (2)–(4) of Table 3.5 show that inventory turnover is positively correlated with sales ratio in model (3.4). The coefficient of $\log g_{sit}$ is 0.67 and the coefficient of $\max\{0, \log g_{sit}\}$ is -0.48. This implies that a 1 % increase in g_{sit} is associated

Table 3.6 Regression estimates for model (3.4) obtained after partitioning firms based on size

	Small firms (first quartile)		Large firms (second to fourth quartiles)		Small firms (below median)		Large firms (above median)	
(1)	Estimate (2)	Std. error (3)	Estimate (4)	Std. error (5)	Estimate (6)	Std. error (7)	Estimate (8)	Std. error (9)
R-square (%)	89.31		94.24		90.49		94.84	
log GM_{sit}	-0.349^{***}	0.039	-0.332^{***}	0.019	-0.371^{***}	0.029	-0.317^{***}	0.019
log CI_{sit}	0.343^{***}	0.048	0.432^{***}	0.023	0.391^{***}	0.033	0.387^{***}	0.026
log $S_{si,t-1}$	0.108^{***}	0.021	0.001	0.008	0.063^{***}	0.013	0.004	0.011
log g_{sit}	0.773^{***}	0.057	0.502^{***}	0.042	0.712^{***}	0.042	0.497^{***}	0.055
max{0, log g_{sit}}	-0.593^{***}	0.084	-0.283^{***}	0.052	-0.533^{***}	0.056	-0.232^{***}	0.069

Statistically significant at $^{***}p < 0.01$

Table 3.7 Segment-wise coefficients' estimates for model (3.4)

Retail segment	log GM_{sit}	log CI_{sit}	log $S_{si,t-1}$	log g_{sit}
Apparel and accessory stores	-0.166^{***}	0.848^{***}	0.016	0.243^{***}
Catalog, mail-order houses	-0.319^{***}	0.195^{***}	0.148^{***}	0.429^{***}
Department stores	-0.334^{***}	1.049^{***}	−0.008	0.414^{***}
Drug and proprietary stores	-0.212^{***}	0.321^{***}	0.158^{***}	0.562^{***}
Food stores	-0.393^{***}	1.287^{***}	−0.029	0.492^{***}
Hobby, toy, and game shops	-0.894^{***}	0.307	−0.024	0.408^{***}
Home furnishings and equip stores	−0.024	0.680^{***}	0.129^{***}	0.508^{***}
Jewelry stores	-0.683^{***}	0.439^{***}	-0.223^{***}	0.308^{***}
Radio, TV, cons electr stores	-0.330^{***}	0.389^{***}	0.062^{***}	0.307^{***}
Variety stores	-0.187^{***}	0.122^{***}	0.009	0.223^{***}

Statistically significant at ***$p < 0.01$, **$p = 0.05$, and *$p = 0.1$

with a 0.67 % increase in inventory turns in the sales contraction region and with a 0.19 % (=0.67 − 0.48) increase in inventory turns in the sales expansion region. Both these coefficients are statistically significant at $p < 0.0001$. Thus, we find that inventory turnover is positively correlated with sales ratio in both the regions, providing support for Hypotheses 2(a). Moreover, the coefficient of max{0, log g_{sit}} is negative and statistically significant, providing strong support for Hypothesis 3. The average value of the coefficient of log g_{sit} obtained by doing a regression omitting the variable max{0, log g_{sit}} is 0.38.

Columns (5)–(7) in Table 3.5 show the coefficient estimates for sales ratio when the model is quadratic in firm size. We find that the estimates and standard errors of these coefficients are similar to those obtained when the model is linear in size. Therefore, they also support Hypotheses 2 and 3. The results from the separate regressions for small and large firms in Table 3.6 also support our hypotheses.

The coefficients of log g_{sit} and max{0, log g_{sit}} in Tables 3.5 and 3.6 show that the effect of a change in sales ratio on inventory turnover is significantly lower when $g_{sit} > 1$ than when $g_{sit} \leq 1$. In Table 3.5, the coefficient of log g_{sit} is lower in the sales expansion region than in the sales contraction region by 0.48 in the linear model and by 0.454 in the quadratic model. This result confirms our intuition that firms would find it harder to improve inventory turnover during periods of sales decline than during periods of sales growth. Further, Table 3.6 shows that the coefficient estimates of log g_{sit} differ significantly across small and large firms in the sales contraction region, but are statistically similar in the sales expansion region. For example, when the smallest 25 % of firms are classified as small, the coefficient estimates for small and large firms are 0.773 and 0.502, respectively, in the sales contraction region, and 0.180 (=0.773 − 0.593) and 0.219 (=0.502 − 0.283), respectively, in the sales expansion region. Thus, we observe that during periods of sales decline, inventory turns for small firms are more sensitive to sales ratio than for large firms. But during periods of sales expansion, there is no significant difference in the coefficient of sales ratio between small and large firms. The coefficients' estimates for the case in which small and large firms are defined by the median tell the same story.

Table 3.8 Example showing the effect of volatility in sales ratio on expected inventory turnover

	Probability distribution of g_{sit}	Expected multiplicative effect on inventory turnover due to variation in sales ratio[a]	
		Firm classified as small	Firm classified as large
Scenario A	$g_{sit} = 1.2$ with probability 0.5 $g_{sit} = 0.8$ w. p. 0.5	$[(1.2)^{0.18} + (0.8)^{0.77}]/2 = 0.938$	$[(1.2)^{0.22} + (0.8)^{0.50}]/2 = 0.968$
Scenario B	$g_{sit} = 1.1$ with probability 0.5 $g_{sit} = 0.9$ w. p. 0.5.	$[(1.1)^{0.18} + (0.9)^{0.77}]/2 = 0.970$	$[(1.1)^{0.22} + (0.8)^{0.50}]/2 = 0.985$

[a]For the purpose of this table, we classify a firm as small if its size belongs to the first quartile of its retail segment and as large otherwise. Thus, we use the coefficients' estimates in Columns 2 and 4 of Table 3.5 for our computations. All computations are done assuming that (1) the effects of GM_{sit} and CI_{sit} are normalized to zero, (2) the effect of diminishing returns to scale is negligible for small changes in size, and (3) the firm size and sales ratio are normalized to 1.0 in the base case

We explain this result with an example. Consider the effect of volatility in sales growth on the inventory turnover of a firm over a period of 1 year. Table 3.8 shows two growth scenarios for the firm and their effects on inventory turnover. In both scenarios, the firm's expected sales ratio is zero (i.e., $E[g_{sit}] = 1$). The scenarios differ in the standard deviation of sales ratio. We examine each scenario using the coefficients' estimates for a small firm and for a large firm obtained from Table 3.6. For example, in scenario A, we find that the expected inventory turnover of the firm is 93.8 % of what it would have been if g_{sit} were a constant equal to 1. We make the following observations by comparing all the cases in this example:

1. The firm's expected inventory turnover declines in each case even though its total expected sales are equal to the sales in the previous year. Thus, volatility in sales has a negative effect on inventory turnover.
2. The decline in expected inventory turnover is higher if the firm experiences more variation in g_{sit} (i.e., Scenario A) than if the firm experiences less variation in g_{sit} (i.e., Scenario B). For example, for a small firm, expected inventory turns decline by 6.2 % in Scenario A and by 3.0 % in Scenario B.
3. The decline in inventory turnover is higher if the firm is small than if the firm is large. Further, the difference between large and small firms increases as the standard deviation of g_{sit} increases.

Thus, this example shows the effect of volatility in sales ratio on inventory turnover using our results. Interestingly, the inferences from the example are analogous to those from the newsboy model in inventory theory. Further, it shows that a firm with more volatile sales has two ways to improve its inventory turnover: either it should target a sufficiently high growth rate that compensates for the effect of volatility in sales ratio on inventory turnover, or it should reduce its inventory and offer a lower service level.

As with firm size, we analyze whether the coefficient of log g_{sit} is consistent across segments. Table 3.7 shows the coefficients' estimates obtained for each segment. We find that the coefficient of log g_{sit} varies significantly across segments ($p < 0.0001$). However, sales growth consistently has a large positive coefficient for

each segment. Its value ranges between 0.22 for variety stores to 0.56 for Drug and Proprietary stores.

In summary, we find strong support for the hypotheses that inventory turnover is positively correlated with sales ratio and that inventory turnover is more sensitive to sales ratio in the sales contraction region than in the sales expansion region. We also find that the latter effect is stronger for small firms than for large firms.

8 Conclusions and Directions for Future Research

Our paper highlights the importance of understanding inventory turnover performance of retailers. Like GFR, we find that inventory is a significant proportion of the assets of a retailer. However, inventory turnover varies widely across retailers and for a retailer over time. We have shown that a significant proportion of the within-firm variation in inventory turnover is explained by changes in firm size, sales ratio and variables identified by GFR. In particular, inventory turnover of a firm is positively correlated with both size and sales ratio. Our results support the arguments of economies of scale and scope studied in the operations management literature. We use a data set of 353 publicly listed U.S. retailers for the period 1985–2003 in our analysis. This data set is larger and more recent than that used by GFR. Thus, we also examine the hypotheses formulated in GFR regarding the correlations of inventory turnover with gross margin, capital intensity and sales surprise. We find that inventory turnover is strongly negatively correlated with gross margin and positively correlated with capital intensity in our data set. These results are consistent with those obtained in GFR.

Our results are useful to retailers for benchmarking their inventory turnover performance against their peers. Since the correlations estimated by us are based on a large set of firms, they provide estimates of the average change in inventory turnover associated with given changes in gross margin, capital intensity, size and sales ratio. A positive residual for a firm in our model indicates that the firm achieved higher inventory turnover than its peer group after controlling for differences in the explanatory variables, while a negative residual indicates otherwise. Thus, managers may use these residuals to investigate reasons for differences in inventory turnover performance across firms or for a firm over time. The fixed effects in our model may be used similarly by managers for benchmarking. Another application of our results is related to the difference between the coefficients of sales ratio during periods of sales growth and sales decline. This result shows that aggregate retail inventory changes with sales in a manner that is consistent with the newsboy model in inventory theory. This result also implies that managers should pay more attention to managing inventory when a firm is small, or when a firm is going through a period of sales decline, or when a firm faces more volatility in sales.

Our paper suggests three possible directions for future research on aggregate-level inventory management in retailing.

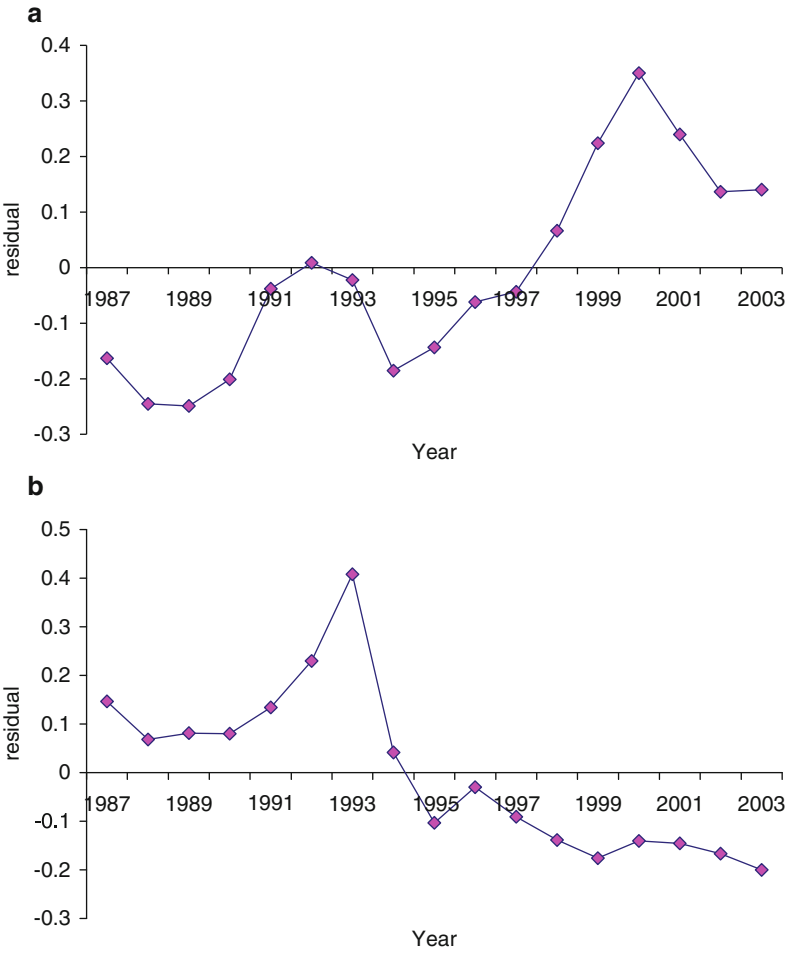


Fig. 3.3 (**a**) Time-series plot of residuals from model (3.4) for Best Buy Stores, Inc. (**b**) Time-series plot of residuals from model (3.4) for Jennifer Convertibles, Inc.

1. *Modeling of aggregate-level inventory decisions*: Even though the variables in our model are statistically significant, there is still a considerable amount of variation in inventory turnover that remains unexplained. For example, we find that the residuals from our model show differing patterns across firms. There are firms whose residuals have consistently improved over time after controlling for changes in all the explanatory variables, and other firms whose residuals show a consistently declining trend. To illustrate this, Fig. 3.3a, b show time-series plots of residuals from our model for Best Buy Stores, Inc. and Jennifer Convertibles, Inc., respectively. Notice that the residuals for Best Buy trend upwards with time while those for Jennifer Convertibles trend downwards. These unexplained but systematic differences suggest that there is scope for future research to better

understand retailers' inventory turnover performance. There has been considerable advancements in econometric models of inventory in the recent years, which could be applied to retailers to help them decipher variations in inventory turnover.

2. *Explaining the drivers of inventory productivity using augmented data sets*: Several operational factors can be said to contribute to the relationships of gross margin, capital intensity, firm size and sales ratio with inventory turnover. Since public financial data do not capture these operational factors, it is not possible to identify the drivers of inventory turnover using these data. A richer data set may be used in future research to examine the aforementioned relationships more closely. For example, the discussion in Sect. 5 identifies many variables that may be included in such a data set, for example, number of store locations, their store formats and square footage, number of warehouses and their square footage, same stores sales growth rates, etc. In a recent paper, Kesavan et al. (2010) construct such a data set by incorporating number of store locations, accounts payables, and several other variables. They apply a simultaneous equations model to estimate causal effects of sales, inventory and gross margin on each other. They further show that their model provides more accurate forecasts of sales than standard time-series models as well as equity analysts.
3. *Examining the effects of firm lifecycle and bankruptcies on model estimation*: Our data set consists of only publicly listed firms that have at least five consecutive years of data available. Since these firms would be above a certain size, our coefficient estimate for size could be subjected to selection bias. Also our coefficient estimates could be subjected to survival bias since slow growing firms could exit from our data set. Future research may examine how these factors affect the relationship of inventory management with other performance variables.

Acknowledgement The authors are thankful to the series editors, Naren Agrawal and Stephen Smith, and anonymous reviewers for many helpful comments on this manuscript. The questions of the effects of firm size and sales growth rate on inventory turnover were suggested to the first author by Marshall Fisher and Ananth Raman.

References

Alan, Y., Gao, G., Gaur, V. (2014). Does inventory turnover predict future stock return? A retailing industry perspective. *Management Science*. (Forthcoming).

Balakrishnan, R., Linsmeier, T. J., & Venkatachalam, M. (1996). Financial benefits from JIT adoption: effects of consumer concentration and cost structure. *Accounting Review, 71*, 183–205.

Billesbach, T.J., Hayen, R. (1994). Long-term impact of JIT on inventory performance measures. *Production and Inventory Management Journal*, 62–67, First Quarter.

Bray, R., & Mendelson, H. (2012). Information transmission and the bullwhip effect: an empirical investigation. *Management Science, 58*(5), 860–875.

Cachon, G., Randall, T., & Schmidt, G. (2007). In search of the bullwhip effect. *Manufacturing & Service Operations Management, 9*(4), 457–479.

Cachon, G., & Olivares, M. (2010). Drivers of finished goods inventory in the U.S. automobile industry. *Management Science, 56*(1), 202–216.

Chang, D., & Lee, S. M. (1995). Impact of JIT on organizational performance of U.S. firms. *International Journal of Production Research, 33*, 3053–3068.

Chen, H., Frank, M. Z., & Wu, O. Q. (2005). What actually happened to the inventories of American companies between 1981 and 2000? *Management Science, 51*, 1015–1031.

Chen, H., Frank, M. Z., & Wu, O. Q. (2007). U.S. retail and wholesale inventory performance from 1981 to 2004. *Manufacturing & Service Operations Management, 9*(4), 430–456.

Eppen, G. (1979). Effect of centralization on expected costs in a multi-location Newsboy problem. *Management Science, 25*(5), 498–501.

Eppen, G., & Schrage, L. (1981). Centralized ordering policies in a multi-warehouse system with lead times and random demand. In L. Schwarz (Ed.), *Multi-level production/inventory control systems: theory and practice* (Vol. 16). North Holland, Amsterdam: TIMS Studies in the Management Sciences.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics, 33*(1), 3–56.

Gaur, V., Fisher, M. L., & Raman, A. (1999). *What explains superior retail performance?* (Working paper). Ithaca, NY: Cornell University.

Gaur, V., Fisher, M. L., & Raman, A. (2005). An econometric analysis of inventory turnover performance in retail services. *Management Science, 51*, 181–194.

Huson, M., & Nanda, D. (1995). The impact of just-in-time manufacturing on firm performance. *Journal of Operations Management, 12*, 297–310.

Jain, N., Girotra, K., Netessine, S. (2013). Managing global sourcing: inventory performance. *Management Science*. (Forthcoming)

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 future earnings for U.S. public retailers. *Manufacturing & Service Operations Management, 15*(1), 6–23.

Lieberman, M. B., & Demeester, L. (1999). Inventory reduction and productivity growth: linkages in the Japanese automotive industry. *Management Science, 45*, 466–485.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: methods and applications* (3rd ed.). New York, NY: Wiley.

Olivares, M., & Cachon, G. (2009). Competing retailers and inventory: an empirical investigation of General Motors' dealerships in isolated U.S. markets. *Management Science, 55*(9), 1586–1604.

Olivares, M., Terwiesch, C., & Cassorla, L. (2008). Structural estimation of the newsvendor model: an application to reserving operating room time. *Management Science, 54*(1), 41–55.

Rajagopalan, S., & Malhotra, A. (2001). Have U.S. manufacturing inventories really decreased? An empirical study. *Manufacturing & Service Operations Management, 3*, 14–24.

Raman, A., Gaur, V., Kesavan, S. (2005). *David Berman*. Harvard Business School Case 605-081.

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

Sack, K. (2000). *Retailing: general industry survey*. New York, NY: Standard & Poor's.

Schroeder, L., Sqoquist, D., & Stephan, P. (1986). *Understanding regression analysis*. London: Sage Publications.

Silver, E. A., Pyke, D. F., & Peterson, R. (1998). *Inventory management and production planning and scheduling* (3rd ed.). New York, NY: Wiley.

Chapter 4
The Role of Execution in Managing Product Availability

Nicole DeHoratius and Zeynep Ton

1 Introduction

Several surveys show that a significant number of customers leave retail stores because they cannot find the products for which they are looking (e.g. Emmelhainz et al. 1991; Andersen Consulting 1996; Gruen et al. 2002; Kurt Salmon Associates 2002). Most research in operations management focuses on two factors to explain suboptimal product availability—poor assortment and poor inventory planning. Our research with several retailers during the last few years highlights a third factor, poor execution, or the failure to carry-out an operational plan. We find that even with the application of algorithms to select the appropriate stocking quantity and appropriate store assortment, the right product may be still unavailable to retail customers. For example, after auditing 50 products at ten different stores, management at a specialty retailer found that only 16 % of the stockouts could be attributed to statistical stockouts (cited in Ton 2002). Instead, 24 % of the stockouts were due to inventory record inaccuracy, discrepancies between the recorded and actual on-hand inventory quantity, and 60 % were due to misplaced products, products that were physically present at the store but in locations where customers could not find them.

Inventory record inaccuracy and misplaced products are two examples of poor store execution. These problems affect product availability in two ways. First, they lead to stockouts and hence compromise retailers' service levels. When the actual level of inventory for a particular product is lower than the planned level due to either inventory record inaccuracy or product misplacement, the actual service level will be

N. DeHoratius (✉)
Booth School of Business, University of Chicago, 5807 S. Woodlawn Avenue, Chicago, IL 60637, USA
e-mail: Nicole.DeHoratius@ChicagoBooth.edu

Z. Ton
MIT Sloan School of Management, E62-581 100 Main Street, Cambridge, MA 02142, USA

N. Agrawal, S.A. Smith (eds.), *Retail Supply Chain Management*,
International Series in Operations Research & Management Science 223,
DOI 10.1007/978-1-4899-7562-1_4

lower than the planned service level. At Borders Group Inc., a formerly large retailer of entertainment products such as books, CDs, and DVDs, lost sales due to misplaced products reduced profits by 25 % (Raman et al. 2001). Andersen Consulting (1996) estimates that sales lost due to products that are present in storage areas but not on the selling floor amount to $560–$960 million per year in the US supermarket industry.

Second, for retailers that rely on automated replenishment systems to manage store inventory, execution problems affect future product availability through the distortion of historical sales and inventory data stored in these systems. Distortion of inventory data may prevent the triggering of a replenishment order when the system inventory is greater than the actual inventory or may unnecessarily trigger an order when the system inventory is less than actual inventory. Moreover, when a product that is actually out of stock is reported as in stock, the automated replenishment system may wrongly conclude there is no demand. The system observes no sales for that item because it is not available to the customer. Thus, even when multiple customers are willing to purchase that item, the system may automatically reduce the forecast of future demand which in turn causes the retailer to stock less of it or even to drop the item from the assortment entirely.

Despite their prevalence and impact, research on execution problems is limited. Much of the work in the retailing context focuses on the drivers of these problems and only recently have researchers attempted to incorporate these problems into existing planning models. In this chapter we summarize the existing research on store execution and identify future research opportunities in this area. The chapter is organized as follows. In Sect. 2, we describe the magnitude and root causes of the two execution problems, based on specific well-researched case studies. In Sect. 3, we describe the findings of the empirical studies that have identified factors that exacerbate the occurrence of execution problems. In Sect. 4, we describe the effect of execution problems on inventory planning and summarize how researchers have incorporated these problems into existing inventory models. Finally, in Sect. 5, we conclude with a discussion of future research opportunities.

2 Retail Execution Problems

Evidence of execution problems exists in a number of different contexts. Distribution centers,[1] manufacturing firms,[2] financial services,[3] utility companies,[4] hospitals,[5] and government agencies[6] have all faced problems with misplaced products

[1] See, for example, Bayers (2002), Millet (1994) and Rout (1976).

[2] See, for example, Hart (1998), Sheppard and Brown (1993), Tallman (1976), Brooks and Wilson (1993), Bergman (1988), Krajewski et al., (1987) Flores and Whybark (1986; 1987), and Woolsey (1977).

[3] See, for example, Cassady and Mierzwinski (2004) and Capital Market Report (2000).

[4] See, for example, Woellert (2004) and Redman (1995).

[5] See, for example, McClain et al. (1992) and Young and Nie (1992).

[6] By the Numbers (2005), McCutcheon (1999), Galway and Hanks (1996), Laudon (1986), Schrady (1970) and Rinehart (1960).

and/or record inaccuracy. The costs pertaining to the inability to execute an operational plan in these contexts, as in retailing, have been shown to be substantial. We describe below the extent of such problems in retailing and identify how they arise.

2.1 Inventory Record Inaccuracy

At Gamma Corporation,[7] a leading retailer with hundreds of stores and over $10 billion in sales, physical audits revealed that inventory record inaccuracy was pervasive throughout the chain (DeHoratius and Raman 2008). Discrepancies were found in 65 % of the nearly 370,000 audited inventory records with the absolute difference between system and actual inventory quantity per item per store ranging from 0 to 6,988 units (Fig. 4.1). The average absolute discrepancy between system and actual inventory was nearly five units, or 36 % of the average target quantity. Of those records that were inaccurate, approximately 58 % of them had positive discrepancies where the recorded quantity exceeded the actual and nearly 42 % of them had negative discrepancies where the on-hand quantity

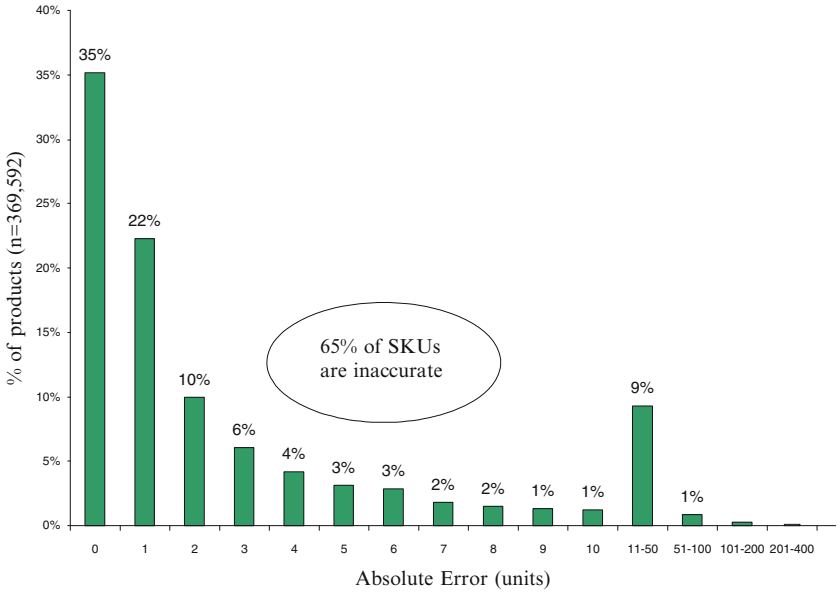


Fig. 4.1 Histogram of the absolute value difference between system and actual inventory measured in units (*Source*: Raman et al. 2001)

[7] Name disguised to preserve confidentiality.

exceeded the recorded quantity. Interestingly, nearly each product that was stocked out in the store at the time of the audit showed a positive on-hand amount recorded in the inventory management system. In other words, these stockouts were invisible to corporate merchandise and inventory planners.

2.2 Misplaced Products

Misplaced products, whether they are mis-shelved or left in storage areas, lead to stockouts if customers are unable to locate the inventory they seek. At Borders, two surveys showed that approximately 18 % of customers who approached a salesperson for help experienced a phantom stockout (Ton and Raman 2003). That is, the product was physically present at the store but could not be found even with the help of a salesperson. Physical audits at 242 Borders stores showed that, on average, 3.3 % of a store's assortment (over 6,000 products per store) was placed in storage areas and had no presence on the selling floor (Fig. 4.2). At some stores, nearly 10 % of the assortment was missing from the selling floor. Note that these estimates of misplaced products are conservative because they do not include those products that have been mis-shelved either by customers or employees.

2.3 Root Causes of Execution Problems

We identify three sources of poor execution: (1) poor process design, (2) an operating environment that makes it challenging for employees to conform to

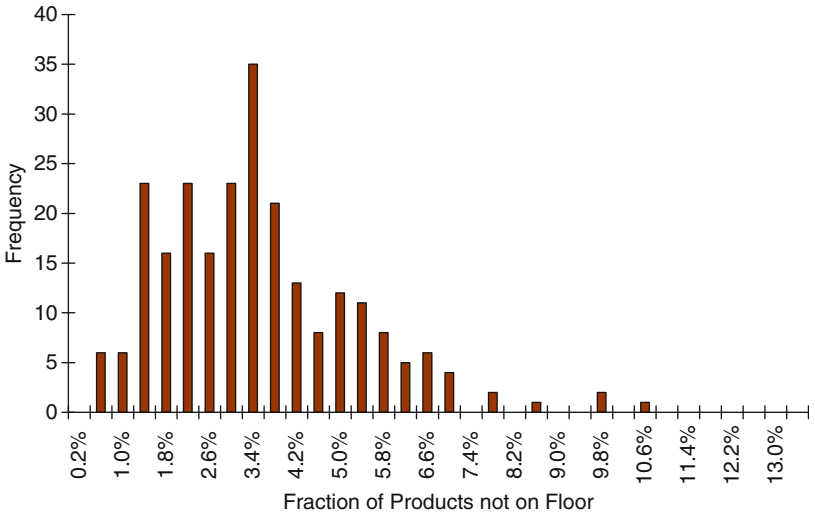


Fig. 4.2 Histogram for the fraction of products that are not available on the sales floor (*Source*: Raman et al. 2001)

prescribed processes, and (3) employee errors. Poor process design may result from, among others, poorly specified work content, poorly specified sequence of activities, inadequate time given to perform work, and an absence of feedback on process quality. At most retail chains for example, extra units of inventory that do not fit into display shelves are kept in storage locations. When the level of inventory of a particular product on the shelf approaches zero, employees are supposed to replenish units of that product from the storage locations. Since existing systems do not track the location of stored products, store employees have to rely on their memory to determine where products are stored in order to replenish the shelves. Not surprisingly, this poor process design often leads to product misplacement. Furthermore, at many retail chains, employees have to manually enter the price lookup (PLU) into the registers. This process requires employees to remember the PLU codes of hundreds of different products, making errors inevitable. Executives at one supermarket chain, for example, told us that, on average, they sold 25 % more "medium tomatoes" than the total amount shipped to their stores because store employees often entered the PLU code for "medium tomatoes" even when customers were buying other types of tomatoes, such as "organic tomatoes" or "vine ripe tomatoes". It is reasonable to expect inventory record inaccuracy to be more accurate at retail stores where electronic point-of-sale scanning is used.

Even when processes are well-designed, employees may deliberately choose not to follow them. In an operating environment where nonconformance to designed processes is not monitored or punished, employees may choose not to carry-out activities requiring substantial effort. In some cases, the operating environment may make it challenging for employees to follow designed processes. In numerous retail chains we have observed that instead of placing merchandise that does not fit into the display shelves into storage locations where extra merchandise is supposed to be stored, store employees hide it in places within the selling floor so that they do not have to travel all the way to specified storage areas. This nonconformance causes misplaced products. Similarly, during the checkout process in a supermarket store employees sometimes choose not to scan two products that are identical in price but different in flavor (e.g., two liter bottle of Diet Coke and two liter bottle of Coke) separately, scanning one product twice instead. While the employees do not create a discrepancy between the value of the inventory sold and the amount due the store from the customer since the products are identically priced, this action does create a discrepancy in two inventory records. The recorded on-hand quantity for one product will be unnecessarily depleted by two units while the other product's record will remain at its current level despite the product leaving the store. Similar discrepancies arise when store employees do not properly record a product returned or exchanged by a customer.

Even when processes are well designed and employees have the intention of carrying out store processes, they may commit errors which lead to execution problems. Many retail activities are prone to employee error. For example, at some stores, standard operating procedures could dictate that all products have presence on the selling floor. In shelving new merchandise, employees may fail to place some products on the selling floor and instead mistakenly take them to storage

areas, leading to misplaced products. At another store, distribution center employees may pick and ship the wrong product to the store leading to discrepancies between recorded and actual inventory quantities in that store. Numerous other examples of employee errors could be observed when examining retail store processes. Cognitive psychologists have studied human error for a long time and have identified the mechanisms by which errors are generated and how they can be reduced (for more information on human error, see Reason 2002).

Finally, execution problems within the context of retail stores can also be caused by customers. Customer shopping habits, for example, contribute to misplaced products. At many stores, when customers remove products from the shelves and subsequently decide not to purchase them, they may not return the products to their appropriate location but rather place them in the wrong location in the store. These products remain misplaced until store employees find them and place them in their proper locations. Customer or employee theft is another contributor to inventory record inaccuracy. Hollinger and Langton (2003) estimate that inventory theft costs US retailers close to 1.3 % of annual sales or more than 26 billion dollars. Products that are removed from the store illegally are not removed from the inventory record until an audit is performed, the missing products identified, and the record corrected.

3 Factors That Exacerbate Execution Problems

Two empirical studies exist which examine specific drivers of misplaced products and inventory record inaccuracy. Research by DeHoratius and Raman (2008) and Ton and Raman (2006) consider these issues by comparing performance across retail stores within a chain. Both studies show large variation in execution performance across stores that are owned and operated by the same parent company, have the same incentives for store employees, use the same information technology systems, and are instructed to use the same standard operating procedures for shelving and replenishing inventory within the stores. As a result, these factors cannot explain the variation in performance. DeHoratius and Raman (2008) and Ton and Raman (2006) identify several alternative drivers of poor execution, namely inventory levels, product variety, employee turnover, lack of training, employee workload, and employee effort.

Note that the factors identified by DeHoratius and Raman (2008) and Ton and Raman (2006) contribute to execution problems by creating an operating environment that makes process conformance challenging or by making it more likely for store employees to make errors. How each of these factors contributes to execution problems is the subject of this section. We refer readers to the appendix for a description of the research methodology used including a precise identification of the independent variables used, a list of the control variables, and a brief description of the model estimation.

3.1 Inventory Levels

Proponents of Just-in-Time (JIT) manufacturing have argued repeatedly that inventory hides process problems and thus inhibits process improvements (Schonberger 1982; Hall 1983; Krafcik 1988). Production systems with high inventory levels have fewer learning opportunities and hence achieve lower quality over long term. A similar effect is observed at retail stores. Stockouts at retail stores that result from poor inventory planning or from poor execution are similar to production problems. Although these stockouts are not desirable, as they often lead to lost sales, like production problems they present opportunities for improvement. Since the likelihood of a stockout is higher at lower inventory levels, stores with lower inventory levels are likely to have more learning opportunities.

In retail stores where each product is given a specific space on the selling floor, a visual inspection of the shelves would allow store employees to identify the products that stocked out. When a product on the selling floor is stocked out, store employees could check whether the recorded inventory level for the product matches the actual quantity observed in the store. If there is a discrepancy between the recorded inventory level and the actual inventory on the selling floor, employees could investigate whether the discrepancy is due to product misplacement or record inaccuracy. If the former, employees could attempt to locate the extra units and bring them back to the appropriate location. If the latter, the retailer can create a formal quality process that lets employees adjust system inventory manually while also investigating the reason for the mismatch.

Retailers can learn from observing companies in other industries that maintain high levels of record accuracy. Arrow Electronics, a distributor of electronic parts and equipment, has close to 100 % inventory record accuracy, takes advantage of periods when inventory levels are low. Specifically, Arrow has a mechanism that triggers counts when either system or physical inventory reaches zero. If a part is physically stocked out in a location, the picking operators are instructed to verify that the system inventory for that part is also zero. Similarly, if the system inventory is zero, the picking operators are instructed to verify that the physical inventory for the part is also zero. When there is a discrepancy between the system inventory levels and physical inventory levels, warehouse operators investigate the source of the problem and when necessary make inventory adjustments to the system (Raman and Ton 2003).

Maintaining high inventory levels at retail stores can cause execution problems not only by reducing opportunities to easily identify discrepancies but also by increasing the complexity in the operating environment. All else being constant (e.g., the size of the selling area), stores with higher levels of inventory often have more units stored in storage areas. Since the replenishment process from storage areas, like most operational processes, is prone to employee errors, there are more opportunities to make errors in replenishing merchandise to the shelf. Thus, we expect more product misplacements in operating environments with high inventory levels.

Both DeHoratius and Raman (2008) and Ton and Raman (2006) provide empirical evidence to support the relationship between inventory levels and store execution. DeHoratius and Raman (2008) show that stores with higher inventory levels in a given selling area also have greater inventory record inaccuracy. Similarly, Ton and Raman (2006) show that stores with higher inventory levels per product also have a greater percentage of phantom products, defined as the products in storage areas but not on the selling floor. Ton and Raman (2010) confirm this finding using 4 years of data from the same research site.[8]

3.2 *Product Variety*

As with earlier claims that higher product variety increases the complexity in manufacturing settings (e.g., Skinner 1974; Anderson 1995; Fisher et al. 1995; MacDuffie et al. 1996; Fisher and Ittner 1999), more variety at a retail store increases the confusion and complexity in the operating environment and hence causes more process nonconformance or employee errors that lead to execution problems. Increasing product variety, for example, increases the difficulty of differentiating products during the checkout process. Consequently, store employees may scan one product multiple times without recognizing or caring that the customer is purchasing multiple different products, causing inventory record inaccuracy. Increasing product variety at a store also increases the number of steps performed in inventory replenishment at the stores. Given that stores have limited shelf space, store employees are required to move more units of products to storage areas at stores that have higher product variety. Since each step in replenishment is prone to errors, higher product variety is associated with more products that are in storage areas and not on the selling floor.

Both DeHoratius and Raman (2008) and Ton and Raman (2006) provide empirical evidence to support the relationship between product variety and store execution. DeHoratius and Raman (2008) show that stores with higher product variety also have greater inventory record inaccuracy. Similarly, Ton and Raman (2006) show that stores with more products in a given area also have a greater percentage of phantom products. Ton and Raman (2010) confirm this finding in their longitudinal study.

3.3 *Employee Turnover and Training*

The average employee turnover for US businesses in general is about 10–15 % (White 2005). Retail stores, however, experience much higher rates of employee

[8] See appendix for details of this study.

turnover. According to the National Retail Federation, the average part-time and full-time employee turnover in the retail industry is 124 % and 74 % respectively. Ton and Raman (2006) report an average employee turnover of 112 % for part-time employees and 65 % for full-time employees at Borders stores. Interestingly, the authors show that stores with higher employee turnover also have a greater percentage of phantom products, suggesting these problems may be linked.

High levels of employee turnover affect store execution in numerous ways. First, employee turnover disrupts existing operations (Dalton and Todor 1979; Bluedorn 1982). When a store employee quits the store, there is often a period of finding and training a replacement. During this period, workload for existing employees is generally higher. Higher workload may lead to more errors and consequently more execution problems. Moreover, the departure of employees often causes demoralization of existing employees (Staw 1980; Steers and Mowday 1981; Mobley 1982). Demoralization may cause existing employees to make more errors in performing their jobs.

Second, employee turnover leads to a loss of accumulated experience (Argote and Epple 1990; Nelson and Winter 1982). As employees spend more time at the stores, they become better at performing their jobs and consequently make fewer errors. Ton and Raman (2006), for example, state that as employees spend more time at Borders stores, they become more familiar with the products in their sections and as a result become better at noticing those that are missing from the selling floor.

Third, because store employees typically leave their job within a year, retailers often choose not to invest in their training. In fact, new employees receive, on average, only 7 h of training in the retail industry (Managing Customer Service 2001). As a result of limited training, new employees often start performing their jobs without a full understanding of the existing processes and their impact on store operations. Hence, they regularly commit process nonconformance (e.g., the checkout scanning example in Sect. 2.3). Ton and Raman (2006) provides empirical evidence for the positive effect of training on store execution. The authors find a negative association between percentage of phantom products and the amount of training offered at the stores.

3.4 Employee Workload

For most retailers, store labor represents the largest controllable expense at retail stores. For example, in 2003, selling, general, administrative expenses, which consist largely of store employee payroll expenses, represented approximately 20 % of retail sales.[9] Consequently, many store managers are evaluated based on how well they manage payroll expenses at their stores. When store managers reduce

[9] Source: Standard & Poor's Compustat, 427 public firms with SIC Codes between 5200 and 5999.

payroll expenses—either by reducing the number of employees at the stores or reducing the number of hours worked—the amount of workload per employee increases. With increased workload, store employees are more likely not to conform to designed processes. They are also likely to make more errors in performing their tasks. For example, a salesperson is more likely to scan two similar products that have the same price together instead of separately if he or she sees a long line of customers waiting to be checked out. It is often more difficult to observe the accuracy of scanning than to observe the speed of scanning both by customers and store managers. Ton and Raman (2006) show that stores that have higher employee workload, measured as payroll expenses as a percentage of sales, also have higher percentage of phantom products.

3.5 Employee Effort

DeHoratius and Raman (2008) argue that employee effort affects inventory record inaccuracy. When store employees exert more effort into monitoring select products, the inventory records for these products are expected to be more accurate. Employee effort, however, is unobservable. Thus, the authors use two proxies, item cost and shipping method, for employee effort. They posit that employees exert more effort into monitoring expensive than inexpensive products and thus expensive items should be more accurate than inexpensive ones. Similarly, they argue that store employees monitor items shipped directly to the retail store from the vendor more closely than those items shipped to the store from the retailer's own distribution center. We discuss each of these proxies and their findings in turn.

Inventory shrinkage is a common problem at retail stores and store employees often spend considerable effort in shrink prevention activities (DeHoratius and Raman 2007). Inventory shrinkage has a direct impact on store operating profits and shrinkage of expensive products affects store profitability more than shrinkage of less expensive products. Given that store managers are often evaluated on their financial performance, controlling inventory shrinkage of expensive products is often a key priority for store personnel. Consequently, it is not unusual for store employees to monitor expensive and inexpensive items differentially. DeHoratius and Raman (2008) show that this differential treatment leads to lower levels of record inaccuracy for expensive items relative to inexpensive ones.

DeHoratius and Raman (2008) also show that the magnitude and likelihood of inventory record inaccuracy is lower for those products shipped directly to the retail store from the vendor compared to those products shipped to the store from the retailer's own distribution center. They posit that store employees pay more attention to checking shipments that arrive from vendors than other shipment types. They do so because when the value ordered by the store exceeds the value shipped from the vendors, stores receive a credit from the vendor. Stores do not, however, receive a credit from the distribution centers unless the discrepancy between what was shipped and what was ordered exceeds a threshold more than 30 times the

average cost of a single product. Consequently, store employees pay more attention to checking shipments that arrive from vendors to ensure invoice accuracy. Moreover, shipments from the vendors tend to contain fewer products and hence easier for store employees to inspect.

4 How Execution Problems Affect Inventory Planning

Inventory planning at retail stores requires two main decisions, how much inventory to stock and when to replenish. The policies retailers establish with respect to these decisions have been shown to be critical determinants of store performance (see Tayur et al. 1999 and Graves and de Kok 2003). We use two examples to demonstrate how inventory record inaccuracy and misplaced products affect each of these decisions. These examples are described in detail in DeHoratius (2002) and Ton (2002). We then summarize research that incorporates execution problems into existing inventory planning models.

4.1 Inventory Record Inaccuracy and Inventory Planning

Management at Gamma received a letter of complaint from a regular customer noting that a specific product he sought was persistently out of stock (DeHoratius 2002). He stated that the product failed to be replenished even after bringing the stockout to the attention of the store manager. After researching the problem, Gamma management discovered that, although the product was out of stock, inventory records showed 42 units on-hand in that store. Because the inventory record showed that there was a sufficient amount of on-hand inventory to meet demand, the automatic replenishment system failed to release additional inventory to the store even though, in reality, there were no products on the shelf.

Sales records also revealed that this store had not sold a single unit of this product, a product that typically sold one unit per week per store, during the past 7 weeks. The demand forecast was then automatically updated to reflect the recent low levels of sales, namely zero sold in 7 weeks. Therefore, not only were customers unable to find the product on the shelf during the time when the product was out of stock but, even after re-stocking the shelf, their demand was less likely to be met in the future since the adjusted demand forecast reduced the target stocking quantity that needed to be maintained at the store. Moreover, it is important to note that the product might have remained out of stock until the next physical audit or cycle count had this customer not written to Gamma. Without inventory to sell the recorded quantity would remain at 42 units, never falling below the reorder point for this product.

DeHoratius and Raman (2008) found the lost revenue due to stockouts caused by record inaccuracy problems at Gamma amounted to 1.09 % of Gamma's retail sales

and 3.34 % of its gross profit. They derived this estimate from examining those items similar to the one above—items that were out of stock at the store but with a positive on-hand quantity sufficiently large so as to prevent the automated replenishment system from triggering an order.

4.2 Misplaced Products and Inventory Planning

Figure 4.3 shows the cumulative number of customers who entered a store and the cumulative sales for a particular product, a type of bread, between 8:00 am and 8:00 pm. As shown in the figure, the cumulative number of customers entering the store steadily increased from 8:00 am to 8:00 pm. The particular product, on the other hand, was selling well until about 12:30 pm, did not sell at all from 12:30 to 4:00 pm, started selling again after 4:00 pm, and stopped selling after 6:00 pm.

During both of these periods when there were no sales for this particular product, the system inventory level for this product was positive. As a result, a simple interpretation of these sales and inventory data would be that the in-stock for this product was 100 %, and that there was no demand for this product between 12:30 pm and 4 pm, and after 6 pm. The reality, however, was quite different. Between 12:30 and 4 pm, the inventory was located in the backroom, and was not available to the customers. At 6 pm the product stocked out.

Although one could argue that even if the product was available for sale no customer would have chosen to purchase it during 12:30 and 4 pm, we believe this

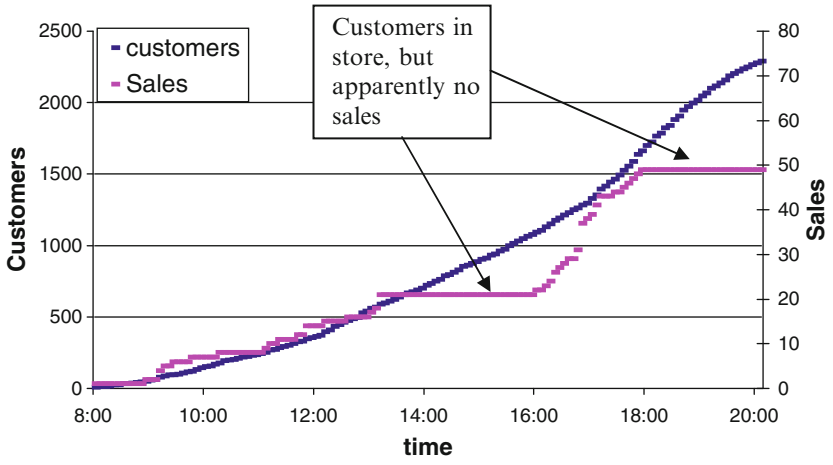


Fig. 4.3 Store sales versus customer entrances (*Source*: Ton 2002)

to be highly unlikely. Given that there was not change in the rate at which customers entered the store during this period, it is likely that the store lost sales as a result of this product misplacement.

4.3 *Incorporating Execution Problems into Existing Research Streams*

The two examples above demonstrate the challenge faced by retailers when deciding how much and when to replenish. Because retailers are unable to observe all customer actions, they rely on data to infer the actions of customers and plan accordingly. Yet, as the two examples reveal, these data can be misleading. In reality, execution errors lead to additional uncertainty and recently several researchers have begun to incorporate such uncertainty into both tactical and strategic planning models. This research shows the impact execution problems have on tactical decisions such as safety stock calculations, ordering policies, and the timing of inventory counts as well as strategic decisions such as channel coordination and technology choice.

Among the papers addressing tactical decisions in the presence of execution problems, Iglehart and Morey (1972) were the first to determine the optimal buffer stock that protects against shortages caused by record inaccuracy. Moreover, they determine the optimal frequency of inventory counts by taking into account the cost of holding buffer stock and the cost of conducting inventory counts to correct record inaccuracies. More recently, Kök and Shang (2007) derive the optimal joint inspection and replenishment policy by minimizing the total inventory and inspection cost. They create an inspection adjusted base stock policy which adjusts the replenishment order according to the level of inaccuracy and the chosen inspection policy. They argue that the order quantity needs to be increased to accommodate the additional uncertainty caused by record inaccuracy and that inaccuracy accumulates over time but that it can be corrected through inspection. Thus, if inspection cost is high, their model suggests auditing less frequently and carrying additional inventory to buffer against record inaccuracy.

Iglehart and Morey (1972) and Kök and Shang (2007) assume error in inventory records are random with a mean of zero. Thus, the discrepancy between the inventory record and actual inventory can take either sign. Moreover, both papers allow for a correlation between an item's demand and the magnitude of its inventory discrepancy. Emma (1966) and DeHoratius and Raman (2008) show empirically that the more frequently an item sells (i.e., the greater the demand) the greater the record inaccuracy. By taking into account these empirical findings, Iglehart and Morey (1972) and Kök and Shang (2007) offer practical solutions to record inaccuracy. However, one factor that limits the applicability of Kök and Shang's (2007) findings to the retail context is their backlogging assumption. In most retail settings, unfilled demand is lost rather than backlogged.

Offering an alternative solution to record inaccuracy, DeHoratius et al. (2008) propose the maintenance of a probabilistic inventory record to account for the presence of inventory record inaccuracy in retail systems. This probabilistic inventory record would take the place of the point estimate commonly used in retail to track inventory holdings. They model inventory inaccuracy through an "invisible" demand process that can either deplete or replenish physical but not recorded inventory. Using a periodic review process with unobserved lost sales, they demonstrate that the impact of inventory record inaccuracy can be mitigated through this probabilistic approach to inventory planning. Furthermore, they do so while taking into account the product characteristics that have been shown to impact record inaccuracy such as the cost of a product and its annual selling quantity (DeHoratius and Raman 2008).

Kang and Gershwin (2005) focus on one source of inventory record inaccuracy, namely theft, whereby the quantity of units recorded ends up being greater than that actually found on the retail shelf for a given item. Fleisch and Tellkamp (2005) also analyze inventory shortages caused by either record inaccuracy or misplaced products. Through simulation, these studies demonstrate that even small rates of inventory record inaccuracy and misplaced products can result in substantial lost sales and suboptimal retail performance.

Camdereli and Swaminathan (2005), Rekik et al. (2008), Atali et al. (2005), and Gaukler et al. (2007) focus primarily on strategic planning in the face of execution problems. Camdereli and Swaminathan (2005) not only derive the optimal inventory policy for a retailer that knows the proportion of its inventory that is misplaced but also show that decreasing the proportion of misplaced products impacts channel parties differently. They identify conditions for channel coordination in the face of reduced product availability due to misplaced products. Rekik et al. (2008) also examine the impact of reduced product availability due to misplaced products. Unlike Camdereli and Swaminathan (2005), the objective of Rekik et al. (2008) is to explore how the use of RFID technology can mitigate the cost of product misplacement. Gaukler et al. (2007) also examine the role of RFID by evaluating whether the use of this technology can improve in-store, shelf replenishment processes and hence product availability. They also discuss the impact of execution problems on channel coordination and the differential benefit RFID technology has among channel members. Similar to Rekik et al. (2008) and Gaukler et al. (2007), Atali et al. (2005) examine the value of RFID technology in reducing execution errors by comparing an inventory system with and without the visibility such technology provides. However, unlike the previously cited papers, Atali et al. (2005) evaluate not only product misplacement, one-sided errors that deplete inventory levels and reduce product availability, but also execution problems than can result in the actual inventory level exceeding the recorded level.

5 Future Research Opportunities

There are several research opportunities for those interested in the impact of store execution on product availability, in particular, and on retail supply chains, in general. For example, the widely accepted theoretical relationship between inventory levels and product availability is that increasing inventory levels is associated with increased service levels and thus increased store sales. The empirical findings summarized in this chapter, however, suggest that increasing inventory levels also increases the occurrence of misplaced products and inventory record inaccuracy. Thus, through its effect on store execution, increasing inventory levels may also compromise service levels. There is an opportunity for management scholars to develop models where both the direct and indirect effects of increasing inventory levels of product availability are considered and to examine empirically the direct and indirect effects of inventory levels on stores sales. In Ton and Raman's (2010) longitudinal study, the direct positive effect of increasing inventory levels on sales is larger than the indirect negative effect through store execution. There may, however, be settings where the indirect negative effects outweigh the direct positive effects.

There is also opportunity to incorporate execution problems into models that estimate demand and assess forecast accuracy in the presence of stockouts (Wecker 1978; Agrawal and Smith 1996; Nahmias 1994; Raman and Zotteri 2000; Smith and Agrawal 2000). Raman and Zotteri, for example, argue that sales data could be used along with inventory data to incorporate lost sales estimates into the estimation of demand. More specifically, the authors generate an estimate of the lost sales by using inventory data to identify when a stockout occurred. Once it is known when the stockout occurred, the historical sales rate can be appropriately extrapolated to determine lost sales during the stocked out period. Thus, both observed demand when a product is in stock (e.g., sales history) and an estimate of unobserved demand when product is out of stock (e.g., lost sales estimation) can be used to estimate the demand more accurately. Consider a common situation, however, where there are no units of inventory for the product (either due to inventory record inaccuracy or misplacement) while the inventory records show a positive value for inventory. In this situation, sales for the product will be zero, while the inventory for the product will appear positive. Consequently, the above analysis will conclude that there was no demand for the product during the period when the inventory record was inaccurate or when the product was misplaced and the demand estimation will be inaccurate. The demand estimation could be improved by assigning a probability that the inventory record is not a true reflection of reality when the system inventory level for the product is positive and there are no observed sales.

Note that efforts to compensate for execution problems with robust decision support tools and efforts to prevent execution problems through, among others, improved process design, improved conformance to designed processes and error prevention are not mutually exclusive. There is much to gain from better

understanding the ways to eliminate or reduce the prevalence of execution problems while simultaneously designing decision tools robust enough to account for the existence of execution problems. As we reviewed in this chapter, existing research has identified numerous factors that exacerbate the occurrence of execution problems. These factors are largely under the control of retail managers. For example, retail managers can choose to invest more in training or spend more on payroll expenses to reduce the occurrence of execution problems. These actions, however, result in increased costs. Given that higher profitability is the overarching goal, researchers and retail managers can develop models that optimize store profitability given the relevant costs and benefits of changes to training or payroll expenses. Moreover, retailers could institute a process improvement effort that identifies and corrects those employee actions leading to errors in product location or record inaccuracy.

Additional empirical research opportunities exist. For example, the findings of DeHoratius and Raman (2008) and Ton and Raman (2006) need to be tested using data from other retailers in order to determine their generalizability. The effect of human resource variables such as employee turnover, training, and workload on inventory record inaccuracy also needs to be examined. Moreover, opportunities exist for researchers to examine the impact of process design, technology, or employee incentives on execution. Given that most retailers do not alter process design, technology usage, or employee incentives across their own stores, researchers wishing to examine these factors would need to compare execution across several different retailers.

More research is needed to examine the relationship between storage area size and store execution. Proponents of lean production systems have argued that smaller repair areas would force employees to introduce procedures to reduce defects and hence production systems with smaller repair areas would be associated with higher quality. Most retail stores include areas for storing extra merchandise that do not fit into the display shelves. These storage areas are similar to the repair areas in manufacturing plants. Consistent with the lean production system argument, Ton and Raman (2006) hypothesize, but are unable to support with their data, a relationship between smaller storage areas and product misplacement. Nevertheless, the authors cite anecdotal evidence suggesting that smaller storage areas force store employees to better monitor products in the storage areas and to quickly replenish the selling floor with units from storage locations, lowering the level of product misplacements. Specifically, store execution suffered tremendously at one particular store when the store's storage capacity increased from one year to next.

One potential reason for the lack of statistical significance may be the imperfect measure the authors used for storage capacity. Although stores typically have multiple storage areas, Ton and Raman (2006) use size of the backroom as a surrogate for total storage in a store. In addition, it might be *how* the storage area is managed rather than its size that affects percentage of products that are in storage but not on the selling floor. Ton (2002) states that her store visits revealed a great deal of variation in the utilization of the storage areas. There were some backrooms that were very well organized, with products clearly categorized, and each shelf

well displayed with labels that indicated what merchandise was stored on that shelf. In other backrooms there were no labels on the shelves and multiple products were stacked on top of each other. Some backrooms were so messy that there were boxes and carts in between the shelves that prevented employees from gaining access to large areas of the storage space.

There is also opportunity to conduct empirical research on the consequences of poor store execution. DeHoratius and Raman (2008) and Ton and Raman (2010) show the negative effect of poor store execution on store sales. Similar research can be conducted to examine the effect of store execution on other financial or non-financial measures of store performance. This includes the impact of store execution on the performance of retail supply chain initiatives such as vendor managed inventory or collaborative planning, forecasting and replenishment (CPFR) programs.

Note that in this chapter we focused solely on the effect of execution on managing product availability. The execution problems described in this chapter have implications beyond managing product availability. Accurate inventory data may allow a company to make same-day delivery promises or to integrate online and physical store operations (see Raman and Ton 2003 and Ton and Raman 2003 for teaching case examples). Thus, researchers can identify the impact of execution problems on business strategy as well as performance.

Appendix

DeHoratius and Raman (2008)

Research Site: The authors examine the drivers of inventory record inaccuracy using data from Gamma Corporation, a large specialty retailer with over 10 billion dollars in annual sales. Gamma uses electronic point-of-sale scanning for all its sales and an automated replenishment system for inventory replenishment.

Data: The authors collected data from physical audits of 37 Gamma stores in 1999. These data included detailed information about each stock-keeping-unit (SKU) contained in each store, amounting to a total of 369,567 observations, or SKU-Store combinations. Physical audits revealed the recorded quantity (the number of inventory units for each SKU recorded to be on-hand at a specific store) as well as the actual quantity (the number of inventory units actually present at the store for each SKU). In addition to SKU level data, the authors collected both store and product category data and complemented their quantitative analysis with extensive fieldwork.

Dependent Variable: The dependent variable is the inventory record inaccuracy of each SKU in each store. Inventory record inaccuracy (IRI) is measured as the absolute difference between the recorded and actual quantity for each SKU-store combination.

Independent variables: SKU level variables include the cost of the item, its annual selling quantity, and whether the item had been shipped to the store from one of Gamma's distribution centers or directly from the vendor. Store level variables are the number of units in a given selling area, product variety, and the number of days between the current and previous physical audit.

Empirical Model: Because these data have a multi-level structure (SKUs are contained within stores and product categories), the authors fit a series of hierarchical linear models to examine the drivers of IRI. In addition to all independent variables, the empirical model includes control dummy variables for each region ($REGION\_ONE_{k,}$ $REGION\_TWO_k$). Equation (4.1) below summarizes their model.

$$\begin{aligned} IRI_{ijk} = {} & \Theta_0 + b_{00j} + c_{00k} + e_{ijk} + \pi_1{}^*\left(QUANTITY\_SOLD_{ijk}\right) + \pi_2{}^*\left(ITEM\_COST_{ijk}\right) \\ & + \pi_3{}^*\left(DOLLAR\_VOLUME_{ijk}\right) + \pi_4{}^*(VENDOR_i) + \pi_5{}^*\left(VENDOR\_COST_{ijk}\right) \\ & + \gamma_{001}{}^*(REGION\_ONE_k) + \gamma_{002}{}^*(REGION\_TWO_k) + \gamma_{003}{}^*(DENSITY_k) \\ & + \gamma_{004}{}^*(VARIETY_k) + \gamma_{005}{}^*(DAYS_k). \end{aligned} \tag{4.1}$$

where

IRI_{ijk} is the record inaccuracy of item i ($i = 1\ldots,n_{jk}$) in product category j ($j = 1\ldots,68$) and store k ($k = 1, \ldots,37$).
Θ_0 is a fixed intercept parameter.
The random main effect of product category j is $b_{00j} \sim N(0, \tau_{boo})$.
The random main effect of store k is $c_{00k} \sim N(0, \tau_{coo})$.
The random item effect is $e_{ijk} \sim N(0, \sigma^2)$.
τ_{boo}, τ_{coo}, and σ^2 define the variance in IRI between product categories, stores, and items, respectively.
π_1–π_5 are the fixed item level coefficients and γ_{001}-γ_{005} are the fixed store level coefficients.

Each of the variables is defined below:

$QUANTITY\_SOLD_{ijk}$ is the annual selling quantity of item i in product category j and store k.
$ITEM\_COST_{ijk}$ is the cost of item i in product category j and store k.
$DOLLAR\_VOLUME_{ijk}$ is the interaction between the cost of the item and its annual selling quantity.
$VENDOR_i$ is a dichotomous variable that takes the value of one if the item is shipped direct to the store from the vendor and takes the value of zero if the item is shipped to the store from the retail-owned distribution center.
$VENDOR\_COST_{ijk}$ is an interaction term between the way in which an item was shipped to the store and its cost.
$DENSITY_k$ is the total number of units in a store divided by that store's selling area (units per square foot).

$VARIETY_k$ is the number of different merchandise categories within a store
$DAYS_k$ measures the number of days between audits for a given store.

Findings: The authors find significant positive relationships between IRI and an item's annual selling quantity, store inventory density, store product variety, and the number of days since the last store audit. A significant negative relationship exists between IRI and an item's cost as well as its dollar volume. The way in which an item is shipped to the store is a significant predictor of IRI such that items shipped direct to the store from the vendor are more accurate than items shipped from the retail distribution center. This relationship, however, depends on the cost of an item. Specially, the difference between vendor-shipped and DC-shipped items is greater for inexpensive items than for expensive ones.

Ton and Raman (2010)

Research Site: The authors examine the drivers of misplaced products using data from Borders Group, a large retailer of entertainment products such as books, CDs, and DVDs. To ensure product availability, the retailer has invested heavily in information technology and merchandise planning to make sure that the right product is sent to the right store at the right time.

Data: The authors collected data from physical audits of 242 Borders stores in 1999. Physical audits provide data on the total units of inventory at the store, total number of products at the store, and the number and dollar value of the products that were present in storage areas but not on the selling floor. In addition to physical audit data, the authors collected data on store attributes and human resource characteristics. The authors complemented their empirical data with extensive fieldwork.

Dependent Variable: The dependent variable, % phantom products, is the percentage of products that are in storage areas but not on the selling floor. The authors call these products "phantom" because they are physically present in the store and often shown as available in retailers' merchandising systems, but in fact are unavailable to customers.

Independent variables: The authors use the following independent variables: inventory level per product, total number of products in a given area, size of the storage area, employee workload, employee turnover, store manager turnover, and the number of trainers at the store.

Empirical Model: The authors estimate the parameters of Eq. (4.2) using ordinary least square estimator to examine the drivers of % phantom products. In addition to all independent variables, the empirical model includes the following control variables: store sales, store age, seasonality, unemployment rate, and a dummy variable for each region. Note that, one variable, store sales, is an endogenous variable and

hence the authors employ instrumental variable estimation to cope with endogeneity. The authors use corporate sales as an instrument for store sales.

$$\begin{aligned}\%Phantom\ Products_i = {} & \beta_0 + \beta_1 Seasonality_i + \beta_2 Unemployment\ Rate_i + \beta_3 LN(Age)_i \\ & + \beta_4 Sales_i + \beta_5 Wage_i + \beta_{6j} Region_i + \beta_7 Inventory\ Depth_i \\ & + \beta_8 Product\ Density_i + \beta_9 Storage\ Size_i + \beta_{10} Labor\ Intensity_i \\ & + \beta_{101} SM\ Turnover_i + \beta_{12} FT\ Turnover_i + \beta_{13} PT\ Turnover_i \\ & + \beta_{14} Training_i + \varepsilon_i \end{aligned} \tag{4.2}$$

$$\begin{aligned} i &= 1, 2, \ldots, 242 \\ j &= 1, 2, \ldots, 17 \end{aligned}$$

Each of the variables is defined below:

% Phantom Prodcuts$_i$ is the number of products in storage but not on floor in store *i* divided by the total number of products in store *i*.

Seasonalityij is the seasonality index for month *j* in which the audit is conducted at store *i*. The seasonality index for month *j* is calculated as: $\theta_j = \dfrac{\sum_{i=1}^{242} S_{ij}}{\left(\sum_{j=1}^{12}\sum_{i=1}^{242} S_{ij}/12\right)}$

Unemployment Rate$_i$ is the unemployment rate of the metropolitan statistical area in which the store is located in 1999.

ln(*Age*)$_i$ is the natural log of the age of store *i* (in months) during the time of the audit.

Sales$_i$ is the total sales at store *i* in 1999.

Wage$_i$ is the average hourly wage at store *i* in 1999.

Region$_j$ are 17 dummy variables indicating region in which store *i* is located.

Inventory Depth$_i$ is the total number of units in store *i* divided by the number of products in store *i*.

Product Density$_i$ is the number of products in store *i* divided by the total selling area of store *i*.

Storage Size$_i$ is the backroom area of store *i* divided by the total selling area of store *i*.

*Labor Intensity$_i$*is the payroll expenses at store *i* in 1999 divided by sales at store *i* in 1999.

SM Turnover$_i$ is a dummy variable indicating the departure of store manager at store *i* in 1999.

$FT\ Turnover_i$ is the total number of full-time employees in store i that departed in 1999 divided by the average number of full-time employees in store i.
$PT\ Turnover_i$ is the total number of part-time employees in store i that departed in 1999 divided by the average number of part-time employees in store i.[10]
$Training_i$ is the total number of "trainer months" at store i in 1999.

Findings: The authors find significant positive relationships between % phantom products and inventory level per product, total number of products in a given area, employee workload, and store manager turnover. The authors find partial support for the positive relationship between employee turnover and % phantom products. The authors also find a significant negative relationship between % phantom products and the amount of training at the store.

Ton and Raman (2007)

Research Site: The authors examine the effect of product variety and inventory levels on store sales using data from Borders Group.

Data: The authors collected data from physical audits of all Borders stores from 1999 to 2002. The dataset includes 356 stores, some of which opened between 1999 and 2002. As a result the authors do not have 4 years of data for all 356 stores.

Dependent Variables: The authors use two dependent variables. First is the percentage of phantom products, products that are in storage areas but not on the selling floor. The second dependent variable is store sales.

Independent variables: The authors use the following independent variables: inventory level per product, total number of products at a store.

Empirical Model: The authors estimate the parameters of Eq. (4.3) to examine the effect of product variety and inventory levels on % phantom products and estimate the parameters of Eq. (4.4) to examine the effect of % phantom products on store sales. In both equations, the authors control for *factors that vary over time for stores and are different across stores* (seasonality, unemployment rate in the store's metropolitan statistical area, amount of labor used in a month, employee turnover, full-time employees as a percentage of total employees, store manager turnover, and the number of competitors in the local market), *factors that vary over time but are invariant across stores* (year fixed effects), and *factors that are time-invariant for a store but vary across stores* (store fixed effects).

The authors use ordinary least squares (OLS) estimators in estimating both Eqs. (4.3) and (4.4) and report the heteroskedasticity robust standard errors for OLS. In addition to OLS estimators, the authors also treat Eqs. (4.3) and (4.4) as

[10] Full-time and part-time turnover include only employees that were responsible for inventory management.

seemingly unrelated regressions (SUR) allowing for correlation in the error terms across two equations. In addition, because of autocorrelation in the error terms of Eq. (4.4), the authors consider a flexible structure of the variance covariance matrix of the errors with first-order autocorrelation and estimate the parameters of Eq. (4.4) using maximum likelihood estimation.

$$\%Phantom\ Products_{it} = \alpha_i + \lambda_t + \beta_1\ Product\ Variety_{it} + \beta_2 Inventory\ Level_{it} + X_{iy}\beta + \varepsilon_{it} \tag{4.3}$$

$$Sales_{it} = \delta_i + \phi_t + \gamma_1 \%Phantom\ Products_{it} + \gamma_2\ Product\ Variety_{it} + \gamma_3 Inventory\ Level_{it} + X_{iy}\gamma + \varepsilon_{it} \tag{4.4}$$

$$\alpha_i, \delta_i = Fixed\ effect\ for\ store\ i,\ i = 1,\ 2\ \ldots,\ 356,$$
$$in\ equations\ (1)\ \text{and}\ (2)\ respectively$$
$$\lambda_t, \phi_t = Fixed\ effect\ for\ year\ t,\ t = 1999,\ 2000,\ 2001,\ 2002$$
$$in\ equations\ (1)\ and\ (2)\ respectively$$

Each of the variables is defined below:

$\%Phantom\ Products_{it}$ is products that are in storage areas but not on floor at store i in year t at the time of the physical audit divided by the # of products at store i in year t at the time of the physical audit

$Sales_{it}$ is sales during the month preceding the audit at store i in year t

$Product\ Variety_{it}$ is the # of products at the store at the time of the physical audit at store i in year t

$Inventory\ Level_{it}$ is the # of units at the store at the time of the physical audit at store i in year t divided by the # of products at the store at the time of the physical audit at store i in year t

The vector X_{iy} represents the following control variables:

$Seasonality_j$ is the seasonality index for month j in which the audit is conducted at store. Let S_{ijt} =sales at store i in month j in year t. Then the seasonality index for month j is $\dfrac{\sum_{t=1}^{4}\sum_{i=1}^{267} S_{ijt}}{\left(\sum_{t=1}^{4}\sum_{j=1}^{12}\sum_{i=1}^{267} S_{ijt}/48\right)}$.

$Unemployment_{it}$ is the unemployment rate of the metropolitan statistical area in which the store is located during the month preceding the audit at store i in year t.

$Labor_{it}$ is the payroll expenses during the month preceding the audit at store i in year t.

$Employee\ Turnover_{it}$ is the fraction of employees that are charged with managing inventory that had left during the month preceding the audit at store i in year t.

$Proportion\ Full_{it}$ is the fraction of full-time employees during the month preceding the audit at store i in year t.

Store Manager Turnover$_{it}$ is a dummy variable that has a value of 1 if the store manager had left the company voluntarily since the last physical audit at store i in year t.

Competition$_{it}$ is the total number of Barnes & Noble and Borders stores in the area during the month preceding the audit at store i in year t.

Findings: The authors find that increasing both product variety and inventory level per product at a store is associated with an increase in % phantom products. The authors also find that an increase in % phantom products is associated with a decrease in store sales. As a result, their empirical analysis shows that through store execution, increasing product variety and inventory levels has an indirect negative effect on store sales. This indirect negative effect, however, is smaller than the direct positive effect of increasing inventory levels and product variety on store sales.

References

Agrawal, N., & Smith, S. A. (1996). Estimating negative binomial demand for retail inventory management with unobservable lost sales. *Naval Research Logistics, 43*, 839–861.

Andersen Consulting. (1996). *Where to look for incremental sales gains: The retail problem of out-of-stock merchandise*.

Anderson, P. (1995). *Technology. The Blackwell encyclopedic dictionary of organizational behavior* (pp. 557–560). Cambridge, MA: Blackwell.

Argote, L., & Epple, D. (1990). Learning curves in manufacturing. *Science, 247*, 920–924.

Atali, A., Lee, H., Özer, Ö. (2005). If the inventory manager knew: Value of RFID under imperfect inventory information, *M&SOM conference proceedings*.

Bayers, C. (2002). The last laugh. *Business 2.0, 3*(9), 86–93.

Bergman, R. P., (1988). A B count frequency selection for cycle counting supporting MRP II. *Production and Inventory Management Review*, pp. 35–36

Bluedorn, A. (1982). The theories of turnover: Causes, effects, and meaning. *Res. in the Soc. of Org., 1*, 75–128.

Brooks, R. B., & Wilson, L. W. (1993). *Inventory record accuracy. Unleashing the power of cycle counting*. Essex Junction: Oliver Wight Publications Inc.

(2005). By the numbers. *Government Executive*, **37**(6): 12.

Camdereli, A. Z., & Swaminathan, J. M. (2005). Coordination of a supply chain under misplaced inventory. *University of North Carolina Working Paper*. OTIM-2005–02.

Capital Markets Report. (June 22, 2000). http://www.erisks.com.default.asp

Cassidy, A., & Mierswinski, E., (2004), Mistakes do happen: A look at errors in consumer credit reports. *U.S. Public Interest Research Group Report*.

Dalton, D., & Todor, W. (1979). Turnover turned over: An expanded and positive perspective. *Academy of Management Review, 4*(2), 225–235.

DeHoratius, N., (2002). *Critical determinants of retail execution*. Unpublished Dissertation. Harvard Business School.

DeHoratius, N., & Raman, A. (2007). Store Manager Incentive Design and Retail Performance. *Manufacturing and Service Operations Management*, *9*(4), 518–534.

DeHoratius, N., Mersereau, A., & Schrage, L. (2008). Retail inventory management when records are inaccurate. *Manufacturing and Service Operations Management*, *10*(2), 257–277.

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

Emma, C. K. (1966). Observations on physical inventory and stock record error, Interim report 1. Department of Navy Supply Systems Command.
Emmelhainz, L. W., Emmelhainz, M. A., & Scott, J. R. (1991). Logistics implications of retail stockouts. *Journal of Business Logistics, 12*(2), 129–142.
Fisher, M. L., & Ittner, C. D. (1999). The impact of product variety on automobile assembly operations: Empirical evidence and simulation analysis. *Management Science, 45*(6), 771–786.
Fisher, M. L., Jain, A., & MacDuffie, J. P. (1995). Strategies for product variety: Lessons from the auto industry. In B. Kogut & E. Bowmanm (Eds.), *Redesigning the firm* (pp. 116–154). New York, NY: Oxford University Press.
Fleisch, E., & Tellkamp, C. (2005). Inventory inaccuracy and supply chain performance: A simulation study of retail supply chains. *International Journal of Production Economics, 95* (3), 373–385.
Flores, B. E., & Whybark, D. C. (1986). Multiple criteria ABC analysis. *International Journal of Operations and Preproduction Management, 6*(3), 38–45.
Flores, B. E., & Whybark, D. C. (1987). Implementing multiple criteria ABC analysis. *Journal of Operations Management, 7*, 79–85.
Galway, L. A., & Hanks, C. H., (1996). Data quality problems in army logistics. *MR-721-A RAND*.
Gaukler, G. M., Seifert, R. W., & Hausman, W. H. (2007). Item-level RFID in the retail supply chain. *Production and Operations Management, 16*(1), 65–76.
Graves, S. G., & de Kok, A. G. (2003). *Supply chain management: Design, coordination, and operations* (Handbooks in operations research and management science, Vol. 11). Amsterdam: Elsevier Publishers.
Gruen, T. W., Corsten, D. S., & Bharadwaj, S. (2002). *Retail out-of-stocks: A worldwide examination of extent, causes and consumer responses.* Grocery Manufacturers of America, The Food Marketing Institute, and CIES.
Hall, R. V. (1983). *Zero inventories*. Homewood: Dow Jones-Irvin Inc.
Hart, M. K. (1998). Improving inventory accuracy using control charts. *Production and Inventory Management, 11*, 44–48.
Hollinger, R. C., & Langton, L. (2003). *National retail security survey final report*. www.soc.ufl.edu/srp.htm.
Iglehart, D. L., & Morey, R. C. (1972). Inventory systems with imperfect asset information. *Management Science, 18*(8), B388–B394.
Kang, Y., & Gershwin, S. B. (2005). Information inaccuracy in inventory systems: Stock loss and stockout. *IIE Transactions, 37*(9), 843–59.
Kök, G. A., & Shang, K. H. (2007). Inspection and replenishment policies for systems with inventory record inaccuracy. *Manufacturing and Service Operations Management, 9*(2), 185–205.
Krafcik, J. F. (1988). Triumph of the lean production system. *Sloan Management Review, 30*(1), 41–51.
Krajewski, L. J., King, B. E., Ritzman, L. P., & Wong, D. S. (1987). Kanban, MRP, and Shaping the manufacturing environment. *Management Science, 33*(1), 39–57.
Kurt Salmon Associates. (2002). *Biennial Consumer Outlook Survey*.
Laudon, K. C. (1986). Data quality and due process in large interorganizational record systems. *Communications of the ACM, 29*, 4–11.
MacDuffie, J. P., Sethuraman, K., & Fisher, M. L. (1996). Product variety and manufacturing performance: Evidence from the International Automotive Assembly Plant Study. *Management Science, 42*(3), 350–369.
Managing Customer Service. (2001). Hiring and training tips From Fortune's top-rated employer, August 1.
McClain, J. O., Thomas, L. J., & Mazzola, J. B. (1992). *Operations management: Production of goods and services*. Upper Saddle River: Prentice-Hall/Pearson Education.
McCutcheon, C. (1999). Pentagon's ongoing record of billions in lost inventory leads hill to demand change. *Cong. Quart. Week*, **57**(18): 1041.
Millet, I. (1994). A novena to Saint Anthony, or how to find inventory by not looking. *Interfaces, 24*, 69–75.

Mobley, W. (1982). *Employee turnover: Causes, consequences, and control*. Reading: Addison-Wesley.
Nahmias, S. (1994). Demand estimation in lost sales inventory systems. *Naval Research Logistics, 41*, 739–757.
Nelson, R., & Winter, S. (1982). *An evolutionary theory of economic change*. Cambridge: Harvard University Press.
Raman, A., DeHoratius, N., & Ton, Z. (2001). Execution: The missing link in retail operations. *California Management Review, 43*(3), 136–152.
Raman, A., & Ton, Z. (2003). *Operational execution at Arrow Electronics*. Harvard Business School Case.
Raman, A., & Zotteri, G., (2000). *Estimating retail demand and lost sales*. Harvard Business School Working Paper.
Reason, J. (2002). *Human error*. Cambridge: Cambridge University Press.
Redman, T. (1995). Improve data quality for competitive advantage. *Sloan Management Review, 36*, 99–107.
Rekik, Y., Sahin, E., & Dallery, Y. (2008). Analysis of the impact of RFID technology on reducing misplacement errors at the retailer. *International Journal of Production Economics, 12*(1).
Rinehart, R. F. (1960). Effects and causes of discrepancies in supply operations. *Operations Research, 8*(4), 543–564.
Rout, W. (1976). That damn storeroom. *Production and Inventory Management, 17*, 22–29.
Schonberger, R. J. (1982). *Japanese manufacturing techniques: Nine hidden lessons in simplicity*. New York, NY: The Free Press.
Schrady, D. A. (1970). Operational definitions of record accuracy. *Naval Research Logistics Quarterly, 17*(1), 133–142.
Sheppard, G. M., & Brown, K. A. (1993). Predicting inventory record-keeping errors with discriminant analysis: A field experiment. *International Journal of Production Economics, 32*, 39–51.
Skinner, W. (1974). The focused factory. *Harvard Business Review, 53*, 113–121.
Smith, S. A., & Agrawal, N. (2000). Management of multi-item retail inventory systems with demand substitution. *Operations Research, 48*(1), 50–64.
Staw, B. (1980). The consequences of turnover. *Journal of Occupational Behavior, 1*(4), 253–273.
Steers, R., & Mowday, R. (1981). Employee turnover and post-decision accommodation processes. *Research in Organizational Behaviour, 3*, 235–281.
Tallman, J. (1976). A practical approach to installing a Cycle Inventory Program. *Production and Inventory Management, 17*(4), 1–16.
Tayur, S., Ganeshan, R., & Magazine, M. (1999). *Quantitative models for supply chain management*. Boston: Kluwer Academic Publishers.
Ton, Z. (2002). *The role of store execution in managing product availability*. Unpublished Dissertation. Harvard Business School.
Ton, Z., & Raman, A. (2003). Borders Group, Inc. *Harvard Business School Case*.
Ton, Z., & Raman, A. (2006). Cross sectional analysis of phantom products at retail stores. *Harvard Business School Working Paper*.
Ton, Z., & Raman, A. (2010). The effect of product variety and inventory levels on retail store sales: A longitudinal study. *Production and Operations Management, 19*(5), 546–560.
Wecker, W. E. (1978). Predicting demand from sales data in the presence of stockouts. *Management Science, 24*, 1043–1054.
White, E. (2005). New recipe: To keep employees, Domino's decides it's not all about pay. *The Wall Street Journal, 17*, A1.
Woellert, L. (2004). Shortchanged on long distance. *BusinessWeek,* **3889**: 13
Woolsey, G. (1977). The warehouse model that couldn't be and the inventory that couldn't be zero. *Interfaces, 7*(3), 14–17.
Young, S. T., & Nie, W. D. (1992). A cycle-count model considering inventory policy and record variance. *Production and Inventory Management, 33*(1), 11–16.

Chapter 5
Analytics for Operational Visibility in the Retail Store: The Cases of Censored Demand and Inventory Record Inaccuracy

Li Chen and Adam J. Mersereau

1 Introduction

A retail store is a system in which customers, associates, and merchandise interact which each other over time to produce sales and profits for the firm (Fig. 5.1). The store, however, is far from a black box from the manager's perspective. Retail managers have a number of operational levers to influence these interactions, including store design, assortment planning, pricing, inventory control, and staffing. Retail managers also have some visibility into what transpires in the store.

Historically, this visibility has been limited to inventory positions, staff schedules, and, since the emergence of barcode technologies in the 1970s, point-of-sale (POS) data. Recent years, however, have seen a heightened interest among practitioners in store visibility—how a retailer can gain clearer visibility and how it can best use this visibility for operational and marketing advantage. Citing opportunities brought by existing and new retail data sources, a recent report by the McKinsey Global Institute highlights retail's "tremendous upside potential across the industry for individual players to expand and improve their use of big data" (McKinsey Global Institute 2011).

We believe that one factor contributing to this interest in visibility is the continued rise of internet retailing (i.e., e-commerce), which continues to grow as a fraction of the overall retail industry. A commonly cited advantage enjoyed by e-commerce retailers compared with their brick-and-mortar cousins is their visibility into the sales process, given that interactions of customers with the e-commerce retail site (and with associates and inventory, when applicable) can be (and are)

L. Chen
Fuqua School of Business, Duke University, Durham, NC, USA

A.J. Mersereau (✉)
Kenan-Flagler Business School, University of North Carolina, Chapel Hill, NC, USA
e-mail: ajm@unc.edu

N. Agrawal, S.A. Smith (eds.), *Retail Supply Chain Management*,
International Series in Operations Research & Management Science 223,
DOI 10.1007/978-1-4899-7562-1_5

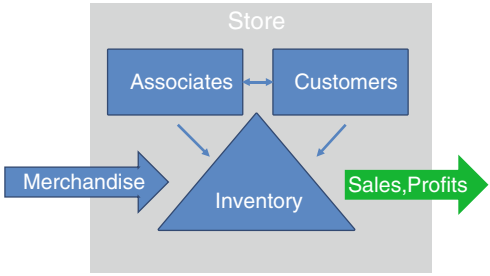


Fig. 5.1 A simplified view of in-store retail operations

recorded and mined for information. Customer clickstreams can reveal detailed insights into customer behavior. Furthermore, the customer experience is largely decoupled from firm operations in e-commerce retail, enabling tighter monitoring of inventory control and customer service. Indeed, a significant challenge of modern in-store retailing, seen in the push for "omni-channel" retailing (Brynjolfsson et al. 2013), is learning how best to compete with, complement, and learn from the e-commerce channel. Part of the answer seems to be finer visibility into and control of the in-store environment.

A second factor behind the increased interest in retail visibility is the emergence of modern technologies for in-store data collection as well as information technologies for capturing, storing, and analyzing data from these sources. These technologies bring the promise of a revolution in retail operations by offering visibility at a more granular level of detail and a finer time scale. Examples of such technologies include radio-frequency identification (RFID) and traffic counters, which have existed for a number of years but whose uses are still being explored and evaluated, and new technologies such as smart shopping carts, video monitoring, and cell phone tracking. In our investigation of these technologies and in discussions with practitioners and academics, we have encountered both optimism and skepticism about them. It is clear that new approaches are needed to translate these data sources into meaningful insights and profitable decisions and to evaluate the technologies. We will discuss some of these new visibility technologies and the associated research opportunities in Sect. 4.

The main goal of this chapter is to provide a review of two substantial literatures on in-store retail management that deal with imperfect visibility, namely demand censoring and inventory record inaccuracy. We believe that these two literatures, though largely disjoint from each other, share common features and themes that make them instructive for other problems involving in-store visibility.

Inventory Management with Censored Demand Observations Retail demand data are typically captured by POS transactions. However, POS data present an imperfect observation of true demand due to the demand censoring effect: when the actual demand exceeds the available inventory level, the excess demand is not captured by the POS data. The demand censoring problem is more prominent in brick-and-mortar stores than in online stores, because the latter can monitor and track customer purchases closely to alleviate such a problem. Academic researchers

have long recognized the need to account for this censoring effect in demand forecasting and inventory management (e.g., Conrad 1976; Wecker 1978). This literature has been primarily centered on methodologies for dealing with the imperfect demand observations.

Inventory Management with Inaccurate Inventory Records Computerized inventory positions, which accumulate POS and store receipts on a daily basis, form the basis of automated replenishment policies for many retailers. There is ample evidence that such logical inventory records do not match the physical inventories on store shelves due to shrinkage, misplacement, and transaction errors (see DeHoratius and Ton 2015). In other words, retail managers have imperfect visibility into inventory in the store. A substantial literature on managing inventories given this reality has grown in the last decade, featuring diverse assumptions on error processes, decisions, and observability.

Given the common challenges in incorporating imperfect information into operational models, it is not surprising that both literatures use overlapping methodologies such as various learning and optimization paradigms. These two literatures also yield some common insights. One such insight is that lack of visibility can be costly, and if not properly accounted for can erase the gains from sophisticated, optimized policies. A second is that intelligent analytics can substitute for visibility in some cases. A third is that analytical models can help measure the return on investment of new visibility technologies by evaluating the best performance possible without visibility.

For ease of reference for readers who may be interested in only one of the topics, we have written the reviews of these two literatures to be largely self-contained. The rest of the chapter is organized as follows. In Sect. 2, we review the literature on the demand censoring problem. We discuss three types of models: Bayesian models with perishable inventory, Bayesian models with nonperishable inventory, and nonparametric models. We conclude the section by comparing the Bayesian and nonparametric models and discussing future research opportunities. In Sect. 3, we review the literature on the inventory record inaccuracy problem. Specifically, we provide a basic illustrative model for the problem, discuss the modeling issues and tradeoffs, review specific models from the literature, and conclude with a discussion of important open research questions. Section 4 discusses emerging visibility technologies and future research opportunities more generally in the general area of in-store visibility.

2 Models of Demand Censoring

In most retail environments, when inventory runs out, the unmet demand is lost and not observed. As a result, the sales data are censored by the available inventory level. When the demand distribution is known, this is a classic inventory problem involving lost sales (see Zipkin 2000 and references therein). However, if the demand distribution is not known, which is often the case for a new product

introduction, one has to rely on potentially censored data to estimate the unknown demand. Intuitively, if this partial observability of demand is not factored into the estimation procedure, the demand estimate will be biased low (Wecker 1978). If the low demand estimate is subsequently used to determine inventory stocking decisions, the resulting inventory level will also be biased low and thus will lead to more lost sales and an even lower future demand estimate. To avoid this potential vicious cycle, it is important to take into account the data censoring effect in demand estimation and inventory control decisions. In other words, we need to develop "intelligent" methods to narrow the performance gap between a system with imperfect demand data and a system with full-visibility.

Consider the case in which the demands in each period, denoted by D_t, are independently and identically distributed (i.i.d.). The demand D_t here could be a residual variable after removing seasonality and promotional effects. Let us further assume that the demand probability density function, denoted by $f(\xi|\theta)$, has an unknown parameter θ, with $\theta \in \Theta$. Let $F(\xi|\theta)$ denote the cumulative distribution function (CDF) and $\overline{F}(\xi|\theta) = 1 - F(\xi|\theta)$ the complementary CDF.

Also let y_t denote the inventory level in period t. Then the sales in period t is given by $\min\{D_t, y_t\}$. If $D_t = \xi_t < y_t$, then the demand information is observed exactly, and the likelihood function is given by $f(\xi_t|\theta)$. On the other hand, if $D_t \geq y_t$, then the demand information is censored by the inventory level y_t; all we know is that the actual demand is greater than or equal to y_t, so the likelihood function is given by $\overline{F}(y_t|\theta)$.

Suppose that there are n historical sales observations. Without loss of generality, let the first j observations be the exact demand observations, i.e., $\xi_1, \ldots, \xi_j$, and let the remaining $n - j$ observations be the censored demand observations, i.e., $y_{j+1}, \ldots, y_n$. We can write the joint likelihood function as

$$\prod_{i=1}^{j} f(\xi_i|\theta) \cdot \prod_{i=j+1}^{n} \overline{F}(y_i|\theta).$$

By maximizing this expression over θ we obtain the maximum likelihood estimator (MLE) of the unknown demand parameter.

Conrad (1976) recognizes the difference between sales and demand data and proposes the above MLE method for Poisson demand. Nahmias (1994) further considers the demand censoring problem for normal demand, and provides three estimators: the MLE estimator, the best linear unbiased estimator, and a simplified estimator based on three sample statistics. He compares the performance of these three estimators by simulation. Agrawal and Smith (1996) find that the negative binomial distribution fits their empirical data significantly better than the Poisson and normal distribution, and develop estimators for the negative binomial distribution under demand censoring. Anupindi et al. (1998) apply the MLE method to estimate the Poisson demands of multiple substitutable products for a vending machine data set. In their problem, product stockouts result in only partial lost sales due to substitution. They develop an expectation-maximization (EM) method

to account for *missing* stockout information in periodical inventory data. For a similar demand estimation problem, Conlon and Mortimer (2013) develop an EM method under a discrete choice model and demonstrate that failing to account for stockouts correctly can lead to biased demand estimates. Vulcano et al. (2012) further develop an efficient EM algorithm under the multinomial logit choice model, where they treat the observed demand as an incomplete observation of the primary demand (i.e., the would-be demand if all products were available for sale). Musalem et al. (2010) develop an alternative Bayesian estimation method based on data augmentation (i.e., imputing the entire sequence of sales) with Markov chain Monte Carlo methods.

To apply a Bayesian method to the estimation problem at hand, let $\pi(\theta)$ denote an initial prior belief on the unknown demand parameter θ. The posterior belief of θ given the same n historical sales observations as before can be written as

$$\pi(\theta|\xi_1, \ldots, \xi_j, y_{j+1}, \ldots, y_n) = \frac{\pi(\theta)\prod_{i=1}^{j} f(\xi_i|\theta) \cdot \prod_{i=j+1}^{n} \overline{F}(y_i|\theta)}{\int_{\Theta} \pi(\theta')\prod_{i=1}^{j} f(\xi_i|\theta') \cdot \prod_{i=j+1}^{n} \overline{F}(y_i|\theta')d\theta'}.$$

As with the MLE case, the ordering of the demand observations does not affect the Bayesian posterior because of the product form of the likelihood function. For an N-point discrete demand distribution with an N-dimensional beta prior, Silver (1993) derives a recursive formula for computing the Bayesian posterior expected values of the N probability masses under demand censoring.

When demand is fully observable, the above Bayesian updating procedure can be greatly simplified with conjugate prior distribution families—one only needs to update the corresponding sufficient statistic of the conjugate prior (see DeGroot 1970 for a detailed discussion of this topic). However, when demand is censored due to unobserved lost sales, most common conjugate prior distribution families do not apply. In particular, Braden and Freimer (1991) conjecture that the distributions that entail a sufficient statistic under demand censoring, termed the "newsvendor distribution," are limited to the following distribution family:

$$\overline{F}(\xi|\theta) = e^{\eta(\theta)b(\xi)},$$

where $\eta(\cdot)$ and $b(\cdot)$ are real-valued functions. Examples of such distributions include the exponential distribution, the Weibull distribution, certain bounded support distributions and certain bimodal distributions (see Braden and Freimer 1991). Specifically, when $\eta(\theta) = -\theta$ and $b(\xi) = \xi^k$ with fixed $k > 0$, the newsvendor distribution takes the form of the Weibull distribution. Below we use

the Weibull distribution to illustrate the Bayesian updating scheme under demand censoring.

Under the Weibull distribution, the demand density function is given by

$$f(\xi|\theta) = k\theta\xi^{k-1}e^{-\theta\xi^k} \quad \text{for} \quad \xi \geq 0.$$

Let us further assume that the initial prior follows a gamma distribution with the shape parameter $a > 0$ and the scale parameter $S > 0$, i.e.,

$$\pi(\theta) = \frac{S^a\theta^{a-1}e^{-S\theta}}{\Gamma(a)} \quad \text{for} \quad \theta \geq 0.$$

Thus, given the same n historical sales observations as before, it is easy to verify that the posterior also follows a gamma distribution with the updated shape and scale parameters given by $a + j$ and $S + \sum_{i=1}^{j} \xi_i^k + \sum_{i=j+1}^{n} y_i^k$, respectively. In other words, the shape parameter increases by one only when an exact demand observation is made, and the scale parameter increases by $(\min\{\xi_t, y_t\})^k$ every period.

An advantage of the Bayesian method over the MLE method is that one can integrate demand estimation together with optimal control, and formulate the joint estimation and optimization problem as a Bayesian dynamic program. In a seminal paper, Scarf (1959) first studies such a joint estimation and optimization problem when demand information is fully observable (i.e., without demand censoring). Scarf (1960) further shows the dimensionality of the Bayesian dynamic program can be reduced for the gamma-gamma conjugate prior distribution family. Azoury (1985) extends Scarf's state-space reduction technique to various conjugate prior distribution families, such as the Pareto-uniform and the gamma-Weibull conjugate priors. Under certain suitable conditions, Lovejoy (1990) shows that the Bayesian dynamic program can be simplified to a single-period optimization problem. When demand is censored due to unobserved lost sales, the joint estimation and optimization problem becomes much more challenging. Below we provide a review of the existing literature on this subject.

2.1 Bayesian Models with Perishable Inventory

Consider a periodic-review inventory control problem for a single product. The product is stocked and sold for T periods. At the beginning of each period t ($t = 1, \ldots, T$), an inventory level y_t is chosen to minimize the total inventory holding and stockout penalty costs. The production leadtime is assumed to be negligible, so the inventory level is achieved immediately after the decision. Here we also assume the product is perishable and cannot be carried over to meet

demands in subsequent periods. In this case, the on-hand inventory at the beginning of a period is always zero.

At the end of each period, a unit holding cost h or a unit penalty cost p is charged for any leftover inventory or unsatisfied demand, respectively. The purchase cost of the product is omitted in our formulation as it can be normalized to zero with the standard technique of Heyman and Sobel (1984). The terminal value at the end of the planning horizon is assumed to be zero.

Let $\pi_t(\theta)$ denote the prior belief of the unknown demand parameter θ at the beginning of period t. The predictive demand density in period t is given by $\int_{\Theta} f(\xi|\theta)\pi_t(\theta)d\theta$. Given the inventory level y, the single-period expected inventory holding and stockout penalty cost, denoted by $L_t(y, \pi_t)$, can be expressed as

$$\begin{aligned} L_t(y,\pi_t) &= h\mathsf{E}_{D_t|\pi_t}\left[(y-D_t)^+\right] + p\mathsf{E}_{D_t|\pi_t}\left[(D_t-y)^+\right] \\ &= h\int_{\Theta}\int_{0}^{y}(y-\xi)f(\xi|\theta)\pi_t(\theta)d\theta d\xi + p\int_{\Theta}\int_{y}^{\infty}(\xi-y)f(\xi|\theta)\pi_t(\theta)d\theta d\xi, \end{aligned}$$

where $(\cdot)^+ = \max\{\cdot, 0\}$.

Let $V_t(\pi_t)$ denote the cost-to-go function from period t given the prior π_t. Then the Bayesian dynamic program optimality equations can be written as, for $t = 1, \ldots, T$,

$$\begin{aligned} V_t(\pi_t) &= \min_{y\geq 0}\{G_t(y,\pi_t)\} \\ &= \min_{y\geq 0}\left\{ L_t(y,\pi_t) + \int_{\Theta}\int_{0}^{y} V_{t+1}\left(\frac{f(\xi|\cdot)\pi_t(\cdot)}{\int_{\Theta} f(\xi|\theta)\pi_t(\theta)d\theta}\right) f(\xi|\theta)\pi_t(\theta)d\theta d\xi \right. \\ &\quad \left. + V_{t+1}\left(\frac{\overline{F}(y|\cdot)\pi_t(\cdot)}{\int_{\Theta}\overline{F}(y|\theta)\pi_t(\theta)d\theta}\right)\int_{\Theta}\overline{F}(y|\theta)\pi_t(\theta)d\theta \right\}, \end{aligned}$$

with $V_{T+1}(\cdot) = 0$. Let $y_t^p = \text{argmin}_{y\geq 0}\{G_t(y,\pi_t)\}$ denote the optimal inventory decision in the above problem. Also let $y_t^m = \text{argmin}_{y\geq 0}\{L_t(y,\pi_t)\}$ denote the myopic inventory decision in the problem. Note that in the case with no censoring, the myopic decision is in fact optimal in each period.

Intuitively, under demand censoring, one would stock more than the myopic inventory level to increase the chance of having an exact demand observation, i.e., $y_t^p \geq y_t^m$ for any common prior π_t. This is indeed true for arbitrary prior and demand distributions. Harpaz et al. (1982) first show this "stock more" insight under a general production output model. The same insight is shown to hold for the multiperiod newsvendor problem as described above by Ding et al. (2002), amended later by Lu et al. (2005) and Bensoussan et al. (2009). This insight is further extended to

price-dependent demand models by Bisi and Dada (2007). Using the unnormalized prior technique developed in Bensoussan et al. (2005), Bensoussan et al. (2007a) show that an optimal policy exists and the "stock more" insight holds for an infinite-horizon problem.

To demonstrate this insight, let us examine the derivative of the dynamic program objective function below (Lu et al. 2008):

$$G_t'(y,\pi_t) = L_t'(y,\pi_t) + \left[V_{t+1}\left(\frac{f(y|\cdot)\pi_t(\cdot)}{\int_\Theta f(y|\theta)\pi_t(\theta)d\theta}\right) - \tilde{V}_{t+1}\left(\frac{f(y|\cdot)\pi_t(\cdot)}{\int_\Theta f(y|\theta)\pi_t(\theta)d\theta}\right)\right] \int_\Theta f(y|\theta)\pi_t(\theta)d\theta,$$

where $\tilde{V}_{t+1}(\cdot)$ is the expected cost when a suboptimal inventory policy, computed along each sample path assuming observation y was censored, is evaluated based on demand beliefs updated assuming y was uncensored. Thus, it is clear that $V_{t+1}(\cdot) \le \tilde{V}_{t+1}(\cdot)$, and we have $G_t'(y,\pi_t) \le L_t'(y,\pi_t)$. Hence, it follows that $y_t^p \ge y_t^m$ for any common prior π_t.

While this is an elegant structural result for the problem, computing the optimal inventory decision is still nontrivial. Easy-to-compute solutions are available only for certain conjugate prior distribution families. For example, Lariviere and Porteus (1999) derive a closed-form formula for the optimal inventory decision under the exponential demand distribution with a gamma prior. Bisi et al. (2011) further obtain a recursive formula for the more general Weibull demand distribution with a gamma prior. For general prior and demand distributions, Chen (2010) shows that the derivative of the dynamic program objective function can be computed by a recursive equation, but the dimensionality of the problem remains an obstacle for solving problems with relatively long time horizons.

2.2 *Bayesian Models with Nonperishable Inventory*

Now let us consider a more general case in which the product is nonperishable and can be carried over to meet demands in subsequent periods. In this case, the on-hand inventory at the beginning of a period is no longer zero, and we need to introduce an additional inventory state into the Bayesian dynamic program.

Let $V_t(x,\pi_t)$ denote the cost-to-go function from period t, given the on-hand inventory level x and the prior π_t. Then the Bayesian dynamic program optimality equations can be written as, for $t = 1,\ldots,T$,

$$V_t(x,\pi_t) = \min_{y\geq x}\{G_t(y,\pi_t)\}$$
$$= \min_{y\geq x}\left\{ L_t(y,\pi_t) + \int_{\Theta}\int_{0}^{y} V_{t+1}\left(y-\xi, \frac{f(\xi|\cdot)\pi_t(\cdot)}{\int_{\Theta} f(\xi|\theta)\pi_t(\theta)d\theta}\right) f(\xi|\theta)\pi_t(\theta)d\theta d\xi \right.$$
$$\left. + V_{t+1}\left(0, \frac{\overline{F}(y|\cdot)\pi_t(\cdot)}{\int_{\Theta}\overline{F}(y|\theta)\pi_t(\theta)d\theta}\right)\int_{\Theta}\overline{F}(y|\theta)\pi_t(\theta)d\theta \right\},$$

with $V_{T+1}(\cdot,\cdot) = 0$. Let $y_t^* = \text{argmin}_{y\geq 0}\{G_t(y,\pi_t)\}$ denote the optimal inventory decision to the above problem. Bensoussan et al. (2008) show that an optimal policy also exists for the infinite-horizon problem.

Extending the derivative result of Lu et al. (2008), Chen (2010) shows that the derivative of the above objective function can be written as

$$G_t'(y,\pi_t) = L_t'(y,\pi_t) + \int_{\Theta}\int_{0}^{y} V_{t+1}'\left(y-\xi, \frac{f(\xi|\cdot)\pi_t(\cdot)}{\int_{\Theta} f(\xi|\theta)\pi_t(\theta)d\theta}\right) f(\xi|\theta)\pi_t(\theta)d\theta d\xi$$
$$+\left[V_{t+1}\left(0, \frac{f(y|\cdot)\pi_t(\cdot)}{\int_{\Theta} f(y|\theta)\pi_t(\theta)d\theta}\right) \right.$$
$$\left. -\tilde{V}_{t+1}\left(0, \frac{f(y|\cdot)\pi_t(\cdot)}{\int_{\Theta} f(y|\theta)\pi_t(\theta)d\theta}\right)\right]\int_{\Theta} f(y|\theta)\pi_t(\theta)d\theta,$$

where $\tilde{V}_{t+1}(0,\cdot)$ is a generalization of $\tilde{V}_{t+1}(\cdot)$ in the perishable inventory case with zero starting inventory. Thus, we have $V_{t+1}(0,\cdot) \leq \tilde{V}_{t+1}(0,\cdot)$. But, on the other hand, we have $V_{t+1}'(y-\xi,\cdot) \geq 0$. Hence, $G_t'(y,\pi_t)$ can be either greater or less than $L_t'(y,\pi_t)$, implying that $y_t{}^* \geq y_t{}^m$ may not hold in this case. Thus, the "stock more" result in the perishable inventory case does not extend to the nonperishable inventory case when the optimal inventory decision is compared with the myopic decision.

Nevertheless, we can show that it is optimal to "stock more" than in a system without demand censoring. Since the myopic decision is optimal in a perishable inventory system without demand censoring, this can be seen as a generalization of the "stock more" result in the perishable inventory case. Let $V_t^o(x,\pi_t)$ denote the cost-to-go function from period t, given the on-hand inventory level x and the prior π_t for a system without demand censoring. Then the Bayesian dynamic program optimality equations can be written as, for $t = 1,\ldots,T$,

$$V_t^o(x,\pi_t) = \min_{y\geq x}\{G_t^o(y,\pi_t)\}$$
$$= \min_{y\geq x}\left\{L_t(y,\pi_t) + \int_{\Theta}\int_{0}^{\infty} V_{t+1}^o\left((y-\xi)^+, \frac{f(\xi|\cdot)\pi_t(\cdot)}{\int_{\Theta} f(\xi|\theta)\pi_t(\theta)d\theta}\right) f(\xi|\theta)\pi_t(\theta)d\theta d\xi\right\},$$

with $V_{T+1}^o(\cdot,\cdot)=0$. Let $y_t^o = \text{argmin}_{y\geq 0}\{G_t^o(y,\pi_t)\}$ denote the optimal inventory decision to the above problem.

Chen and Plambeck (2008) show that $y_t^* \geq y_t^o$ for any common prior π_t under the general discrete demand distribution. For a general continuous demand distribution, it is easy to verify that the derivative of $G_t^o(y,\cdot)$ is given by

$$G_t^{o'}(y,\pi_t) = L_t'(y,\pi_t) + \int_{\Theta}\int_{0}^{y} V_{t+1}^{o\,'}\left(y-\xi, \frac{f(\xi|\cdot)\pi_t(\cdot)}{\int_{\Theta} f(\xi|\theta)\pi_t(\theta)d\theta}\right) f(\xi|\theta)\pi_t(\theta)d\theta d\xi$$

By backward induction, we can show that $V_{t+1}^{o\,'}(y-\xi,\cdot) \geq V_{t+1}'(y-\xi,\cdot)$. Hence, it follows that $G_t'(y,\pi_t) \leq G_t^{o'}(y,\pi_t)$, and we have $y_t^* \geq y_t^o$ for any common prior π_t.

Computing the optimal inventory decision for this problem is even more complex than for the perishable inventory case. Leveraging the dimensionality reduction technique developed by Scarf (1960) and Azoury (1985), Lariviere and Porteus (1999) show that this problem can be reduced to a two-dimensional dynamic program under the Weibull demand distribution with a gamma prior. Bisi et al. (2011) further show that the Weibull distribution is the only distribution that allows for such a dimensionality reduction technique for the problem. They also show that the dynamic program objective function is convex under the exponential demand distribution (a special case of the Weibull distribution when $k=1$), but is generally non-convex under other demand distributions.

From the generalized "stock more" result, a natural lower bound for the optimal inventory decision is given by y_t^o. This can be relatively easy to compute, benefiting from the fact that the corresponding Bayesian dynamic program is convex (see Scarf 1959). By the dimensionality reduction technique developed by Scarf (1960) and Azoury (1985), we can compute y_t^o easily for an array of conjugate prior distribution families. However, for general prior and demand distributions, computing y_t^o is still subject to the curse of dimensionality. Lu et al. (2007) derive an upper bound for the optimal inventory decision based on the first-order condition. However, their upper bound works only for certain prior and demand distributions. Chen (2010) further derives a set of upper bounds for the optimal inventory decision that works for all prior and demand distributions. For a fairly general monotone likelihood-ratio distribution family, he derives relaxed but easy-to-compute lower and upper bounds along any sample path. He also proposes two effective heuristics based on the solution bound results and the first-order condition.

2.3 Nonparametric Models

In addition to the Bayesian (parametric) models reviewed above, there is also a stream of research on the demand censoring problem based on nonparametric approaches. Under the nonparametric models, one makes no parametric assumptions on the underlying demand distribution, but employs an adaptive data-driven ordering policy that ensures the system performance converges to the optimal performance in the long run. It is worth noting here that the expected cost in each period in this literature is typically computed fixing the (unknown) true demand parameter θ. This differs from the Bayesian models, where such cost is integrated over the updated prior belief of θ, which could be influenced by the inventory decisions in the past.

Given the inventory decision y, the single-period newsvendor cost function is given by

$$L_t(y) = h \cdot \mathsf{E}_{D_t}\left[(y - D_t)^+\right] + p \cdot \mathsf{E}_{D_t}\left[(D_t - y)^+\right].$$

Let $y^* = \text{argmin}_{y \geq 0} L_t(y)$, and let L^* denote the resulting optimal cost (note that D_t is i.i.d., so the optimal decision in each period is stationary). It is easy to verify that the derivative of $L_t(y)$ is given by

$$L_t^{'}(y) = h \cdot \Pr(D_t < y) - p \cdot \Pr(D_t \geq y).$$

Thus, an unbiased sample-path estimate of the subgradient of $L_t(y)$ at y can be written as

$$H_t(y) = \begin{cases} h, & \text{if} \quad D_t < y, \\ -p, & \text{if} \quad D_t \geq y. \end{cases}$$

Using the above subgradient estimate, Burnetas and Smith (2000) propose the following simple adaptive ordering policy for the perishable inventory case:

$$y_{t+1} = y_t - \frac{y_t}{(h+p)t} \cdot H_t(y_t).$$

They show that under this ordering policy $\lim_{T\to\infty} \mathsf{E}[\Sigma_{t=1}^{T} L_t(y_t)/T] = L^*$ and y_t converges to y^* with probability one. They further extend this policy to a joint pricing and inventory ordering problem. Godfrey and Powell (2001) propose a similar sample-path subgradient estimate to successively approximate the newsvendor cost function with a sequence of piecewise-linear functions under demand censoring. A variant of their algorithm is shown to be asymptotically optimal under certain conditions (e.g., discrete demands) by Powell et al. (2004).

Huh and Rusmevichientong (2009) propose another adaptive ordering policy based on the sample-path subgradient estimate, and achieve a better rate

of convergence. Specifically, assume that $\bar{y}$ is a known upper bound for the unknown optimal inventory level y^*. For some $\gamma > 0$, their adaptive ordering policy is given by

$$y_{t+1} = \max\left\{\min\left\{y_t - \frac{\gamma\bar{y}}{\max\{h,\ p\}\sqrt{t}} \cdot H_t(y_t), \bar{y}\right\}, 0\right\}.$$

They show that in the perishable inventory case, the long-run average system cost $E[\Sigma_{t=1}^{T} L_t(y_t)/T]$ under the above adaptive ordering policy converges to the optimal cost L^* at a rate of $O(1\sqrt{T})$. For the nonperishable inventory case, with an additional assumption that there is a known positive lower bound for the unknown expected demand $\mathsf{E}[D_t]$, they show that the above ordering policy achieves the same rate of convergence when γ is sufficiently small under some mild technical conditions.

For a general unknown discrete demand distribution with perishable inventory, Huh et al. (2011) propose a data-driven policy based on the Kaplan–Meier (KM) estimator (Kaplan and Meier 1958), termed the "KM-myopic" policy. To apply their policy, one needs to make the following change in the definition of demand censoring: given an inventory level y_t, one can observe the event $\{D_t = y_t\}$ distinctly from the event $\{D_t > y_t\}$. In other words, this equates to a "partial censoring" setting in which one observes an additional lost-sales indicator of whether demand *strictly* exceeds the available inventory level or not. We note that for continuous demand distributions, the notion of the lost-sales indicator is not essential because the events $\{D_t > y_t\}$ and $\{D_t \geq y_t\}$ have the same probability measure. However, for discrete demand distributions, such a notion makes a significant difference in Bayesian updating (see also Huh and Rusmevichientong 2009, Sect. 3.4).

Under this new notion of demand censoring, we provide an illustration of the KM estimator and the corresponding KM-myopic policy below. Given n sorted observations, say, $\xi_1 \leq \xi_2 \leq \xi_3^c \leq \xi_4 \leq \cdots \leq \xi_n$, where the superscript c denotes censored observations such that $D_t > \xi_t$, the KM estimator works as follows. At first, allocate probability equally among n observations. Then, starting from the left, redistribute the probability of a censored observation among higher observations iteratively. For example, in this case, the smallest censored observation is ξ_3. Thus, in the first iteration, the $1/n$ probability originally assigned to ξ_3 is shared equally among $\xi_4, \ldots, \xi_n$, each of which will hence get an updated probability of $1/n + 1/n(n-3) = (n-2)/n(n-3)$. After we pass through the observations in this way, the resulting empirical distribution is given by

$$\overline{F}_n(\xi) = \prod_{i:\xi_i \leq \xi} \left(\frac{n-i}{n-i+1}\right)^{\delta_i},$$

where $\delta_i = 0$ if ξ_i is a censored observation, and $\delta_i = 1$ otherwise. The adaptive KM myopic policy can thus be constructed as follows:

$$y_{t+1} = \min\left\{y \geq 0 : \overline{F}_t(y) \leq \frac{p}{p+h}\right\}.$$

Huh et al. (2011) show that under the KM-myopic policy, y_t converges to the optimal inventory level y^* almost surely.

Besbes and Muharremoglu (2013) study the minimum worst-case regret for nonparametric models with perishable inventory, where they define regret as the difference between the expected cost of an adaptive policy and the full-information optimal cost L^*. They show that for a continuous demand distribution, the minimum worst-case regret under demand censoring grows logarithmically with the number of periods, as in the fully-observable demand case. On the other hand, when the demand distribution is discrete, they show that the minimum worst-case regret under demand censoring grows logarithmically with the number of periods, while regret can be bounded by a constant in the fully-observable demand case. Regret can also bounded by a constant under discrete demand in the "partially censored" setting. Thus, their finding highlights the importance of the availability of the lost-sales indicator in the existing literature of nonparametric models involving *discrete* demand distributions (e.g., Huh and Rusmevichientong 2009; Huh et al. 2011).

2.4 Open Research Areas

We have reviewed both Bayesian and nonparametric models for the demand censoring problem. Each type has its own strengths and limitations. For example, the Bayesian models entail an elegant Bayesian dynamic program formulation of the joint estimation and optimization problem. One can rely on these models to derive interesting structural results that shed light on the value of information and Bayesian learning. However, computing the optimal policy for the Bayesian models is nontrivial for relatively long time-horizon instances due to the curse of dimensionality. To overcome the dimensionality challenge, one typically has to resort to a fairly restrictive newsvendor distribution family that preserves the conjugate prior structure under demand censoring. This limits the applicability of the Bayesian models. The nonparametric models, on the other hand, work well for long time-horizon problems, and there is no need for any prior knowledge of the underlying demand distribution. As illustrated in our review, the adaptive ordering policies are often quite intuitive and easy to implement. The main challenge here, however, is to ensure the adaptive ordering policies converge quickly to the true optimal policy. Otherwise, the system performance in relatively short time horizons could be poor.

Despite the plethora of studies on demand censoring as reviewed above, there remain many open problems for future research. Below we discuss several of them.

1. *Demand Substitution:* Many retailers implicitly rely on demand substitution to mitigate the out-of-stock effect of a particular item at a particular store. There is an extensive literature on demand substitution, which is discussed in the chapters

concerning retail assortment planning in this volume. In our censored demand context, incorporating demand substitution among multiple products into the learning model could be of great practical value. Chen and Plambeck (2008) present a Bayesian model to jointly estimate the demand rate and the substitution probability. However, to keep their problem tractable, they make a simplifying assumption that the excess demand and the resulting substitution quantity are observable. It would be interesting to relax this assumption to investigate how demand censoring would affect the optimal inventory decisions under substitution. This is an open research problem that can be addressed by both the Bayesian and nonparametric approaches.

2. *Non-Stationary Demand:* Another practical consideration is non-stationary demand, which is common in many retail environments. Most of the censored demand models reviewed above assume the demand distribution is stationary. If the systematic variations in demand are deterministic (e.g., known seasonality), then one can simply normalize the demand observation by removing the deterministic variation components, so as to convert the problem to an equivalent stationary-demand one. However, if the systematic variations follow a random process, the problem becomes more complicated. Chen (2013a) shows that some of the results obtained under stationary demand can be extended to the Markov-modulated demand processes when the state transition probabilities are known. The case involving unknown transition probabilities is an open problem, as it is not clear how demand censoring would affect the learning of the unknown probabilities.
3. *Sales Transaction Timing Information:* One could further improve learning under censored demand by incorporating the timing of sales transactions. Jain et al. (2015) study such a Bayesian inventory control problem. They find that, when stockout timing information is available, the system performance improves significantly compared with the case without such information. Given that modern POS data include transaction timestamps, it would be interesting to further understand how timing information impacts some of the results reviewed here.
4. *Pricing Decisions:* One could also incorporate pricing decisions into the demand learning models. Burnetas and Smith (2000) propose an adaptive pricing and ordering policy for a price-dependent demand model with demand censoring. Bisi and Dada (2007) consider the joint pricing and ordering problem for price-dependent models in the Bayesian framework. Chen (2013a) studies a Bayesian dynamic pricing problem with an unknown customer willingness-to-pay distribution. In this case, if a customer buys a product, her willingness to pay must be greater than or equal to the posted price; if she does not buy the product, her willingness to pay must be below the posted price. Thus, the posted price serves as either a left- or right-censoring point of the customer's willingness to pay. Chen (2013a) proposes several approximation techniques to tackle this two-sided censoring problem. Applying the nonparametric approach to this two-sided censoring problem could be another interesting future research direction.

5. *Positive Replenishment Lead Times:* Both the Bayesian and nonparametric models in the literature assume zero lead time. Extending the existing models to the case of positive lead times would be an interesting and important contribution to the literature. However, we envision that such an extension could be technically challenging, because the lost-sales problem with a positive lead time is a known hard problem even when the demand distribution is known (see Zipkin 2000).

3 Models of Inventory Record Inaccuracy

There is ample evidence that the inventory available to customers on retail shelves is not correctly reflected in the retailers' computerized inventory records. In other words, retail managers have imperfect visibility into inventory in the store. DeHoratius and Raman (2008) examine the physical audit of a large, anonymous retail chain and observe that only 35 % of the retailer's inventory records match the physical inventory in the store. The extent of the problem is corroborated by other authors. Kang and Gershwin (2005) observe only 51 % record accuracy at a second anonymous retailer, and Gruen and Corsten (2008) find 32 % record accuracy at a third. We do not review in detail the literature on empirical measurement of inventory record inaccuracy; we refer the interested reader instead to the survey of DeHoratius and Ton (2015) in this volume.

Our focus instead is on potential analytical responses to the record inaccuracy phenomenon. Nearly all classical research on inventory management research assumes that the customer-available inventory level is known at every point in time, and landmark results in inventory theory rely on known inventory positions as a core (if not always explicit) assumption. A few analytical models of record inaccuracy date to the 1970s (e.g., Iglehart and Morey 1972), motivated by warehouse applications. There has been a surge of interest in inventory record inaccuracy in the past decade, particularly specialized to retail contexts, coinciding with new empirical studies and the rise of inventory tracking technologies—most prominently, item-level RFID tags which potentially offer real-time information on inventory locations and movements.

DeHoratius et al. (2008) outline three possible, non-exclusive responses of a retailer to inventory record inaccuracy: prevention, correction, and integration. Prevention refers to the elimination of root causes of inventory record inaccuracy, correction refers to inspection efforts, and integration refers to decision tools that account for the possible presence of inventory record inaccuracies. Our focus here is on "integrative" analytical approaches to inspection and replenishment, which we view as complementary to efforts towards prevention.

Analytical models are valuable for a few reasons. First, record inaccuracy is a significant feature of real inventory systems, and accounting for it has the potential to improve the matching of supply with demand and reduce inventory-related costs. Automated replenishment systems that assume accurate inventory records may not

live up to their billing when this assumption is violated. Second, modeling record inaccuracy helps measure the return on investment of inventory tracking technologies such as RFID. By comparing the inventory management cost of a "full-visibility" retailer equipped with an inventory tracking technology (an idealized model of which affords perfect inventory visibility) with the best-possible performance of an "intelligent" or "informed" retailer with distributional information about errors, one obtains a measure of the value of inventory visibility (Rekik et al. 2008; Kök and Shang 2007; Lee and Özer 2007). In addition, many papers also consider as a benchmark the performance of a "naive" or "ignorant" retailer who is oblivious to errors. Models of "intelligent" retailers are the focus of this review. A common theme in the literature on inventory record inaccuracy is that "intelligent" inventory models that account for record inaccuracy can recapture a significant fraction of the benefits of visibility without the substantial physical investment in tracking technologies.

The purpose of this section is to review the analytical literature on inventory record inaccuracy with an eye towards how analytical models can make best use of available information in the absence of inventory visibility afforded by tracking technologies or process improvement initiatives. We begin by presenting an example model of inventory record inaccuracy to illustrate some basic insights and challenges. We then discuss key modeling considerations before discussing relevant papers in more detail. We conclude the section with a discussion of open research directions.

3.1 A Basic Model

Consider a basic, single period inventory model in which a decision maker (DM) chooses an inventory quantity to stock in the face of uncertain demand. As a benchmark, assume a newsvendor setup in which the DM has full knowledge of an initial stock x. The DM places an order for y items at unit cost c and the items arrive immediately with no lead time. Random demand D then arrives according to probability distribution F, yielding sales $S = \min\{D, x + y\}$. A penalty cost of p per unit is charged for unsatisfied demand $D - S$, and leftover inventory $x + y - S$ is salvaged for $c_s - h$ per unit. If inventory records are perfect and the initial inventory x is known, we can write the problem as

$$\min_{y \geq 0} \; L(x, y) - c_s \mathsf{E}_D\left[(x + y - D)^+\right], \tag{5.1}$$

where

$$L(x, y) = cy + p\mathsf{E}_D\left[(D - x - y)^+\right] + h\mathsf{E}_D\left[(x + y - D)^+\right].$$

The solution is well-known to be of the critical fractile type:

$$y^* = \min\left\{y \geq 0 : F(x+y) \geq \frac{p-c}{p+h-c_s}\right\}. \tag{5.2}$$

Now suppose that inventory records are inaccurate, which we model by replacing the initial inventory position x by a random variable X with distribution P. We can write the new problem as

$$\min_{y \geq 0} \overline{L}(P,y) - c_s \mathsf{E}_{X,D}\left[(X+y-D)^+\right], \tag{5.3}$$

where

$$\overline{L}(P,y) = cy + p\mathsf{E}_{X,D}\left[(D-X-y)^+\right] + h\mathsf{E}_{X,D}\left[(X+y-D)^+\right].$$

The solution retains the critical fractile form (see Mersereau 2013),

$$\overline{y}^* = \min\left\{y \geq 0 : W(y) \geq \frac{p-c}{p+h-c_s}\right\}, \tag{5.4}$$

but the demand distribution F is replaced by a new distribution $W(y) = \Pr(D-X \leq y)$. The distribution W reflects demand less available inventory and can be computed as a convolution of F and P.

It is intuitive that in many realistic cases the solution to (5.4) should exceed that of (5.2) in order to make up for inventory lost in the error process (assuming that $\mathrm{E}[X] \leq x$) and to buffer the additional newsvendor uncertainty introduced by the distribution P (assuming the fractile $\frac{p-c}{p+h-c_s}$ is sufficiently large). Indeed, a number of authors (e.g., Kök and Shang 2007; DeHoratius et al. 2008; Atali et al. 2011) observe either analytically or numerically that record inaccuracy does indeed tend to increase stocking quantities under reasonable assumptions on demand and/or error distributions.[1] We revisit this "uncertainty effect" on replenishment in Sect. 3.3.1.

We note that this single-period model can also be viewed as a random yield model with additive yield uncertainty. See Yano and Lee (1995) for a detailed review of the literature on inventory management with random yield. In the random yield literature, errors are typically connected to incoming replenishments and are typically immediately observed by the DM. Therefore, inventory uncertainty does not persist or accumulate over time. With record inaccuracy, however, errors generally persist until the retailer performs an inspection. This is a significant

[1] The result is difficult to prove generally. Song (1994) includes a detailed analysis of the conditions required to rank newsvendor stocking quantities for different probability distributions, and these conditions are difficult to verify here.

challenge in moving from a single-period model to a multiperiod model of inventory record inaccuracy.

It is natural to formulate a multiperiod inventory problem as a Markov decision process (MDP). With perfect inventory records, we can formulate a T-period lost-sales version of problem (5.2) as a MDP with a one-dimensional state representing the current inventory position. Let x_{t-1} indicate the inventory position at the beginning of period t, let D_t denote random demand in period t (drawn from a potentially time-varying demand distribution F_t), and indicate by $V_t(\cdot)$ the cost-to-go from period t through the end of the horizon. The Bellman equation is as follows for $t = 1, \ldots, T$:

$$V_t(x_{t-1}) = \min_{y \geq 0} \{L(x_{t-1}, y)\} + \mathsf{E}_{D_t}[V_{t+1}(U(x_{t-1} + y - D_t))], \tag{5.5}$$

where $V_{T+1}(x_T) = -c_s x_T$ Here, $U(x) = (x)^+$ is an update function that specifies the inventory carried to the next period. With record inaccuracy, the inventory record is no longer a sufficient summary of the system state and the inventory optimization becomes a "partially observed" MDP (POMDP). Define $P_t(x) = \Pr(X_t \leq x | \mathcal{H}_t)$ as the probability distribution of the inventory random variable X_t conditional on the observed process history $\mathcal{H}_t$. We may consider P_t to be the system state of a modified dynamic programming formulation for $t = 1, \ldots, T$,

$$\begin{aligned} \overline{V}_t(P_{t-1}) = \min_{y \geq 0} & \left\{\overline{L}(P_{t-1}, y)\right\} \\ & + \mathsf{E}_{X_{t-1}, D_t}\left[\overline{V}_{t+1}\left(\overline{\mathcal{U}}_t(P_{t-1}, y, \min\{X_{t-1} + y, D_t\})\right)\right], \end{aligned} \tag{5.6}$$

where $\overline{V}_{T+1}(P_T) = -c_s \mathsf{E}[X_T]$. Here, the update operator $\overline{\mathcal{U}}_t$ transforms P_{t-1} to P_t given replenishment y and observed sales $S_t = \min\{X_{t-1} + y, D_t\}$. We do not express the $\overline{\mathcal{U}}_t$ operator here explicitly, but we note that it can be complicated, in general depending on probability distributions of both paying demand and unobserved errors. It must shift the inventory distribution up and down to reflect observed inventory inflows (replenishments) and observed outflows (sales). It must accumulate potential errors occurring in period t. Finally, as we discuss later, the update may also account for inferences the DM can make about customer-available inventory based on sales or other side observations. DeHoratius et al. (2008) and others derive $\overline{\mathcal{U}}_t$ using Bayes law. In other models (e.g., Kök and Shang 2007) the classical inventory record and the number of periods of error accumulation serve as sufficient statistics for P_{t-1}, in which case the update operator is simpler to express.

POMDPs are provably difficult to solve in general (Papadimitriou and Tsitsiklis 1987), suggesting that a problem like (5.6) is unlikely to be solvable without restrictions or approximations.

3.2 Modeling Considerations

While we have attempted in Sect. 3.1 to frame a fairly general model of replenishment under inventory record inaccuracy, this model already makes a number of strong assumptions, in particular about the decisions available to the DM, the modeling of errors, and the DM's observations of the system. These three dimensions represent key distinctions among papers in the literature, and we briefly discuss each one in turn.

1. **Decisions:** Section 3.1 formulates the problem of *replenishment* under inventory record inaccuracy, but inventory *inspection* (also referred to as counting or auditing) is another control available to a decision-maker operating with inventory record errors. Traditionally, retailers do periodic (often annual) inventory counts for accounting purposes. More frequent inspections, referred to as "cycle counts," may follow a fixed schedule based on an ABC-type categorization of stock-keeping units (SKUs), in which SKUs judged to be particularly at risk of inaccurate records, or of strategic or financial importance, are scheduled for cycle counts more frequently. An alternative to such static counting schedules are dynamic versions of cycle counts in which the retailer chooses which SKUs to inspect each day based on real-time information.

   Conceptually, it is straightforward to extend (5.6) to dynamically trigger inspections. We add a binary decision variable z_t each period which is an input to the update operator $\overline{\mathcal{U}}_t$. An inspection in period t resolves the uncertainty around X_t, which we model with an update that sets P_t to a distribution with all its weight at a single value (or to an appropriate probability distribution that represents an imperfect inspection).
2. **Error Process:** A key distinction among models of inventory record inaccuracy is the modeling of the error process. Most authors work in a periodic review setting and assume an error random variable (sometimes referred to as "invisible" or "non-paying" demand) that contributes to the discrepancy between available and recorded inventory each period. These discrepancies are not directly observed, and they accumulate over time between inventory inspections. A modeler of inventory record inaccuracy must make a number of decisions about the error process. Errors can be modeled as additive (e.g., DeHoratius et al. 2008) or multiplicative (e.g., Rekik et al. 2008) relative to the inventory level, and dependent on or independent of demand, replenishment, and/or inventory levels. Errors can be modeled as occurring before or after demand within a period, or interleaved with demand (e.g., Atali et al. 2011). Errors themselves may be directly costly in that they imply a physical loss or gain of saleable units (e.g., Kang and Gershwin 2005) or costless (e.g., Camdereli and Swaminathan 2010). Errors may be modeled as deterministic or associated with a probability distribution. Typical assumed probability distributions are one-sided (e.g., Huh et al. 2010), implying that customer-available inventory

is always less than or equal to recorded inventory, or symmetric around zero (e.g., Kök and Shang 2007).

In order to appreciate these modeling decisions, we can categorize the sources of inaccurate inventory levels, following Atali et al. (2011), into shrinkage (i.e., physical loss of inventory, typically through theft or damage), transaction errors (i.e., scanning, receipt, or counting errors that impact inventory records but not physical inventory), and misplacements (i.e., in which inventory is temporarily unavailable to the customer but still physically present in the store).[2] These different error sources suggest different assumptions about inventory dynamics and cost accrual. For example, it is common to model shrinkage using a one-sided error process inducing direct stock losses, and transaction errors using an additive, symmetric error process that incurs no direct cost.

Modeling all sources of errors in detail (as in Atali et al. 2011) is arguably truest to retail realities, given that all three types of errors are presumably present in retail settings. (DeHoratius and Raman 2008 report discrepancies of both signs in the audit data they analyze, with 58 % of errors such that physical inventory is less than recorded inventory.) However, such a model may be difficult to estimate from data, and it may require additional state variables for tracking the different types of error accumulations to allow for proper accounting of costs. Instead, most authors model a single error process that either reflects a single error source (shrinkage, transaction errors, or misplacement) or an aggregation of error sources.

Assuming that demand and errors occur interleaved within a period is also desirable but complicates modeling because of the different accounting of lost sales and "lost errors." Such a model must account for all possible sequences of demand and errors within a period. Instead, many authors model errors as occurring together, either before or after demand within a period.

Another common simplification is to assume errors arise from a stochastic process independent of demand and inventory levels. In many retail contexts, we would expect this not to be the case; for example, the same underlying factors leading to high or low demand would seem to also impact the volume of shrinkage, misplacement, and transaction errors. Because demand and inventory levels are not directly observed, modeling this dependency can bring complications that destroy problem structure. In some models, these complications take the form of an additional layer of conditioning in the update operator (e.g., DeHoratius et al. 2008). In others, the dynamic program state may need

[2] Here we depart slightly from DeHoratius and Ton (2009) in terminology. DeHoratius and Ton (2009) define "inventory record inaccuracy" as the difference between a store's recorded inventory position and the physical inventory in the store. Misplaced inventory, which is physically present in the store, does not contribute to inventory record inaccuracy in this definition. In our discussion, we will liberally use the term "inventory record inaccuracy" to refer to the difference between customer-available inventory and recorded inventory. That is, we consider misplaced inventory to be part of inventory record inaccuracy.

to include the history of sales observations, leading to a curse of dimensionality (e.g., Kök and Shang 2007).

3. **Observability:** A critical modeling choice is what the DM observes about sales and stockouts. In the lost sales model of Sect. 3.1, sales are the minimum of demand and customer-available inventory and are therefore statistically dependent on customer-available inventory. In such a model, the DM can in theory use sales observations to make inferences on available inventory; for example, if a DM observes a sequence of periods with zero sales, this may signal that there is no customer-available stock. Such inferences can be modeled using Bayes law. While potentially powerful, these inferences yield a complicated $\overline{U}_t$ operator in problem (5.6) (DeHoratius et al. 2008) that depends on sales observations and demand distributions.

   An alternative, which seems reasonable especially when stockouts are rare, is to ignore the signalling potential of sales observations. Such an assumption greatly simplifies the $\overline{U}_t$ operator, as errors accumulate independently of sales. In such cases, the inventory record and the number of periods since the last inspection typically serve as sufficient statistics for the multiperiod dynamic optimization (Kök and Shang 2007).

   A third possibility is to assume that customer-available inventory levels become observed whenever they reach zero (e.g., Bensoussan et al. 2007b). This can be practically motivated by assuming that customers who find an empty shelf request a "rain check" that is recorded, or by the practice of "zero-balance walks" in place at some retailers, in which employees periodically look for empty shelves in the store.

3.3 Review of Existing Literature

With this backdrop, we now review the operations management literature on store-level analytical models of inventory record inaccuracy. Given the challenges inherent in problems like (5.6), we believe that a fruitful way to categorize the existing literature on inventory record inaccuracy is by the modeling assumptions and analytical approximations employed to enable tractable analysis and computation. We put the literature into four categories: single-period models, classical multiperiod models, multiperiod models featuring low-dimensional sufficient statistics for P_t, and "partially observed" multiperiod models employing Bayesian updating.

3.3.1 Single Period Models

Single-period models of optimal stocking under inventory record inaccuracy yield some basic insights while avoiding some of the complexities inherent in multiperiod POMDP formulations like (5.6). For this reason, single-period models

are often employed as starting points upon which more complex models are built (e.g., Kök and Shang 2007; Huh et al. 2010; Mersereau 2013), or as stylized building blocks within more complex systems. For example, Heese (2007), Gaukler et al. (2007), Sahin and Dallery (2009) and Camdereli and Swaminathan (2010) employ single-period models to study the impact of inventory record inaccuracy on supply chain coordination. As our focus is on in-store operations, we do not review the supply chain aspects of these papers.

Rekik et al. (2008) analyze a single-period model that modifies the classical newsvendor problem by allowing for multiplicative misplacement errors to occur before paying demand arrives. A parameter θ is defined as the ratio between customer-available inventory and the inventory record, and is considered to be both deterministic as well as uniformly distributed on [0,1]. The authors explicitly look at the profit of naive, intelligent, and full-visibility retailers and conclude that the intelligent retailer achieves significant benefits over the naive retailer. The authors also examine stocking quantities: for the deterministic case stocking quantities first increase with θ (to make up for reduced yield) and then decrease with θ (to reduce misplaced inventory and associated overage charges).

Heese (2007) uses a multiplicative error model with uniformly distributed yields and makes similar observations about stocking quantities to Rekik et al. (2008). (His "centralized" model can be viewed as a single-location model.) Furthermore, even when setting the mean error ratio to one, Heese (2007) finds that the DM orders more than without inventory uncertainty for sufficiently high target service levels. We alluded to this "uncertainty effect" on stocking quantities in Sect. 3.1. Mersereau (2013) suggests an uncertainty effect in a model similar to (5.3). Mersereau (2013) also finds that optimal stocking levels can decrease if the DM anticipates physical errors to occur after stocking levels are chosen. This "direct loss" effect can be understood as reducing the stock available for theft or damage.

Single-period models therefore yield three insights into the effects of inventory record inaccuracy on optimal stocking levels: (1) optimal stocking levels may increase to make up for reduced yield, (2) they may also increase to buffer additional uncertainty brought by record inaccuracy; and (3) they may decrease in order to reduce the inventory available for misplacement or shrinkage.

3.3.2 Classical Multiperiod Models

A prevalent approach to modeling inventory record inaccuracy is to assume that inventory errors follow a pre-determined probability distribution that is independent of sales observations. As mentioned in Sect. 3.2, this greatly simplifies the update operator $\overline{\mathcal{U}}_t$ in (5.6), because the number of periods of error accumulation often serves as a sufficient statistic for the shape of P_t. Despite this simplification, optimal policies appear to be difficult to characterize in these systems except in specific cases.

An early stream of literature on inventory record inaccuracy, dating to Iglehart and Morey (1972), views error accumulation in the inventory system as a renewal process and seeks an auditing trigger that achieves a pre-specified probability of a "warehouse denial;" i.e., an event in which there is a physical stockout even though the inventory record appears sufficient to cover demand. This is an appropriate service metric in a warehouse context in which denials are observed by the firm and trigger a reconciliation of the inventory record with the physical inventory state.

Iglehart and Morey (1972) assume that errors are additive, stationary, mean zero random variables and that the DM maintains a buffer stock to account for them. Given a fixed buffer stock, the authors derive an asymptotic normal distribution approximation for the probability of cumulative errors exceeding the buffer stock. Their model decouples the classical safety and cycle stocks from the buffering of inventory inaccuracies, and the payoff is a joint inspection and replenishment policy expressed in closed form. The model of Rekik et al. (2009) is related in that the DM minimizes holding cost subject to a constraint on the probability of stockout during a finite horizon.

Morey (1985) uses a similar framework to Iglehart and Morey (1972) to establish "back-of-the-envelope" expressions for service levels as functions of error parameters, buffer stocks, and audit frequencies. Morey and Dittman (1986) generalizes Iglehart and Morey (1972) to determine audit frequencies in more general internal control settings, not necessarily inventory-related.

Kang and Gershwin (2005) present a detailed motivation for the problem of inventory record inaccuracy, including empirical evidence from an anonymous retailer. The paper's analysis is largely based on a numerical simulation of a (Q, R)-based stochastic inventory model with additive one-sided errors (called "stock loss" in the paper). One insight is that "freezing" of replenishment is possible; this occurs when the inventory record is above the reorder point yet there is no customer-available inventory on the shelf, in which case no sales occur and an automated replenishment system places no orders. The authors conclude that inventory inaccuracy may be especially costly in naive lean systems which carry little stock to buffer the additional uncertainty. This can be viewed as a corollary of the "uncertainty effect" discussed in Sect. 3.3.1. The paper goes on to numerically evaluate several remediation heuristics.

3.3.3 Multiperiod Models Featuring Sufficient Statistics

A number of papers analyzing multiperiod inventory optimization problems feature conditions or assumptions under which the multidimensional state P_t of a POMDP like (5.6) can be represented by a low-dimensional set of sufficient statistics. While such representations can incur a cost in terms of model generality, they have significant analytical and computational benefits.

Kök and Shang (2007) focus on joint replenishment and dynamic inspection triggering in a model in which errors are additive and have mean zero. They assume that both demand and errors are backlogged and that errors accumulate irrespective

of backlogs. As a result, the error process decouples from inventory levels, as in Iglehart and Morey (1972), and the accumulated discrepancy between physical and recorded inventory is the sum over periods of individual errors. That is, if an error ε_t occurs each period, the accumulated error after j periods of no inspections is $\overline{\varepsilon}_j \equiv \sum_{t=1}^{j} \varepsilon_t$. Its distribution is a j-fold convolution of the one-period error distribution. As a result, the authors are able to formulate the joint replenishment-inspection problem using a two-dimensional state (z_t, j_t), where z_t is the inventory record at time t (maintained by adding replenishments and subtracting observed sales each period) and j_t is the number of periods since the last audit.

Unfortunately, even with these simplifications the authors show that the multiperiod problem is non-convex. The authors suggest an "inspection adjusted base-stock" (IABS) policy that replenishes according to a j_t-dependent base-stock policy and inspects when the inventory record falls below a j_t-dependent cutoff. An IABS policy is optimal for the single-period problem, and an IABS policy seems to perform well as a heuristic for the multiperiod problem.

Atali et al. (2011) provide a detailed model of inventory errors, explicitly distinguishing among shrinkage, transaction errors, and misplacements in their model. Furthermore, they model demand and errors using a "random disaggregation" approach that splits an overall demand random variable into components for paying demand and various error sources. As a result, their model allows for demand and errors to be interleaved within a period. In solving their intelligent ("informed") retailer model, the authors approximate the distribution of total errors by a distribution that depends only on the inventory record and the number of periods since the last audit, as in Kök and Shang (2007). A state-dependent base-stock replenishment policy results from this approximation. A numerical study shows that the intelligent retailer achieves cost close to a full-visibility one and that detailed modeling of errors can achieve significant gains over aggregate error models for some parameter choices. A related model appears in Avrahami et al. (2012), who find through a numerical study that a "static" informed policy that knows only mean error information does nearly as well as an intelligent policy based on distributional error information.

Huh et al. (2010) show that a similar two-dimensional state to the one in Kök and Shang (2007) is sufficient for a particular model in which inventory inaccuracy is driven by additive shrinkage only, replenishments are only possible immediately after an inspection is made, and stockouts induce automatic inspections (akin to a "zero-balance walk"). In a given period, the DM knows that the true inventory level has only decreased since the last inspection (since errors only reduce physical inventory and since replenishments require inspections). If a stockout has not occurred, then the most recent post-inspection inventory level less recorded demand must exceed the accumulated errors (whose distribution is determined by the number of periods since the last inspection). The inventory distribution *conditional* on there being no stockout can therefore be computed given the inventory record, the number of periods since the last audit, and the error distribution. The authors present a rigorous dynamic programming formulation based on this result

and show that a threshold-based inspection policy, coupled with an order-up-to replenishment policy, is optimal for an infinite-horizon problem satisfying a number of technical assumptions.

3.3.4 Multiperiod Models Using Bayesian Updating

A set of authors studying inventory record inaccuracy has chosen to consider the partial observability of inventory levels more directly. These models require minimal assumptions on the inventory error distribution. In particular, the DM's belief P_t around inventory positions can be updated based on POS data. These models are more complex, however, in that the state space of the MDP is the space of possible distributions on X_t. Because of this complexity, optimal policies have only been computed for some simplified cases; otherwise, results are limited to heuristics and approximations.

DeHoratius et al. (2008) consider a multiperiod lost sales inventory system with discrete additive errors drawn from an arbitrary discrete distribution. The authors propose maintaining an explicit inventory belief P_t they call a "Bayesian inventory record" or "BIR." P_t is updated according to Bayes rule, using sales observations as signals of the underlying inventory levels. In particular, the Bayes update reflects that no sales may indicate a stocked out situation, and positive sales indicate that the inventory could not have been fewer than what was sold. The authors prove that such a solution avoids the problem of inventory "freezing" identified by Kang and Gershwin (2005).

DeHoratius et al. (2008) suggest a myopic replenishment policy and a BIR-based heuristic for dynamic triggering of inspections. The authors discuss the estimation of necessary parameters and report on a simulation study calibrated with retailer data that compares the performance of naive, intelligent ("Bayes"), and full-visibility ("Full") retailers. They demonstrate that the intelligent solution achieves a service-inventory tradeoff that captures a substantial portion of the benefits of the full-visibility solution.

DeHoratius et al. (2008) demonstrate that the updates can be performed efficiently in closed form when inventories and demands are discrete, but partial observability of inventory levels clearly adds analytical and computational complexity as discussed in Sect. 3.1. Mersereau (2013) analyzes in detail the problem of replenishment optimization for the model of DeHoratius et al. (2008), identifying both uncertainty and loss effects in a single-period version of the model. In a two-period version of the model, the author also identifies an "information effect:" stocking less can actually reduce the variance of the BIR and enhance information content for future periods. Mersereau (2013) proceeds to approximate the POMDP using an approach borrowed from the machine learning literature. A key finding is that an intelligent myopic policy is near-optimal in numerical trials.

Bensoussan et al. (2011b) formulate a related model to DeHoratius et al. (2008) in that excess demand is unobservable. Errors are one-sided, and demand and inventory are permitted to be continuous. Continuous inventory and demand

complicates the updating process; the resulting inventory belief is a mixture of continuous and discrete distributions. The authors prove the existence and uniqueness of an optimal policy, present a lower-bounding approach, and propose an iterative approximation algorithm.

A separate series of papers considers similar "partially observed" inventory systems with continuous inventory levels where the DM only observes whether or not the physical inventory level is strictly positive. In particular, sales are not observed. Errors are not explicitly modeled but can be assumed to be a component of the (unobserved) demand process. Bensoussan et al. (2007b) considers such a model with lost sales. As in DeHoratius et al. (2008), the state of the system is represented by a distribution around the customer-available inventory level that is updated in a Bayesian fashion. The resulting replenishment problem is therefore defined on a functional state space, and the authors focus on finding conditions for an optimal solution to exist and to be unique. Bensoussan et al. (2008) perform related analyses for a variation of the Bensoussan et al. (2007b) model in which backorders (i.e., "rain checks") are permitted and the DM only observes the inventory level when it is negative. Bensoussan et al. (2011a) use a value function approximation to approximate the problem of Bensoussan et al. (2008). In a numerical study, they observe both an uncertainty and an information effect with interpretations related to those in Mersereau (2013).

Finally, Chen (2013b) considers the problem of dynamic cycle count triggering using a simplified POMDP in which the system can switch from a "normal" state in which the inventory level is known to a "faulty" state in which the system is stocked out. This results in a partial decomposition of the replenishment and inspection decisions. The inspection policy is an easily computed threshold policy based on the number of consecutive zero-sales periods, and the optimal replenishment is a base-stock policy with base-stock levels depending on the time since the last positive sale. The author finds a loss effect; the error process drives the retailer to stock less to limit the inventory made unavailable by errors. Chuang and Oliva (2013) also use a two-state model of record accuracy to determine the inspection frequency in a fixed inspection policy.

3.4 Open Research Areas

Despite numerous and varied analytical approaches to modeling retail inventory inaccuracy in recent years, there remain a number of open opportunities for future research.

1. *Multi-SKU and Multi-Location Models:* As with much of classical inventory theory, single-SKU models dominate the analytical literature on inventory record inaccuracy. Kök and Shang (2014) consider coordinated inspection policies in a serial supply chain. We are aware of little research, however, on models that use data across stores or SKUs. Consider the following inspection

trigger policy: inspect a SKU at a store when its recent sales fall significantly below sales for the same SKU at neighboring stores. It is intuitive that similar SKUs and stores, used in this way, could serve as useful benchmarks for detecting deviations from normal operations. Substitution is also potentially relevant to include in models of inventory record inaccuracy. For example, a retailer might suspect that a SKU has too little customer-available inventory after detecting increased sales of substitute SKUs. Extending models like (5.6) to multiple SKUs adds considerable complexity to the update operator and dimensionality to the state space, however.

2. *Estimation of Model Parameters:* Despite a fairly rich body of empirical research into the presence of record inaccuracy, there remain a number of open questions surrounding the estimation of the daily or weekly error processes assumed by most analytical models. DeHoratius et al. (2008) present a basic estimation approach, and Chuang and Oliva (2013) provide a structural approach for estimating error incidence at the SKU level. Nevertheless, we believe that detailed estimation of error processes remains an unresolved issue. As a result, the existing papers make use of a wide range of assumptions on error distributions. Furthermore, estimation of other model parameters may be confounded by record inaccuracy. Mersereau (2015) shows that the presence of inventory record inaccuracy can introduce biases into the estimation of paying demand.
3. *Analytical and Computational Tractability:* Efficient solutions, much less complete characterizations, of problems like (5.6) have proved elusive without approximations or restrictive assumptions. There is apparent in the existing literature a tradeoff between model realism and tractability, with no clear dominant approach. This leaves room for continued analytical and algorithmic work on both optimal solutions and useful approximations and heuristics.
4. *Comparison of Models and Prescriptions:* Despite the large number of competing models of inventory inaccuracy and solutions for replenishment and inspection, we are not aware of any efforts to compare them. One advantage of Bayesian models like DeHoratius et al. (2008) and Chen (2013b) is that they make use of sales information as signals about inventory levels. It is intuitive that this information should be most useful when stockouts are relatively common. It would be interesting to examine under what conditions a POMDP-based model like DeHoratius et al. (2008) outperforms a sufficient statistic model like Kök and Shang (2007), and vice versa.
5. *Pilot Testing of Policies:* Given the eminent practicality of inventory models integrating inventory inaccuracy, implementations of responses to inventory record inaccuracy would be especially interesting. Such reports have started to emerge. Chuang et al. (2012) report on a field experiment in which a data-driven heuristic was used to trigger inspections. Hardgrave et al. (2013) report on two controlled field experiments measuring the reduction in record inaccuracy enabled by real RFID implementations. Both papers suggest that the potential improvements to retail operations can be substantial.

4 Visibility Technologies and Research Opportunities

Both the literatures on demand censoring and inventory record inaccuracy formulate and solve problems of decision-making under uncertainty, and it is therefore not surprising that these literatures pull from a common set of methodologies including statistical decision theory, stochastic (and partially observed) dynamic programming, and Bayesian and nonparametric inference. We have proposed several specific research directions related to demand censoring and inventory record inaccuracy in Sects. 2 and 3, respectively. We add that demand censoring and inventory record inaccuracy tend to occur simultaneously in many retail stores, and their interaction leads to additional challenges. For example, when records are inaccurate, the retailer no longer receives a reliable indicator of when stockouts occur. Mersereau (2015) is the one paper we are aware of that considers both features together. One unique insight is that if demand censoring is accounted for but inventory record inaccuracy is not, then the retailer will tend to underestimate demand over time. We believe that there is room for further examination of this interaction as well as other interactions involving multiple sources of uncertainty, even though considering multiple uncertainties together brings obvious modeling complications.

We conclude the chapter by looking to other interesting directions for future research on in-store visibility that extend beyond demand censoring and inventory record inaccuracy. We believe that exciting research opportunities abound if we consider other types of information made available by new in-store visibility technologies. Below we discuss some of the modern and emerging technologies developed for the retail industry, categorized by the three main components of the store as illustrated in Fig. 5.1.

1. *Inventory Information.* We introduced RFID in Sect. 3. As the price of RFID tags decreases, attaching RFID tags to individual items (as opposed to cases or pallets) becomes increasingly feasible. The application of RFID technology has received strong interest among individual retailers, technology providers (e.g., Tyco Retail Solutions), trade journals (e.g., RFID Journal), and academics (e.g., the University of Arkansas Walton College's RFID Research Center). Waller et al. (2011) list a full 60 uses of RFID in apparel retail supply chains. Fisher and Raman (2010), who use RFID as a case study to illustrate the opportunities and risks inherent in new retail technology, call RFID "revolutionary." Beyond RFID, new crowdsourcing platforms such as Quri and Gigwalk enlist shoppers to report the status of inventory levels and displays via smartphone, offering retailers a true customer view of their store operations. Interestingly, these technologies also appear to be used by brand managers to monitor retailers' execution and adherence to the brand's promotion plans.
2. *Customer Flow Information.* Traffic counters—sensors that measure traffic in retail stores (e.g., ShopperTrak)—have become common in retail. Knowing how many potential customers are in the store at a time enables retailers to estimate conversion from traffic to sales and to match staffing with customer traffic.

Technologies are increasingly able to track customer movements within the store; for example, by detecting "pings" of customer cell phones (e.g., Euclid Analytics), by attaching RFID tags and mobile devices to shopping carts (e.g., MediaCart!), and by "seeing" customer bodies using infrared technology (e.g., Irisys). Video footage is increasingly analyzed by software to detect and record customer locations and customer engagement (e.g., SCOPIX Solutions, Envysion, RetailNext). By identifying highly trafficked areas of the store these technologies can assist with store layout decisions, and by measuring queue lengths and wait times they can inform queue management. Mobile devices also offer the opportunity for retailers to address individual customers as they shop with store maps, inventory information, and promotions (e.g., Apple's iBeacon).

3. *Store Associates Task Information.* Store associates increasingly carry mobile devices (i.e., smartphones and tablets) to communicate with each other, to give them real-time access to product, sales, and inventory information and to enable them to perform checkout, inspection, and replenishment functions (e.g., Motorola Retail 2008). Such devices offer the possibility of enhanced visibility to associates on the store floor in addition to management.

One possibility is that some of the estimation and inference problems reviewed in Sects. 2 and 3 may become less important as these visibility technologies become more reliable and inexpensive and retailers learn to make use of the information they provide. Nevertheless, we believe that new data sources will also inspire new research problems, and that visibility technologies and analytical methodologies may complement each other in many cases. For example, perhaps a retailer's response to demand censoring can be enhanced by using customer traffic data to make inferences about lost sales in the event of a stockout. Perhaps models of inventory record inaccuracy can be improved using information from an RFID reader that detects whether items are in the front- or backroom of a store. Ultimately, analytical methodologies form the link between new visibility technologies and better decisions. Below we suggest two broad categories of new analytical research opportunities in store operations that could complement the new visibility technologies.

New Insights from Combining Data Sources While it is common to simplify analytical operations management models by assuming a single location, SKU, or customer segment, we believe that there may be significant gains from leveraging data across stores and SKUs to impute missing in-store data. For example, as discussed in Sect. 2.4, sales data from multiple SKUs can be used to estimate substitution probabilities and to determine the optimal stocking policy for multiple SKUs. Another example was suggested in Sect. 3.4: data from other stores and SKUs may be used as benchmarks against which deviations can be detected for the purpose of process control. Given the large number of emerging visibility technologies listed above, there may also be significant value to considering multiple visibility technologies together; for example, recall from Sect. 3 that POS data can be used to make inferences on uncertain inventory levels. By modeling the interactions between different processes in a store, we believe that both better

empirics and improved analytical decisions may be possible. To give two recent examples from the literature, Perdikaki et al. (2012) and Mani et al. (2015) use traffic counting and conversion data to measure the impact of labor staffing on sales performance, with clear implications on labor planning. Lu et al. (2013) use video data to measure queue lengths and thereby quantify the impact of queue lengths on customer purchase behavior, with clear implications on queue design and staff scheduling.

New Parameters and New Decisions As in any operational context, the parameters of an analytical model must be estimated before a model can be used for decision-making. Retail environments are especially complex and non-stationary, heightening the need for estimation. Though we have not attempted to review it in this chapter, there is a growing empirical literature gaining ever finer insights into retail operations from richer datasets using more sophisticated methodologies. The rise of new visibility technologies expands the set of operational parameters that can conceivably be estimated. As an example, customer tracking technologies, by identifying more and less trafficked locations in the store, potentially allow for more detailed, location-specific assortment planning. Furthermore, new technologies offer retail managers new levers in the store. To give just one example, new digital price tags (e.g., Altierre Corp.) and customized mobile phone offers (e.g., Retailigence's adPop) allow for dynamic pricing that can potentially depend on real-time traffic and inventory states.

In conclusion, we believe that the study of visibility in retail stores exemplifies the trend towards business analytics more generally. Inventory management with censored demand observations and record inaccuracy represent just two examples of what is possible. The interplay between information, technology, inventory optimization, customer behavior, and human resources suggest a range of fresh analytical questions that have the potential to make a real impact on practice. Our hope is that our surveys and discussion here encourage further research on these topics.

Acknowledgements The authors thank the editors and an anonymous reviewer for constructive comments that greatly improved the chapter. They thank Jan Davis, Nicole DeHoratius, Saravanan Kesavan, Marcelo Olivares, Ariel Schilkrut, and participants in the 2013 Consortium for Operational Excellence in Retailing annual conference and in the 2013 UNC Retail Conference for valuable discussions and input.

References

Agrawal, N., & Smith, S. (1996). Estimating negative binomial demand for retail inventory management with unobservable lost sales. *Naval Research Logistics, 43*, 839–861.

Anupindi, R., Dada, M., & Gupta, S. (1998). Estimation of consumer demand with stock-out based substitution: An application to vending machine products. *Marketing Science, 17*(4), 406–423.

Atali, A., Lee, H., & Özer, Ö. (2011). *If the inventory manager knew: Value of visibility and RFID under imperfect inventory information*. Stanford University, Working Paper.

Avrahami, A., Tzimerman, A., Herer, Y. T., & Shtub, A. (2012). *The value of inventory accuracy in supply chain management*. Technion: Israel Institute of Technology, Working Paper.

Azoury, K. S. (1985). Bayes solution to dynamic inventory models under unknown demand distribution. *Management Science, 31*(9), 1150–1160.

Bensoussan, A., Cakanildirim, M., Sethi, S. P., & Shi, R. (2011a). Calculation of approximate optimal policies in a partially observed inventory model with rain checks. *Automatica, 47*, 1589–1604.

Bensoussan, A., Cakanyildirim, M., Li, M., & Sethi, S. P. (2011b). *Inventory control with a cash register: Sales recorded but not demand or shrinkage*. University of Texas at Dallas, Working Paper.

Bensoussan, A., Cakanyildirim, M., Minjárez-Sosa, J. A., Sethi, S. P., & Shi, R. (2008). Partially observed inventory systems: The case of rain checks. *SIAM Journal on Control and Optimization, 47*(5), 2490–2519.

Bensoussan, A., Cakanyildirim, M., & Sethi, S. (2005). On the optimal control of partially observed inventory systems. *Comptes Rendus Mathematique, 341*, 419–426.

Bensoussan, A., Cakanyildirim, M., & Sethi, S. (2007a). A multi-period newsvendor problem with partially observed demand. *Mathematics of Operations Research, 32*(2), 322–344.

Bensoussan, A., Cakanyildirim, M., & Sethi, S. (2007b). Partially observed inventory systems: The case of zero balance walk. *SIAM Journal on Control and Optimization, 46*(1), 176–209.

Bensoussan, A., Cakanyildirim, M., & Sethi, S. (2009). A note on 'the censored newsvendor and the optimal acquisition of information'. *Operations Research, 57*(3), 791–794.

Besbes, O., & Muharremoglu, A. (2013). On implications of demand censoring in the newsvendor problem. *Management Science, 59*(6), 1407–1424.

Bisi, A., & Dada, M. (2007). Dynamic learning, pricing, and ordering by a censored newsvendor. *Naval Research Logistics, 54*, 448–461.

Bisi, A., Dada, M., & Tokdar, S. (2011). A censored-data multiperiod inventory problem with newsvendor demand distributions. *Manufacturing Service Operations Management, 37*(11), 1390–1405.

Braden, D. J., & Freimer, M. (1991). Informational dynamics of censored observations. *Management Science, 37*(11), 1390–1405.

Brynjolfsson, E., Hu, Y. J., & Rahman, M. S. (2013). Competing in the age of omnichannel retailing. *MIT Sloan Management Review, 54*(4), 23–29.

Burnetas, A., & Smith, C. (2000). Adaptive ordering and pricing for perishable products. *Operations Research, 48*(3), 436–443.

Camdereli, A. Z., & Swaminathan, J. M. (2010). Misplaced inventory and radio-frequency identification (RFID) technology: Information and coordination. *Production Operations Management, 19*(1), 1–18.

Chen, L. (2010). Bounds and heuristics for optimal Bayesian inventory control with unobserved lost sales. *Operations Research, 58*(2), 396–413.

Chen, L. (2013a). *Bayesian dynamic pricing with two-sided censored customer willingness-to-pay data*. Duke University, Working Paper.

Chen, L. (2013b). *Fixing phantom stockouts: Optimal data-driven shelf inspection policies*. Duke University, Working Paper.

Chen, L., & Plambeck, E. L. (2008). Dynamic inventory management with learning about the demand distribution and substitution probability. *Manufacturing Service Operations Management, 10*(2), 236–256.

Chuang, H., & Oliva, R. (2013). *Empirical modeling of inventory record audit policies*. Texas A&M University, Working Paper.

Chuang, H., Oliva, R., & Liu, S. (2012). *On-shelf availability, retail performance, and external audits: A field experiment*. Texas A&M University, Working Paper.

Conlon, C. T., & Mortimer, J. H. (2013). Demand estimation under incomplete product availability. *American Economic Journal: Microeconomics, 5*(4), 1–30.

Conrad, S. A. (1976). Sales data and the estimation of demand. *Operational Research Quarterly, 27*(1), 123–127.

DeGroot, M. (1970). *Optimal statistical decisions*. New York: McGraw-Hill.

DeHoratius, N., Mersereau, A., & Schrage, L. (2008). Retail inventory management when records are inaccurate. *Manufacturing Service Operations Management, 10*(2), 257–277.

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

DeHoratius, N., & Ton, Z. (2015). The role of execution in managing product availability. In N. DeHoratius & Z. Ton (Eds.), *Retail supply chain management. Quantitative models and empirical studies*. New York: Springer.

Ding, X., Puterman, M., & Bisi, A. (2002). The censored newsvendor and the optimal acquisition of information. *Operations Research, 50*(3), 517–527.

Fisher, M., & Raman, A. (2010). *The new science of retailing*. Boston: Harvard Business Press.

Gaukler, G. M., Seifert, R. W., & Hausman, W. H. (2007). Item-level RFID in the retail supply chain. *Production and Operations Management, 16*(1), 65–76.

Godfrey, G., & Powell, W. (2001). An adaptive, distribution-free algorithm for the newsvendor problem with censored demands, with applications to inventory and distribution. *Management Science, 47*(8), 1101–1112.

Gruen, T. W., & Corsten, D. (2008). *A comprehensive guide to retail out-of-stock reduction in the fast-moving consumer goods industry*. Arlington (VA): Food Marketing Institute.

Hardgrave, B. C., Aloysius, J. A., & Goyal, S. (2013). RFID-enabled visibility and retail inventory record inaccuracy: Experiments in the field. *Production and Operations Management, 22*(4), 843–856.

Harpaz, G., Lee, W., & Winkler, R. (1982). Learning, experimentation, and the optimal output decisions of a competitive firm. *Management Science, 28*, 589–603.

Heese, H. S. (2007). Inventory record inaccuracy, double marginalization, and RFID adoption. *Production Operations Management, 16*(5), 542–553.

Heyman, D., & Sobel, M. (1984). *Stochastic models in operations research, volume II: Stochastic optimization*. New York: McGraw-Hill.

Huh, W. T., Levi, R., Rusmevichientong, P., & Orlin, J. B. (2011). Adaptive data-driven inventory control with censored demand based on Kaplan-Meier estimator. *Operations Research, 59*(4), 929–941.

Huh, W. T., Olvera-Cravioto, M., & Özer, Ö. (2010). *Joint audit and replenishment decisions for an inventory system with unrecorded demands*. Columbia University, Working Paper.

Huh, W. T., & Rusmevichientong, P. (2009). A non-parametric asymptotic analysis of inventory planning with censored demand. *Mathematics of Operations Research, 34*(1), 103–123.

Iglehart, D. L., & Morey, R. C. (1972). Inventory systems with imperfect asset information. *Management Science, 18*(8), B338–B394.

Jain, A., Rudi, N., & Wang, T. (2014). Demand estimation and ordering under censoring: Stockout timing is (almost) all you need. *Operations Research* (forthcoming).

Kang, Y., & Gershwin, S. (2005). Information inaccuracy in inventory systems: Stock loss and stockout. *IIE Transactions, 37*, 843–859.

Kaplan, E., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*(282), 457–481.

Kök, A. G., & Shang, K. (2007). Inspection and replenishment policies for systems with inventory record inaccuracy. *Manufacturing Service Operations Management, 9*(2), 185–205.

Kök, A. G., & Shang, K. (2014). Evaluation of cycle-count policies for supply chains with inventory inaccuracy and implications on RFID investments. *European Journal of Operational Research, 237*(1), 91–105 (Note: This reference was formerly titled "Evaluation of supply chains with inventory record inaccuracy and implications on RFID investments").

Lariviere, M., & Porteus, E. (1999). Stalking information: Bayesian inventory management with unobserved lost sales. *Management Science, 45*(3), 346–363.

Lee, H., & Özer, Ö. (2007). Unlocking the value of RFID. *Production and Operations Management, 16*(1), 40–64.
Lovejoy, W. (1990). Myopic policies for some inventory models with uncertain demand distributions. *Management Science, 36*(6), 724–738.
Lu, X., Song, J.-S., & Zhu, K. (2005). On 'the censored newsvendor and the optimal acquisition of information'. *Operations Research, 53*(6), 1024–1027.
Lu, X., Song, J.-S., & Zhu, K. (2007). *Inventory control with unobservable lost sales and Bayesian updates*. Duke University, Working Paper.
Lu, X., Song, J.-S., & Zhu, K. (2008). Analysis of perishable-inventory systems with censored demand data. *Operations Research, 56*(4), 1034–1038.
Lu, Y., Musalem, A., Olivares, M., & Schilkrut, A. (2013). Measuring the effect of queues on customer purchases. *Management Science, 59*(8), 1743–1763.
Mani, V., Kesavan, S., & Swaminathan, J. M. (2014). Estimating the impact of understaffing on sales and profitability in retail stores. *Production and Operations Management* (forthcoming).
McKinsey Global Institute. (2011). Big data: The next frontier for innovation, competition and productivity. Accessed April 9, 2013, from www.mckinsey.com/mgi
Mersereau, A. J. (2013). Information-sensitive replenishment when inventory records are inaccurate. *Production and Operations Management, 22*(4), 792–810.
Mersereau, A. J. (2014). Demand estimation from censored observations with inventory record inaccuracy. *Manufacturing and Service Operations Management* (forthcoming).
Morey, R. C. (1985). Estimating service level impacts from changes in cycle count, buffer stock, or corrective action. *Journal of Operations Management, 5*(4), 411–418.
Morey, R. C., & Dittman, D. A. (1986). Optimal timing of account audits in internal control. *Management Science, 32*(3), 272–282.
Motorola Retail. (2008). *The next revolution in retail technology*. Accessed July 23, 2013, from www.motorolasolutions.com
Musalem, A., Olivares, M., Bradlow, E. T., Terwiesch, C., & Corsten, D. (2010). Structural estimation of the effect of out-of-stocks. *Management Science, 56*(7), 1180–1197.
Nahmias, S. (1994). Demand estimation in lost sales inventory systems. *Naval Research Logistics, 41*, 739–757.
Papadimitriou, C., & Tsitsiklis, J. N. (1987). The complexity of Markov decision processes. *Mathematics of Operations Research, 12*, 441–450.
Perdikaki, O., Kesavan, S., & Swaminathan, J. M. (2012). Effect of traffic on sales and conversion rates of retail stores. *Manufacturing Service Operations Management, 14*(1), 145–162.
Powell, W., Ruszczynski, A., & Topaloglu, H. (2004). Learning algorithms for separable approximations of discrete stochastic optimization problems. *Mathematics of Operations Research, 29*(4), 814–836.
Rekik, Y., Sahin, E., & Dallery, Y. (2008). Analysis of the impact of RFID technology on reducing product misplacement errors at retail stores. *International Journal of Production Economics, 112*(1), 264–278.
Rekik, Y., Sahin, E., & Dallery, Y. (2009). Inventory inaccuracy in retail stores due to theft: An analysis of the benefits of RFID. *International Journal of Production Economics, 118*, 189–198.
Sahin, E., & Dallery, Y. (2009). Assessing the impact of inventory inaccuracies within a newsvendor framework. *European Journal of Operational Research, 197*, 1108–1118.
Scarf, H. (1959). Bayes solution of the statistical inventory problem. *Annals of Mathematical Statistics, 30*, 490–508.
Scarf, H. (1960). Some remarks on bayes solutions to the inventory problem. *Naval Research Logistics, 7*, 591–596.
Silver, E. (1993). Bayesian updating of an arbitrary discrete distribution under a special case of partial information. *Communications in Statistics Stochastic Models, 9*(4), 615–638.
Song, J.-S. (1994). Leadtime uncertainty in a simple stochastic inventory model. *Management Science, 40*(5), 603–613.

Vulcano, G., van Ryzin, G., & Ratliff, R. (2012). Estimating primary demand for substitutable products from sales transaction data. *Operations Research, 60*(2), 313–334.

Waller, M. A., Cromhout, D. B., Patton, J. B., Williams, B. D., & Hardgrave, B. C. (2011). *An empirical study of potential uses of RFID in the apparel retail supply chain*. University of Arkansas Information Technology Research Institute, Working Paper ITRI-WP156-0111.

Wecker, W. E. (1978). Predicting demand from sales data in the presence of stockouts. *Management Science, 24*(10), 1043–1054.

Yano, C. A., & Lee, H. L. (1995). Lot sizing with random yields: A review. *Operations Research, 43*(2), 311–334.

Zipkin, P. (2000). *Foundations of inventory management*. Boston: McGraw-Hill.

Chapter 6
An Overview of Industry Practice and Empirical Research in Retail Workforce Management

Saravanan Kesavan and Vidya Mani

1 Introduction

In the highly competitive retail environment, many retailers consider in-store experience critical to converting incoming traffic into sales and future visits. Superior in-store experience requires having not only inventory in place but also a skilled store workforce to ensure an efficient and pleasant visit for the customers. Numerous studies in marketing have shown that store associates play a critical role in driving customer satisfaction (Parasuraman et al. 1988; Zeithaml et al. 1996). Anecdotal evidence of financial distress resulting from mismanagement by retailers of their labor force is abundant. One recent example involves Circuit City, a consumer electronics company, which undertook several drastic changes under its new management, including revamping its store labor by letting-go of its highest paid sales associates. Retail observers claim that firing such experienced sales associates caused customer satisfaction to decline precipitously and contributed to Circuit City's subsequent bankruptcy (Mui 2007).

While retailers care deeply about providing high service levels to customers through increased labor in their stores, they are also mindful about the expenses associated with this practice. Payroll-expenses are about 10 % of sales in the retail industry and can often be the largest component of a store's variable costs. Kesavan et al. (2013) study a big-box retailer whose labor expenses account for 85 % of total

S. Kesavan (✉)
Kenan-Flagler Business School, University of North Carolina at Chapel Hill, Campus Box 3490, McColl Building, Chapel Hill, NC 27599-3490, USA
e-mail: skesavan@unc.edu

V. Mani
Supply Chain and Information Systems, Penn State Smeal College of Business, 461 Business Building, University Park, PA 16802, USA
e-mail: vmani@psu.edu

N. Agrawal, S.A. Smith (eds.), *Retail Supply Chain Management*, International Series in Operations Research & Management Science 223,
DOI 10.1007/978-1-4899-7562-1_6

controllable expenses[1] in the store. As a consequence, retailers need to balance the need to drive sales by using more labor against the need to control expenses that can increase commensurately. This task is challenging and requires careful workforce management. In this chapter, we review the literature on workforce management; provide a detailed overview of labor planning practice at one retailer; review new and upcoming technologies in the retail landscape that can potentially impact labor practices; and conclude with areas of future research.

The raison d'être for store labor is fairly similar across most retailer settings. First, stores need sufficient labor to ensure customer service. Service entails dealing directly with the customers during the purchase process: answering customer questions about the product and any services or warranty associated with them, and indirectly affecting their in-store experience by ensuring a neat and clean store. Second, store labor needs to manage the inventory in the store. Managing inventory involves receiving merchandise from delivery trucks while ensuring that it complies with the bill-of-materials, stocking the shelves to ensure that customers can find the products they are looking for, and finally, keeping the price current so it reflects the discounts or pricing changes that the corporate office may mandate. Third, store labor is required to maintain the signage within a store. Corporate office announcements of a new promotional event require that store labor update store signage to be consistent with the marketing activity. Finally, store labor is required for cashiering.

Broadly, retail labor falls into three categories depending upon the employment contract with the retailer: full-time workers, part-time workers, and temporary or seasonal workers. According to the Bureau of Labor Statistics (BLS), only 70 % of the estimated 15 million strong retail workforce in 2013 is full-time. Further, the retail industry added more than 700,000 seasonal employees for the holiday season in 2013 (BLS). Full-time workers are year-round employees who are typically employed for fixed hours per week, typically 35–40 h. They can be employed for a few more hours with overtime pay. Part-time employees are also year-round employees but face variable hours of employment in a week. BLS defines part-time workers as those who usually work less than 35 h per week. For example, the retail organization studied in Kesavan et al. (2013) guaranteed 10 h of employment per week for its part-time employees and deployed them for an average of 22 h per week. While retailers may increase the hours of the part-time employees to 40 h per week, they can do so only for a short-period of time before these workers get reclassified as full-time employees. Finally, temporary employees, sometimes called seasonal employees, are deployed for shorter-periods of time to manage seasonality or short-term demand fluctuation. Seasonal workers can be a large proportion of the total workforce for retailers during the peak period. For example, Home Depot planned to hire 70,000 seasonal employees to augment its 320,000 regular employees to meet seasonal demand in Spring 2012. Seasonal employees

[1] This retailer had identified the controllable component of each of the expenses based on historical data for each store.

have not only varying lengths of employment but also can be deployed for varying hours in a week. Typically, full-time employees are provided with other benefits, such as sick pay, vacation pay, and health care benefits. Some retailers tend to provide benefits to part-time employees but temporary employees rarely receive such benefits.

These different classes of workers offer various advantages to retailers to manage their stores. Typically, full-time workers are considered to have the highest capability amongst the three classes of workers since full-time employment often draws the most qualified candidates.[2] Further, literature on learning curve effects have shown that performance improves with cumulative experience (Lapré and Nembhard 2011) so full-time workers who spend more time in their jobs compared to part-time and seasonal workers are likely to have greater capabilities. Finally, full-time employees' incentives may be better aligned with that of the organization compared to those of part-time and seasonal workers. So, apart from playing an important role in driving sales through superior customer service, they may also reduce organizational costs by having lower turnover compared to part-time and seasonal workers. Annual turnover for the retail sector can be as high as 100 % (National Retail Foundation) but the break-down for part-time and seasonal workers is not available.

Part-time and seasonal workers, on the other hand, provide other important benefits to retailers. The wage rates and other benefits tend to be lower than that of full-time workers. In addition, they provide volume flexibility (upside flexibility and temporal flexibility) (Kesavan et al. 2013) to retailers that could enable them to manage demand less expensively, at least up to a certain point.

Labor planning involves determining the right number of full-time, part-time, and seasonal workers in the stores and allocating the forecasted hours across those workers. We observe considerable differences in the way labor planning is performed in the retail industry. One important dimension in which retail organizations can vary is the level of sophistication used to manage payroll. At one end of the spectrum, payroll decisions are completely driven by store managers without the support of decision-making tools. This practice is typical of smaller retailers, but we have observed that even retailers with annual revenues exceeding a billion dollars may follow such an ad hoc process. At the other end of the spectrum, several retailers have invested millions of dollars in workforce management tools that plan how much labor each store must carry. Some examples of firms developing workforce management tools are RedPrairie, Kronos, Reflexis, and Ceridian.

Another area of difference is the degree to which different departments within a retail organization are involved in labor planning. Several departments within retail organizations commonly want a say in the amount of labor in the store. Sometimes these departments have different goals. For example, the finance department in a

[2] There are exceptions to this generalization. For example, it is common to witness well qualified plumber or a sales associate who pursues part-time opportunities to balance non-work related activities. About 65 % of part-time workers choose to work part-time (BLS).

retail organization cares about controlling labor expenses in the stores, so it sets a ratio of sales to labor as a target for store managers to achieve. Merchandising departments, on the other hand, have their incentives tied to sales of product categories. Since sales of certain product categories, such as appliances or shoes, would be sensitive to labor, the merchandising department may want appropriate coverage of those departments with labor presence. Finally, store operations care about having sufficient labor to cover the large number of non-customer-facing tasks in a store. While the different groups provide feedback on the amount of labor in the store, many retailers ultimately let the store manager determine the right amount of labor in their stores. By tying the store managers' bonuses to profits, the corporate office tries to overcome the classic agency problem that arises in these situations.

In this book chapter, we present an overview of industry practice around workforce management and empirical research on this topic. Even with such a narrowly defined goal, it was necessary to add further restrictions to strike a balance between the depth and breadth of the topics covered. This book chapter is largely restricted to U.S. public retailers. The industry practice explained here is based on our experience with several specialty and big-box retailers and workforce management software providers, and has been validated through presentations to numerous retail practitioners. However, there are likely to be deviations between the labor planning practices described in this chapter and those followed in other retail settings. Consistent with the contemporaneous nature of the empirical research in this area, the literature survey weighs recent papers more.

Next we explain workforce management planning practice in detail for one of the retailers in Sect. 2. In Sect. 3, we review the literature around labor planning, with emphasis on empirical research in retail labor in response to the emerging interest in this area. In Sect. 4, we discuss some of the new technologies shaping the retail landscape that have implications for retail labor. We conclude with directions for future research in Sect. 5.

2 Labor Planning in Practice: Case Study of HomeRetail

In this section, we explain the labor planning practice at HomeRetail, a pseudonym for the retailer with whom we interacted. HomeRetail is a big-box retailer with annual revenues exceeding $1 billion. This retailer is in the home goods industry and carries over 10,000 items in its stores. This retailer employs year-round full-time and part-time employees and seasonal employees for a shorter duration of time to meet its annual sales spike. The labor planning practice is similar to that of many other big-box retailers with whom we have interacted. Specialty retailers tend to have smaller stores, and their labor planning process tends to be much simpler than the one described in this section.

Due to the large sizes of its stores (over 100,000 sq. feet with more than 100 employees), this retailer has a deep organizational structure for each of its

store. Each store was divided into multiple departments based on the product category, and each of those departments have a department manager who is responsible for managing labor associated with that department. The labor within each department is divided into various roles such as sales associates, specialists, cashiers, backend delivery, and assembly, etc. The department managers are incentivized based on sales and profits in their department. The different department managers report to assistant store managers, who in turn, report to the store manager. Store managers also had a human resources (HR) manager to help them with recruiting workers. HR managers play a vital role during the peak season, when they need to hire a large number of seasonal employees for the store, train them, and manage their exit at the end of the season.

Next we explain the labor planning process at HomeRetail in detail. We divide the labor planning process into long-term and short-term planning, where long-term planning refers to planning for 1 year and short-term planning is the planning for near term, such as the next month or two.

Long-term planning: Long-term planning is typically done at the beginning of the fiscal year when retailers revisit the organizational structure for each store and the minimum staff required to manage a store. HomeRetail groups stores based on their sales volume into different tiers. Stores in each tier are allocated base hours, that is the minimum hours per week, for different roles, such as assistant store managers, human resources (HR) manager, cashiers, sales associates, and department managers. These base hours guide the store managers to determine the number of full-time and part-time workers to have in the store on an ongoing basis. Though store managers are given some direction on the proportion of part-time to full-time workers to have in their stores, we observe that they have considerable leeway to deviate from this suggested proportion. If store managers need to recruit additional workers for their stores, they do so with the help of the HR manager in the store.

Short-term planning: While long-term planning enables store managers to get the right number of full-time and part-time workers in place, short-term planning involves balancing the labor hours required in a given month to the workforce in place. This stage begins with the determination of labor hours that need to be staffed for a given month. At HomeRetail, the store managers, the district managers, and the corporate finance team jointly forecast sales for a month, usually 30 days or more in advance. The sales forecast is then used as an input to a regression model that was estimated using historical sales and labor data to predict the labor hours required to satisfy the forecasted sales. These labor hours are communicated to the store manager, who needs to ensure that a sufficient number of workers exist to cover those hours.

Store managers would then schedule full-time and part-time workers to ensure coverage. Typically, managers use software tools to match worker availability with the workload requirements of the store. The workload requirements are driven based on the number of operational activities they need to perform as well as the labor required to support sales tasks, as predicted by the sales forecast. This tool also takes into account several restrictions imposed by minimum labor

requirements set by corporate, local labor laws, union rules, quality of life considerations etc., while determining the final schedule. These schedules are generally posted a week or more in advance so that associates can plan accordingly. At HomeRetail, full-time workers typically worked 8-h shifts for 5 days a week and were asked to work a minimum number of weekend days in a month.[3] So, store managers had some limited flexibility on shift lengths and shift days for full-time employees. Part-time workers offered more flexibility, as they could work for variable shift lengths and for different days of the week.

While the above approach works for most of the year, the forecasted hours could exceed the capacity provided by full-time and part-time workers during peak periods. We find that many stores double their sales during the peak period, so even a fully cross-trained staff would not be able to handle the demand surge necessitating hiring of seasonal workers. While, by convention, peak period coincide with the holiday season, some retailers such as Home Depot and Lowe's begin their peak periods in the spring.

Recruiting and onboarding seasonal workers are challenging tasks for retailers and consume a lot of the attention of store management. Because demand during peak period can be twice as large as that during the non-peak period, stores need to aggressively recruit seasonal workers to maintain service quality. For example, HomeRetail invests heavily in building relationships with local colleges as well as the community as a whole to ensure sufficient supply of seasonal workers to its stores. Many store managers mention that they often start planning for recruitment for the next peak season right at the end of the previous peak season. However, for a majority of stores, the active planning stage for the seasonal workers begins 4 months before the beginning of the peak period, when the area HR manager in consultation with the store manager identifies the approximate number of seasonal workers that the stores may need for the upcoming peak period. This process aligns the corporate managers with the needs of the stores. However, formal recruiting does not begin at this stage. The actual recruiting process takes anywhere between 1 and 3 months for HomeRetail. We explain this process of recruiting seasonal workers next.

When the store manager is ready to recruit seasonal workers, they request approval from the district manager. Once approved, the store's HR manager creates a job description depending upon whether the seasonal worker is required for cashiering, sales, stocking shelves, unloading trucks, or some other role. This job description is posted internally before being communicated to local colleges and other sources of seasonal workers. HomeRetail, for instance, requires its stores to interview three candidates for every position. In addition to multiple rounds of interviews, the candidates also need to undergo drug testing and background checks

[3] At another major apparel retailer that we worked with, full-time workers were asked to work four long shifts of 9 h each and one short shift of 4 h. Shorter shift lengths can increase store profitability significantly (Mani et al. 2014), however associate dissatisfaction could also increase.

before they receive an offer. Thus, even if candidates are readily available, the process of bringing a candidate to a store could take at least a month.

Once workers are recruited, stores follow the essential step of onboarding them by providing appropriate training. The extent of training can vary considerably from retailer to retailer and store to store and by whether workers are full-time, part-time, or seasonal. For example, Fisher and Krishnan (2005) document the case of Wawa convenience stores, where store managers are responsible for training the associates. At HomeRetail, this training is provided partly by the corporate office through centralized web tools and supplemented by store manager and department managers. Unsurprisingly, we find that full-time and part-time workers receive longer periods of training compared to seasonal workers.

3 Literature Review

Labor planning is not new to operations management; indeed, a long history of mathematical models and scheduling algorithms has evolved to optimize staffing requirements. Most of these models have been developed (and successfully applied) in the context of a manufacturing setting. However, some key differences exist between a manufacturing and a retail operation that prevent direct application of these models in a retail store. Since the manufacturing setting is well known to the operations management audience, we begin by highlighting the key differences in labor planning between retail and manufacturing industries. We then discuss the emerging area of empirical research on retail labor in detail.

3.1 Differences Between Manufacturing and Retail Settings

Early work on labor planning in operations management literature concentrated mainly on determining labor requirements in manufacturing environments. The main focus was on determining optimal (or near optimal) solutions to labor requirements in the context of aggregate planning. Aggregate planning is an intermediate-range capacity planning process that typically covers a time horizon of 2–12 months and involves simultaneous determination of a firm's production, inventory, and employment levels over this time horizon to meet the total demand for all products that share the same limited resources. The objective is to minimize the total cost (or expected cost in case of uncertain demand) while taking into consideration constraints on the production rates and changeovers as well as inventory and workforce levels. The cost parameters would include cost of production, inventory and shortage costs, cost of adjusting the production rate through over-time or under-time, and cost of adjusting workforce through hiring and firing employees. In most cases, all available workers were treated as equally productive and cost parameters have to be determined from actual financial data. Subsequent

research has dealt with incorporating labor flexibility as well as short-term decisions like workforce scheduling into the aggregate planning framework. Below we highlight a few relevant papers in this domain.

Starting with the seminal paper by Holt et al. (1956), several papers have developed mathematical models to find the aggregate production rate and size of workforce to meet demand. Linear programming and integer programming techniques are used to get the optimal decision rule that minimizes the total cost of regular payroll and overtime, hiring and layoffs, and inventory and shortages incurred during a given planning interval of several months (Lippman et al. 1967). Continuing studies on the problem of determining labor requirements in job shops, later researchers have also used stochastic programming techniques to cope with non-stationary stochastic demand for labor (Dill et al. 1966; Anderson 2001). In many of these papers, quadratic or convex cost functions are used to represent the cost of hiring, firing, and use of overtime (Kunreuther and Morton 1974). Quadratic cost functions are used to penalize deviation of key variables from target levels. The advantage of using quadratic functions is that they result in linear production rules that can be easily applied in a repetitive manner once the constants in the model are determined (e.g. the linear decision rule in Holt et al. 1956). Convex cost structures arise when marginal hiring (firing) costs increase with the number of employees hired (fired). This could arise when there are steep increases in costs with addition of a new shift, technological and productivity changes, labor slowdowns, etc. These cost structures are usually approximated by piecewise linear cost functions and add substantial complexity and computational effort to the problem.

In contrast to continuous assembly line manufacturing environments, job shops are characterized by batch-processing and may require additional skilled labor for specific type of jobs. Thus, all labor units cannot be treated equal in the aggregate planning problem. Subsequent work in this field has looked at incorporating labor flexibility into the aggregate production and workforce planning in the context of job shop planning (Fryer 1974; Brownell and Lowerre 1976). Later work has looked at impact of different cross-training policies on performance of serial production systems with an objective of minimizing the costs of cross-training while meeting staffing requirements (Daniels et al. 2004; Hopp et al. 2004; Bard and Wan 2008). Extensive work also exists on determining detailed shift schedules for employees. A large body of academic literature has developed mixed integer linear programming techniques for scheduling full-time workers to minimize labor hours while satisfying variable workforce requirements of a service delivery system (Dantzig 1954; Morris and Showalter 1983). Considerable work has also been done on modeling workforce requirements based on multiple shifts, by incorporating the effect of constraints on the changing of shifts over the planning period. The common approach in these papers is to use integer programming to determine optimal shift schedules that include flexible rest or meal-breaks, and allow for alternate shift starting times, shift lengths, and break placement (Bechtold and Jacobs 1990; Thompson 1995).

Several differences exist between labor planning processes in manufacturing and retail. The most important source of these differences is that customers often interact with a service provider to jointly produce the outcome, a process known as customer co-production (Karmarkar and Pitbladdo 1995). Co-production requires the real-time involvement of the customer with the store associate for its successful completion; as such, inventorying service in anticipation of future demand is typically not possible. Thus, a key challenge faced by retailers when planning labor is to ensure that their workforce is available when customers walk into their stores. Significant variabilities in customer arrival process occur within a day, across days of week and across months of the year and make it hard for retailers to match labor with demand. The Figs. 6.1, 6.2 and 6.3 depict these three types of variabilities based on traffic data from 41 stores of a women's apparel retail chain. The low, med and high lines depict the 10th, 50th and 90th percentile of average traffic across these 41 stores. Unlike manufacturing settings in which the order lead time enables manufacturers to mitigate the effect of forecast errors by shifting orders across time or facilities, under- or over-staffing in retail settings can have immediate impacts on financial performance of the stores. Mani et al. (2014) find stores to be understaffed about 41 % of the time in their retail setting and find the impact on lost sales and profitability to be managerially significant.

Another implication of co-production is that labor affects not only costs but also sales directly. In manufacturing, although labor is a part of the production process, it is not a part of the end-product. Thus, the quality of a product in manufacturing can be made independent of labor through proper oversight and inspection. Defects can be identified and reworked ahead of sale. On the other hand, co-production in retail implies that store labor has a direct impact on sales through the customer-observed service quality. Marketing research shows service quality to be an important determinant of customer satisfaction. Maxham et al. (2008) describe a retail

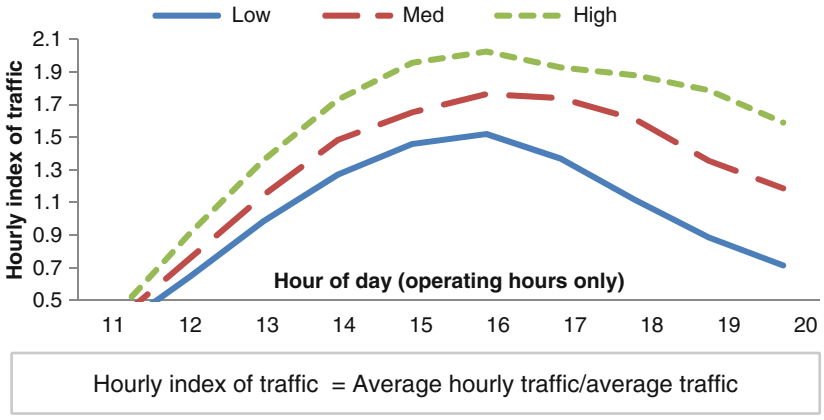


Fig. 6.1 Plot of hourly variation in traffic

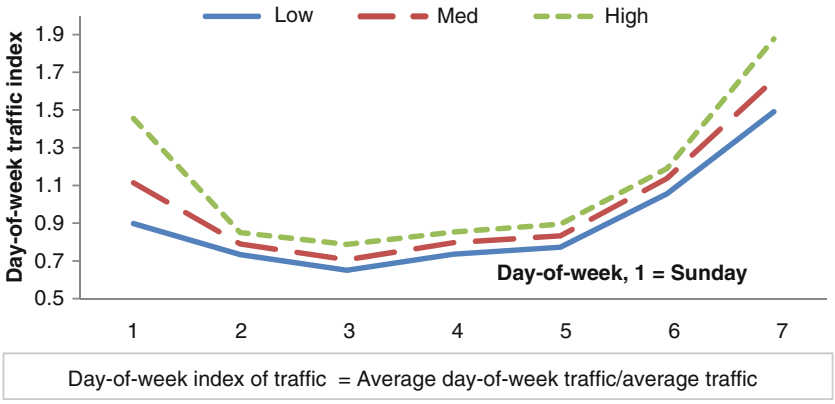


Fig. 6.2 Plot of day-of-week variation in traffic

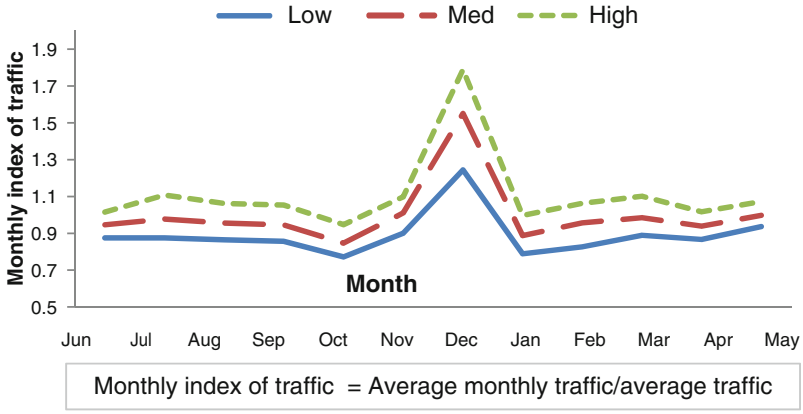


Fig. 6.3 Plot of monthly variation in traffic

value chain wherein perceptions and behaviors of front-line store employees influence customer satisfaction and intent, and ultimately store performance. Through empirical models that examine systems of relationships among employee job perceptions, employee performance, customer evaluations, and store performance, they find that employee perceptions exert a direct influence on customer evaluations, and that customer evaluations affect retail store performance (customer spending and comparable store sales growth). For this reason, retailers need to consider the skills of the associates before letting them perform customer-facing tasks.

Another important difference arising from co-production is that customers in a retail setting impose additional externality costs. While queue length of jobs can be

optimized in manufacturing to minimize costs, customers at a retail store may join or leave queues in a way that is sub-optimal to the whole system. Finally, from a skill-set standpoint, retailers can often recruit workers with limited or no background in retail and train them before introducing them in the store. While this practice increases the pool of workers to recruit from, it can also lead to higher turnover, as these positions are often low paying ones. In contrast, the manufacturing setting tends to use smaller proportions of part-time and temporary workers as compared to the retail industry (BLS 2008; Lockard and Wolf 2012).

3.2 Empirical Research on Retail Labor in Operations Management

In this section, we provide a review of the empirical research on retail labor. Empirical research examining the impact of labor on retail store performance has been gaining importance in recent years. In order to empirically investigate the impact of labor on retail store performance, it is necessary to collect data on the amount of labor available in the store, the demand for labor, and store performance measures.

The amount of labor available in the store depends on the type of labor (e.g. full-time, part-time, temporary workers), the job description (e.g. store managers, sales associates, stockroom employees etc.), and the number of hours available for each employee (e.g. maximum number of hours available, break time, vacation time etc.). This information is usually available from personnel records. It is typically gathered at an aggregate level (e.g. bi-weekly or monthly) and used to generate wage-payroll data. More detailed information on the exact number of labor hours within each day would be available from a workforce management system or a scheduling system which uses the number of hours available for each employee and breaks them down into shift schedules. While the data from the workforce management systems would give a detailed breakdown of number of labor hours available, it might be aggregated across employees. For instance, the labor hours available for a given day would be the total manager-hours, full-time labor hours and part-time labor hours available for each hour of the day. Depending on the kind of store operation, the labor hours might be broken down into stockroom labor hours and sales associate hours.

The demand for labor depends on the type of retail store and their product characteristics. For instance, the level of sales assistance provided to shoppers in a specialty furniture store is significantly different from that provided at a discount super store. In addition, the demand for retail labor is also dependent on the level of store execution activities like unloading delivery trucks, stocking shelves, tagging merchandise, and maintaining the overall store ambience. While it is possible to estimate time requirements for standard activities such as unloading delivery trucks and stocking shelves, it is very difficult to set a time for sales-related activities,

especially in a service-intensive store (Fisher and Raman 2010). Thus, unlike manufacturing or call-center settings, it is relatively much harder in a retail setting to estimate the exact number of hours required in the store from store activities alone. Hence, it is common practice to estimate the demand for retail labor from the level of sales or the level of traffic in the store. Store managers may then add additional hours required for store-execution activities to determine the total labor hours required in the stores. Store sales data are available from point-of-sale (POS) systems that are installed in almost all retail stores today. Data on traffic are harder to collect and requires a traffic counter to be installed in the stores. Stores that have traffic counters would have customer arrival data that can be aggregated and used for analysis. In some instances (e.g. grocery stores), where almost every customer who enters the store leaves with a purchase, the number of transactions from the POS data can be used as a proxy for traffic. However, in specialty stores (e.g. high-end electronics stores and specialty stores), it would be necessary to have access to traffic data to assess the true demand for retail labor.

Store performance measures include both quantitative measures (like revenue, profits, and conversion rate) as well as qualitative ones (like the level of service provided in the stores). Quantitative measures like sales, expenses, and profits are usually available from financial data. While information on store sales can be obtained from POS data, labor expenses are gathered from wage-payroll data, usually at the monthly level. In order to ascertain other expenses like inventory shrink, administrative expenses etc. it is necessary to have access to stores' financial (P&L) statements. These financial measures can be used to construct additional measures like labor productivity (e.g. sales per labor hour) and level of employee turnover. If traffic data are also available, then additional store performance measures like conversion rate (defined as the ratio of number of transactions to traffic), basket value (ratio of sales volume to number of transactions), and traffic to associate ratio (ratio of number of customers to number of sales associates) can be calculated. Qualitative measures on service quality are obtained from customer surveys.

The two most common challenges encountered in conducting empirical research on retail labor are data availability and dealing with endogeneity issues between labor and store performance.

As mentioned earlier, traditional data on retail labor and store performance have been available from POS transactions and wage-payroll. Due to the sensitive nature of these data, many retailers are reluctant to share them.[4] These data are typically aggregated before archiving. The POS data may be available on a daily basis, or even hourly basis, but payroll data are usually available only at a bi-weekly or monthly level. Retailers who have installed workforce management systems typically have labor data at a disaggregate level that are typically more useful for research on labor planning and scheduling. They usually have information on when

[4] In our experience, we find retailers to be particularly sensitive to sharing age and gender information when providing the payroll data.

and which department each person worked on different days of the week. These data can even be available at 15 min intervals. Retailers who have traffic counters can provide customer arrival data. However, unlike call center data where traffic data are usually accurate, retail traffic counters typically have some errors. While it is common for many technology firms to claim that these errors are less than 5 %, it would be useful to verify these data for their accuracy, if possible, before use in research. Finally, we note that while POS and payroll data are available for a long time period for most retailers, granular data on traffic and labor hours are typically only available for a shorter time period as these systems have not been in place for a long time.

Another important issue to consider when examining store labor is that of unobservable factors that may result in omitted variable bias. A key concern with examining the impact of labor on store performance is that labor typically gets scheduled based on certain anticipated events that the manager knows but is unknown to the empirical researcher. For instance, consider the case of store promotions. When store managers run store promotions, they may hire more labor in advance of these promotions. Store promotions lead to an increase in store traffic and store sales. Without data on store promotions, examining the impact of labor on store performance would lead to misleading inferences. So, it is necessary to either control for sales forecast, which account for store promotions and other anticipated events that affect sales, or use appropriate instruments to overcome the endogeneity bias.

In Sect. 3.2.1, we describe the empirical models that have used sales and payroll data to examine the overall relationship between retail labor and store performance. In Sect. 3.2.2, we describe empirical models that deal with customer traffic and staffing issues in retail stores. In these models, traffic data is used along with store sales and labor data to determine the relationship between demand, availability of labor, and store performance. A recent development in retail labor planning literature is the consideration of the type of retail labor available in the store and its impact on store profits. In Sect. 3.2.3, we highlight empirical models that leverage labor-mix data to examine its impact on store productivity and profits.

3.2.1 Relationship Between Retail Labor, Quality, Sales, and Profits Using Sales and Payroll Data

Store labor is an important driver of retail store performance. The benefits of having store labor include providing an increased level of sales assistance to shoppers and improving execution of store operational activities such as stocking shelves, tagging merchandise, and maintaining the overall store ambience (Fisher and Raman 2010), all of which lead to increased sales. Below, we look at two empirical models that examine the relationship between retail labor, service quality, and store performance.

Retail Labor and Basket Values

Netessine et al. (2010) use sales and payroll data from 311 stores of a large retail chain over a 3-year period to study the relationship between store labor and basket values. They collect monthly level data on sales, number of transactions recorded at checkout, the value of shopping baskets, and the total number of employee hours (full-time, part-time, and manager hours) budgeted for the store in a given month.

Based on these data, the authors derive two kinds of mismatches between sales and labor: *Planning mismatch,* which measures the quality of store labor planning using the month-to-month deviations (mismatches) between forecasts of store transactions and planned labor hours, and *Execution mismatch,* which measures the quality of store labor deployment using the month-to-month deviations between planned labor and actual labor deployment. The labor mismatches are calculated as a function of the correlation ($r(.)$) between two time series of corresponding variables. For example, for store i using monthly observations on transactions (TXN_i), labor plan hours ($PLAN\_HOURS_i$) and total employee hours ($TOTAL\_EE_i$), total labor mismatch ($TXNvsTOTAL\_EE_i$), planning mismatch ($TXNvsPLAN\_HOURS_i$) and execution mismatch ($TOTAL\_EEvsPLAN\_HOURS_i$) are calculated as follows:

$$TXNvsTOTAL\_EE_i = 1 - r(TXN_i, TOTAL\_EE_i);$$
$$TXNvsPLAN\_HOURS_i = 1 - r(TXN_i, PLAN\_HOURS_i);$$
$$TOTAL\_EEvsPLAN\_HOURS_i = 1 - r(TOTAL\_EE_i, PLAN\_HOURS_i)$$

The authors use basket value as a measure of store performance and find that the mismatches between store transactions and the total number of employees are negatively associated with basket value (significant at the 5 % level).

Next, they separate the total labor mismatch into planning mismatch and execution mismatch and find that while planning mismatch is negatively associated with basket values (significant at the 1 % level), the association between execution mismatch and basket values is not significant. They further break down execution mismatches based on type of labor (i.e. full-time labor, part-time labor and managers) and find high statistical significance for an association between full-time employee mismatch and average basket value. However, mismatches for part-time employees and store managers were not statistically significant. The regressions are run on cross-sectional data for each store and include control variables for demographics for each store such as household size, proportion of households with no children, and the proportion of the local population that is Asian or Hispanic.

Finally, the authors find that some stores are consistently better at planning staffing levels to meet traffic, while other stores are consistently better at executing a given plan, but no correlation appears between the ability to plan and the ability to execute well. For the retail chain in their study, they conclude that if managers were able to reduce staff planning mismatches by 50 %, the resulting revenue uplift would be 1.8 % of the current chain-level revenue. Eliminating 50 % of execution

mismatches creates an additional revenue uplift of 2.4 %. In conclusion, the authors propose, as ideal, a switch from forecasted sales to forecasted traffic as a basis for labor planning.

Retail Labor, Quality, and Store Profits

Ton (2009) investigates the impact of store labor on store profits through its impact on service quality and conformance quality. In retail stores, increasing the labor level is likely to increase both conformance quality and service quality. For example, when store employees have more time, they are less likely to make errors in activities such as shelving merchandise or placing price tags on display shelves, and more likely to spend time with customers. In turn, sales are likely to be higher when products are shelved properly (Ton and Raman 2010) and salespeople are available to help customers in the purchase process (Fisher et al. 2006). Conformance quality is also expected to increase future sales at retail chains that use centralized merchandise planning systems, as the performance of these systems depends on conformance to in-store merchandising specifications and on accurate point-of-sale and inventory data (Raman et al. 2001). In addition to increasing sales, conformance quality is also likely to improve labor productivity and reduce shrink. Employees can shelve, replenish, and help customers find products more quickly, and fewer products are expected to be damaged or lost. Based on these arguments, Ton (2009) explores the relationship between labor, service and conformance quality, and profitability in retail stores.

Ton (2009) uses monthly data on labor, service quality, and profitability from 1999 to 2002 from 286 stores of large specialty retailer *Beta*. The amount of labor is measured as total labor dollars spent at a store in a given year and includes wages and benefits. The profit margin is defined as the operating income divided by sales. To measure service quality, she uses information from customer surveys that ask questions on five dimensions of service quality: tangibles, responsiveness, assurances, reliability, and empathy (Zeithaml et al. 1990). Ton (2009) also uses three metrics tracked by *Beta* to calculate conformance to the centralized decisions on merchandise planning and display: phantom-products, returns-conformance, and store-conditions. Phantom-products tracks the percentage of products that are in storage areas but not on the selling floor at the time of the physical audit. Returns-conformance tracks whether stores return the products they are supposed to return to the distribution centers. Store-conditions tracks whether stores conform to a wide range of standards related to the flow and storage of products. To create a composite measure of conformance quality, she standardizes each measure of conformance quality for each year by subtracting the mean and dividing by the standard deviation. In the final measure, returns-conformance and store-conditions scores are added and phantom-products scores are subtracted from the average standardized scores.

In the paper, Ton (2009) first tests for the relationship between quality (service and conformance quality) and labor and for the relationship between profit margin

and labor. Next, she tests if the relationship between profit margin and labor is mediated by service quality and conformance quality. The regression models include fixed effects for each store and for each year. Store fixed effects control for time-invariant unobserved heterogeneity across stores, which might otherwise affect store labor, conformance quality and service quality, and profitability. The year effects control for factors, such as economic conditions and corporate policies, which if they change over time, will change for all stores. The control variables in the regressions include planning mismatch (measures the degree of mismatch between a store's payroll plans and its actual workload), and execution mismatch (measures the degree of mismatch between payroll plans and actual labor spending) for the different stores. Store monthly sales are used as a proxy for workload. Also included are full-time employees as a percentage of total employees to control for employee mix, employee turnover to control for tacit knowledge lost when employees leave, store manager turnover to control for management changes, units of inventory in a store to control for level of complexity in the operating environment, unemployment rate in a store's MSA (Metropolitan Statistical Area as defined by the Census Bureau) to control for differences in labor market conditions, and the number of competitors in the local market to control for competition.

The results indicate that increasing labor at a store is associated with higher conformance quality and service quality. Increasing employee turnover and departure of store managers are associated with a decrease in conformance quality and service quality. Higher planning mismatch and increased complexity in operating environment are associated with lower conformance quality but have no effect on service quality. Increasing the proportion of full-time employees has no effect on conformance quality but, surprisingly, a negative effect on service quality. Finally, she finds that a 1 standard deviation increase in labor is associated with a 10 % increase in profit margin.

Netessine et al. (2010) and Ton (2009) both conclude that most stores tend to understaff their stores. Fisher and Raman (2010) cite conversations with many retail managers and conclude that most retailers view labor as a cost, not an asset. To the managers, decisions about staffing trade off a known present cost—paychecks written to employees—against an unknown future benefit, namely, the incremental sales that would result from better staffing. Hence, managers tend to focus more on lowering staffing costs. Further, since the negative effect of having too little labor is often difficult to quantify, they posit that many store managers may place greater emphasis on minimizing payroll expenses to meet short-term performance targets. In the next section, we discuss papers that aim to quantify the impact of labor on store sales and profit by using more detailed data on store traffic, labor, and sales.

3.2.2 Relationship Between Store Traffic, Retail Labor, and Store Sales

In an effort to track the true sales potential in their stores, retailers have recently begun to install traffic counters in their stores. Traffic counters enable retailers to collect data on customer traffic and track conversion rate in their stores. The

availability of this data has also opened up new avenues for research on retail labor. By combining traffic data with point-of-sale (POS) and labor data, it is now possible to estimate the true customer demand and the lost sales due to inadequate labor. Below we look at two papers that leverage traffic data with sales and labor data to examine these issues.

Effect of Traffic on Retail Sales Performance

Perdikaki et al. (2012) use data from 41 stores of an apparel retailer (*Alpha*) to study the relationship between store traffic, labor, and sales performance. They decompose sales volume into conversion rate and basket value. Increase in traffic would lead to an increase in sales, as higher traffic provides more opportunities for sales conversion. However, in the absence of adequate labor, increase in traffic could lead to higher levels of crowding and a decrease in service quality, both of which could lead to a decrease in sales. Thus, having adequate store labor could moderate the impact of traffic on store sales. Based on the above observations, the authors examine the relationship between traffic, labor and store sales; and study if higher store labor leads to greater positive impact of store traffic on store sales performance.

In addition to studying the impact of traffic, the authors also explore the impact of traffic variability on store sales. Stores with higher inter-day traffic variability may face higher traffic uncertainty, which could result in large errors when forecasting labor requirements for stores. Such large forecast errors would result in large mismatches between store labor required to manage in-store customers and actual store labor. Increased intra-day traffic variability could also lead to higher waiting time in queues and result in higher levels of abandonment. Further, higher levels of intra-day traffic variability could cause difficulties in scheduling labor for different hours of the day. This could lead to understaffing during certain hours of the day, resulting in lower service quality and lower sales performance. Hence, the authors test if greater inter- and intra-day traffic variability could lead to lower store sales performance. Finally, they also explore the implications of lower conversion rate on future sales potential by studying the relationship between conversion rate and traffic growth.

For the year 2007, the authors obtain the following types of data: (1) financial data (i.e., the number of transactions and store sales volume); (2) labor data (i.e., employee hours); and (3) traffic data. Sales performance for store i on day t is measured in two different ways: sales volume in dollars and the number of transactions that occur in the stores. These variables are divided by regular business hours for each store on each day of the week to obtain the average number of transactions $(ATXNS_{it})$ per hour and average sales volume per hour $(ASALES_{it})$. Similarly, total traffic and labor hours are divided by regular business hours to obtain average traffic per hour $(ATRAF_{it})$ and average labor hours per hour

($ALBR_{it}$). The authors calculate intraday traffic variability, using hourly data, as the ratio of standard deviation to mean traffic for that day as shown below:

$$\mu_{it} = \sum_{h=1}^{H_{it}} TRAF_{ith}/H_{it}; \sigma_{it} = \sqrt{\sum_{h=1}^{H_{it}} \left(TRAF_{ith} - \mu_{it}\right)^2 / H_{it} - 1}; TRAFVAR_{it} = \sigma_{it}/\mu_{it}$$

where $h = 1 \ldots H$ represent the store business hours. Inter-day traffic variability is calculated using the following AR model for traffic where δ_h denote holiday dummies, δ_m denote monthly dummies, and δ_d denote dummies for days of week.

$$TRAF_{it} = b_{i0} + \sum_{l=1}^{7} b_{il} TRAF_{it-l} + b_{i8}\delta_h + b_{i9}\delta_m + b_{i10}\delta_d + \varepsilon_{it} \tag{6.1}$$

Traffic variability is measured using the residuals (Eq. 6.1) as $TRAFUNC_i \equiv sd\left(\varepsilon_{it}/TRAF_{it}\right)$ where sd(.) denotes the standard deviation. The authors include the following control variables. They calculate labor-traffic mismatch as the ratio of traffic to labor $\left(Mismatch_{it} = \frac{\sum_{h=1}^{H_{it}} TRAF_{ith}/LBR_{ith}}{H_{it}}\right)$. This ratio is used as a proxy for service level. The authors also collect data on daily average temperature of each store location and the Dow Jones Industrial Average and obtain demographic data like averages on median household income and per capita income for 2007 by location. They use the number of other stores in the mall as a proxy for competition and run the following regression to determine the impact of traffic and labor on store sales. W_{it} denotes the vector of control variables

$$\begin{aligned} ASALES_{it} = {} & \vartheta_0 + \vartheta_i + \vartheta_1 ATRAF_{it} + \vartheta_2 ATRAF^2{}_{it} + \vartheta_3 ALBR_{it} \times ATRAF_{it} \\ & + \vartheta_4 ALBR_{it} \times ATRAF^2{}_{it} + \vartheta_5 TRAFVAR_{it} + \vartheta_6 ALBR_{it} \\ & + \vartheta_7 ALBR^2{}_{it} + \vartheta_8 W_{it} + \xi_{it} \end{aligned} \tag{6.2}$$

The authors find that store sales volume is an increasing concave function of traffic. For values of labor corresponding to mean, and traffic at mean plus 1 sd, increasing average traffic per hour by one unit increases average sales volume by $8.14. For values of labor corresponding to the mean, and traffic corresponding to mean minus 1 sd, increasing average traffic per hour by one unit increases average sales volume by $11.80. For values of labor corresponding to mean, mean minus 1 sd, and mean plus 1 sd, the marginal returns to traffic for the store with mean traffic are $10.00, $8.68 and $11.32, respectively. This relationship is shown graphically in Fig. 6.4. Further, the authors find that store sales volume exhibits diminishing returns to labor and increases in intraday traffic are associated with lower sales per hour in stores. Replacing $ASALES_{it}$ with CR_{it} and ABV_{it} in Eq. 6.2 yield similar results, supporting a decreasing nonlinear relationship between traffic and conversion rate. This result is shown graphically in Fig. 6.5.

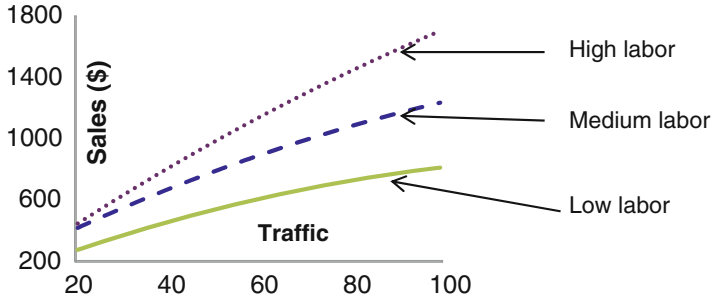


Fig. 6.4 Relationship between store traffic and sales for different levels of labor

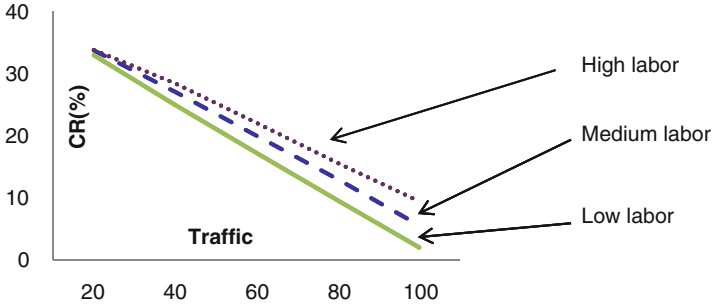


Fig. 6.5 Relationship between conversion rate and store traffic for different levels of labor

In addition, the authors find that increases in inter-day traffic variability are associated with lower sales per hour in stores. Further they find that an increase in conversion rate is associated with an increase in future traffic growth and that this relationship is statistically significant up to 5 months in the future.

Estimating the Impact of Understaffing in Retail Stores

Retailers walk a fine line between having enough labor in their stores to meet service requirements while maintaining low payroll costs. As pointed out by Netessine et al. (2010) and Ton (2009), quantifying the impact of understaffing on store sales and profits is necessary so that managers can make informed decisions on the level of labor to have in their stores. Mani et al. (2014) use detailed traffic data along with labor and sales data to investigate whether retail stores are understaffed and the impact of understaffing on lost sales and profits.

Using hourly traffic data along with POS (point-of-sale) and labor data for 41 stores of an apparel retail chain, the authors first calculate the required amount of labor for each store during each hour. They denote positive deviations from this

required labor as understaffing and negative deviations as overstaffing. They follow two different approaches to labor planning. The first approach uses reduced-form estimation of an empirical model to obtain predicted staffing levels and the second approach uses a structural estimation methodology to obtain optimal staffing levels.

In the first approach, the authors use an empirical model motivated by the square-root staffing model from queueing theory to calculate staffing levels. For each store i in time period t, let $TRAF_{it}$ the number of customers arriving to the store. Then, the target staffing level (N_{it}) can be determined based on the following equation:

$$\begin{aligned} N_{it} = {} & \delta_{0i} + \delta_{1i} TRAF_{it} + \delta_{1pi} TRAF_{it} \times \left(1_{p=1}\right) \\ & + \delta_{2i} TRAF_{it}^{1/2} + \delta_{2pi} TRAF_{it}^{1/2} \times \left(1_{p=1}\right) + \xi_{1it} \end{aligned} \tag{6.3}$$

In the above equation, the authors introduce a dummy variable for peak hours to take into account changes in service rate and quality of service between peak and non-peak hours. The peak hours are determined as a 3-h window during which almost 60 % of store traffic arrives during the day. To quantify the impact of understaffing on sales and profits, they use the following sales and profit functions:

$$S_{it} = \alpha_i TRAF_{it}^{\beta_i} e^{-\gamma_i / N_{it}}; \pi_{it} = S_{it} \times g_{it} - N_{it} \times d_i \tag{6.4}$$

where S_{it} is the store sales, β_i is the traffic elasticity, γ_i captures the responsiveness of sales to labor (indirectly measuring labor productivity), and α_i is a store-specific parameter that captures the sales potential in the store, π_{it} is the gross profit net of labor costs, and d_i is the marginal cost of labor. In this model, overall store sales are positively associated with labor, but an increase in traffic and labor increases sales at a diminishing rate, i.e., $0 < \beta_i < 1, \gamma_i > 1$. The difference between required labor and actual labor for each hour is denoted by ΔN_{it}. Let $1_{\Delta N_{it}>0}$ be an indicator function that takes the value of 1 when the store is understaffed $\left(\Delta N_{it} > 0\right)$, 0 otherwise. The lost sales and drop in profits in time period t when the store is understaffed can be represented as:

$$\begin{aligned} \Delta S_{it} &= \left[\hat{\alpha}_i TRAF_{it}^{\hat{\beta}_i} \left(e^{-\hat{\gamma}_i / \hat{N}_{it}}\right) - S_{it}\right] \times \left(1_{\Delta N_{it}>0}\right); \\ \Delta \pi_{it} &= \left(\Delta S_{it} \times g_{it} - \Delta N_{it} \times d_i\right) \times \left(1_{\Delta N_{it}>0}\right) \end{aligned} \tag{6.5}$$

where "^" indicates the coefficients estimated from the sales equation in Eq. (6.4). Thus, the authors' estimation of lost sales is based on the sales lift that the store would have experienced if it carried the predicted labor $\left(\hat{N}_{it}\right)$.

The authors perform a cluster analysis based on average traffic as well as average sales and divide their sample into weekdays and weekends based on similarities in traffic patterns (and sales patterns) across different days of the week.

For the stores in their weekdays sub-sample, the authors find that stores are understaffed 40.21 % of the time. When understaffing occurs, the magnitude of

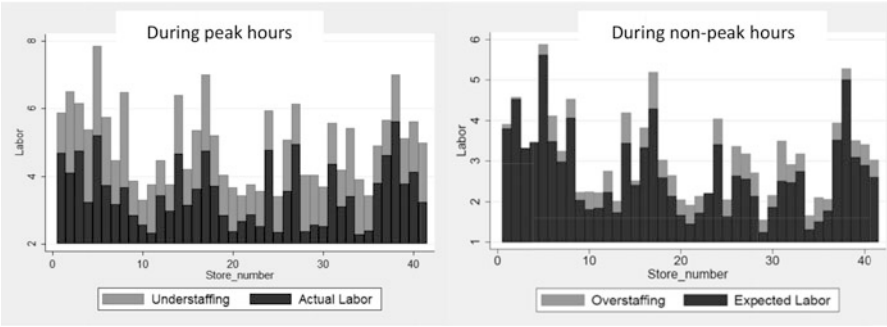


Fig. 6.6 Understaffing during peak and non-peak hours

understaffing is 2.10 persons; this level of understaffing represents a 33.27 % shortage as compared to the predicted labor. During peak hours, they find that the stores are understaffed 64.98 % of the time, and the average magnitude of understaffing is 2.31 persons. Figure 6.6 shows a graphic representation of understaffing during peak and non-peak hours across the 41 stores. Further, they observe a decline of 1.95 % in conversion rate when the store is understaffed (when compared to other peak hours when the store is not understaffed). They determine the average lost sales due to understaffing for the 41 stores during peak hours to be 8.56 %. Approximating g_i by the average gross margin for this retail chain and the labor cost, d_i, by the average wage rate for retail salespersons in that state, the authors find that this retail chain's average profitability will increase by 7.02 % if it eliminates understaffing during peak hours.

Next, the authors investigate the drivers of understaffing by studying the impact of forecast errors and scheduling constraints. They use 1-, 2-, and 3-week-ahead forecasts in place of actual traffic in Eq. (6.3) and calculate the predicted labor. As the forecast horizon increases from 1 to 3 weeks, the magnitude of understaffing as a percentage of predicted labor increases from 5.43 to 17.84 %. The sales lift decreases by 2.61 %, and the profitability improvement lowers by 2.56 % with use of a 1-week-ahead forecast of traffic. To examine how much of the observed understaffing can be explained by scheduling constraints, they consider 2-, 3-, and 4-h shifts in their analysis. They find that when scheduling labor with minimum shift lengths of 4 h, as opposed to 2 h, the magnitude of understaffing as a percentage of predicted labor increases from 7.23 to 28.74 %. The sales lift decreases by 3.76 % and the profitability improvement lowers by 3.52 % when they impose a 2-h shift length constraint. Finally, they explore the impact of the interaction of forecast errors and scheduling constraints on store profitability with the help of a simulation. As shown in Fig. 6.7, scheduling constraints exacerbate the negative impact of forecast error on store profits.

In the second approach, the authors use a staffing model based on a popular practice wherein the cost of labor in store is balanced with the contribution of labor to sales. Assuming that the store managers make optimal labor decisions at an

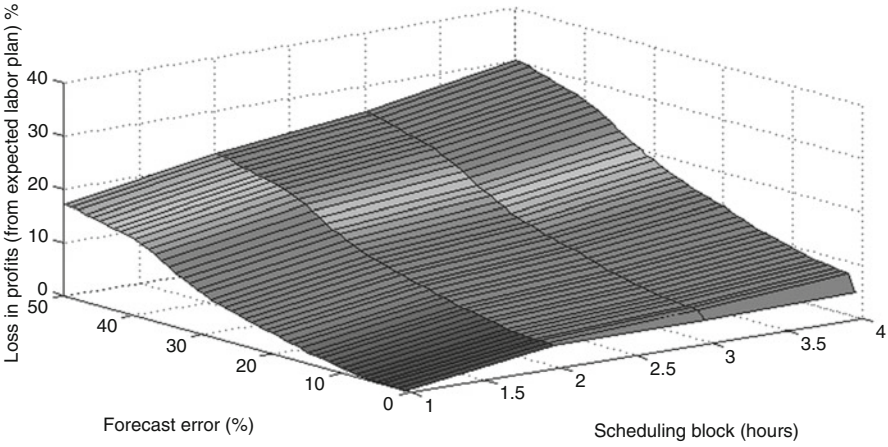


Fig. 6.7 Impact of forecast errors and scheduling constraints on store profits

aggregate daily level, the authors estimate the parameters of the model for each store using historical daily data on sales, traffic, and labor. They use the same sales and profit equations as in Eq. (6.4) but replace d_i with w_i to capture the intrinsic cost of labor that the store manager uses when deciding the amount of labor to have in the store. Each store manager is expected to maximize the profit function in Eq. (6.4), yielding the following first-order condition for amount of labor to have in each store:

$$\gamma_i \alpha_i TRAF_{it}^{\beta_i} e^{-\gamma_i / N_{it}} g_{it} = w_i N_{it}^2 \tag{6.6}$$

The optimal labor plan (N_{it}^*) is the value of labor that is a solution to Eq. (6.6), given α_i , β_i , γ_i , w_i and store traffic ($TRAF_{it}$).

The authors use the generalized method of moments to estimate α_i, β_i , γ_i, and w_i. They find considerable heterogeneity in the estimates of the imputed cost of labor across the 41 stores. For example, the average and standard deviation of w_i are \$58.87 and \$22.43, respectively. Even stores within the same state, having the same average wage rate for retail salespersons, had very different imputed costs of labor. Also, the authors find that the imputed cost of labor is significantly higher during weekdays than weekends ($p < 0.001$). Based on this approach, the authors find that during peak hours, the stores were understaffed 68.21 % of the time, the extent of understaffing was 3.52 persons, and removing understaffing would lead to a sales lift of 7.21 % and increase profitability by 5.87 %.

3.2.3 Relationship Between Employee Turnover, Labor Flexibility, and Store Performance

In Sects. 3.2.1 and 3.2.2, labor mismatches were shown to have a negative impact on store performance. One way retailers can reduce these mismatches is to increase labor flexibility—through use of part-time and temporary workers. This proposition is attractive, as part-time and temporary workers generally incur lower payroll expenses since they do not receive the full benefits of full-time workers and can be given non-standard work schedules. Thus labor flexibility helps retailers handle traffic variability and schedule sales associates for a few hours to meet peak demand. However, employing part-time and temporary workers could impact the quality of sales assistance provided. Also, retailers may have to contend with higher turnover rates among part-time and temporary employees as these employees look to move towards more stable working environments. Below, we describe empirical models that investigate the relationship between labor-mix, employee turnover, and store performance in more detail.

Employee Turnover and Store Performance

Ton and Huckman (2008) posit that performance at mature retail chains depends highly on the successful execution of known activities such as processing inventory, shelving merchandise, responding to customer queries, and transacting sales on cash registers. In such a setting, they expect employee turnover to have a negative effect on firm performance due to operational disruption from employee departures, additional work that must be absorbed by remaining employees, and the loss of tacit knowledge and accumulated experience held by departing employees. However, to the extent that stores operate with a high degree of process conformance, they expect that knowledge concerning task performance will more easily transfer to new employees. Based on the above observations, they propose that in case of settings requiring high levels of knowledge exploitation, turnover will have a negative effect on operating performance, and this effect will be moderated by the level of process conformance present in these settings.

Ton and Huckman (2008) conduct their analysis on data collected on 268 stores of Borders Group superstores over 48 months (1999–2002). The average annual full-time employee and part-time employee turnover across Borders stores ranged from 49 to 69 % and from 94 to 114 % respectively. The authors obtained monthly turnover and performance data for each store from 1999 to 2002. Profit margin is defined as operating income divided by sales, and they exclude temporary workers from the analysis. Turnover is calculated as the number of employees who left a store during that period divided by the average number of employees working at the store during that period.

To develop a composite measure of process conformance, they use the average store conditions score and the average return pull list (RPL) score for each store for

each year. In this setting, retailers are allowed to return unsold books to the publishers for a full refund minus the costs of shipping and handling. The RPL score is a returns conformance score based on the number of units returned divided by the total number of units that were supposed to be returned. The returns process, described in detail in the policy and process book, involves finding the books, packing them, and shipping them to the distribution centers.

Using these scores, they calculate the mean and standard deviation of store conditions and RPL scores across all Borders stores for each year. For each store, they standardize the yearly store conditions and RPL scores by subtracting the mean and dividing by the standard deviations. They combine these standardized scores to create the composite process conformance measure and divide stores into high and low process conformance stores. The authors run a regression of store performance against employee turnover. The regression model controls for several store level variables that vary over time. These include an indicator for turnover by store managers during the current month (to control for management changes); full-time employees as a percentage of total employees (to control for employee mix); total store payroll (to control for the total amount of labor used by the store); and the number of competitors in the local market and unemployment rate in the store's MSA (Metropolitan Statistical Area as defined by the Census Bureau) to control for labor supply. Specifications also include the fixed effects for each store, each year from 1999 to 2002, and each month of the calendar year. Next, to determine whether process conformance moderates the relationship between turnover and performance, they include an interaction variable between level of employee turnover with two categories of process conformance—high and low in the regression model.

The authors find that on average, turnover is associated with decreased store performance, as measured by profit margin and customer service. An increase of 1 standard deviation in full-time turnover at an average store leads to a reduction of 0.5 % in average customer service score and a 2.41 % decrease in average profit margin. The effect of turnover for low-process conformance stores is negative and significant. This negative effect is offset in stores with high levels of process conformance. Thus they conclude that turnover has a non-linear effect on performance.

Employee Turnover and Labor Productivity

Siebert and Zubanov (2009) examine the impact of turnover on labor productivity under two different work systems for sales assistants in a large UK retail organization. In the first system, known as the secondary system, the part-time employees receive less responsibility and specialist training, and have fewer promotion opportunities. Their pay is flat and is determined by salary surveys of similar occupations in the country. In the second system, known as the commitment system, full-time employees are given more responsibility, receive specialist training, and have their pay linked to performance. Managers expect that more turnover will occur under

the secondary system than under the commitment system. The authors suggest that some level of sales assistant turnover is beneficial for performance and test if a negative or an inverted U shaped relationship exists between employee turnover and performance under these systems.

The authors collect data for 325 stores. This data has three parts. In the first part, they collect personnel records of all employees who worked at any time between 1995 and 1999. Individual records include age, gender, date hired, date left employment, and weekly contract hours (measure of hours worked). Since individual productivity was not available, the authors derive employee average data for each store in order to compute store annual average productivity. The second part of the data contains store information such as revenues, store square footage, number of floors, and store environment variables like city center and retail park. The third part of data consists of area-wide wage and unemployment data for the county in which each store was located. They exclude stores that were opened or closed during 1995–1999.

The authors use labor productivity as a measure of store performance. Labor productivity is calculated as annual sales per store, adjusted for inflation, divided by total annual hours worked in the store. Employee turnover is measured as the separation rate. They calculate separations from employee records for the sales assistants working fewer than 30 h per week (defined as part-time separations) and 30 or more hours per week (defined as full-time separations). They calculate the full-time equivalent of separation rate as the number of hours that leavers would have worked had they not left, relative to annual total hours worked. The authors run a regression model of labor productivity against employee turnover and include control variables for capital input, measured as store size in square feet, and labor input, measured as the sum of hours worked by every sales-assistant employed in a given store in a given year. They also include a number of variables relating to store environment and employee characteristics that might also affect productivity. The store environment variables were store location (eight dummies, including city center and retail park), type of product (three dummies, indicating more or less expensive goods), share of children's goods, number of floors, and area wealth and unemployment. By including these variables, the authors aim to control for the fact that it is easier to sell in prime locations, and sales volumes may vary with type of product, store configuration, and customer target group. To control for employee characteristics, they include sales assistants' weekly hours (shares of employees working 0–4, 5–14, 15–29, and 30+ h per week), which determine labor flexibility, the relative wage (sales assistants' pay relative to county average) and sales assistants' average age and tenure which helped control for workforce quality. They use a full-time-equivalent for average age and for employee turnover. Finally, they include 20 regional manager dummy variables to control for possible effects of regional management on store productivity.

The authors obtain two sets of regression results for the turnover-performance link: one—negative—for full-timers, who are managed under a commitment work system, and the other—an inverted U shape—for part-timers, managed under a secondary system. For workers managed under a secondary work system, they find

a clear inverted U-shaped relationship between turnover and performance. The initial positive impact of part-time separations on productivity implies that less productive workers are more likely to separate. However, at the inflexion point, the benefits of improved job-worker match and workplace flexibility become offset by dysfunctional turnover and the loss of firm-specific capital. As for employees managed under a commitment system, the link is purely negative. Thus, the costs of core worker turnover appear to be higher than the costs of secondary worker turnover. Finally, they find that the effect of full-time separations is exacerbated by secondary turnover.

The authors calculate that if all the stores operated at the optimum level of full-time turnover instead of at their observed level, the organization would gain 0.3 % of total sales over the period of 1995–1999. Also, the organization can gain up to 1.4 % in productivity by choosing the optimum levels of full- and part-time turnover, which translates into £0.73 per each hour worked per year on average. Finally, summing up individual productivity gains from moving to optimal part-time turnover for all stores and years, they find that overall gain for the organization would be 0.6 % of the total sales over the period of 1995–1999.

Labor Flexibility and Store Performance

Flexible resources, in the form of part-time or temporary workers, create volume flexibility that can affect profitability through either sales or expenses. Individual flexible labor resources may be less productive than full-time workers, as they might have less capability and fewer qualifications than full-time resources. Having too many flexible resources may lead to an increase in co-ordination costs and poor store execution, which could in turn lower sales. On the other hand, flexible resources may increase sales because they provide firms with a dynamic adjustment option—they offer a greater ability to match staffing within a day or within a week.

Flexible resources might help reduce expenses as flexible labor resources are usually paid less than permanent, full-time labor. They can also be retained for fewer working hours than full-time associates, thereby reducing idle labor expenses. On the flip side, having too many flexible resources could lead to an increase in cost due to more frequent hiring, firing, and training costs. Based on the above, Kesavan et al. (2013) test for the following: an inverted U shaped relationship between flexible labor-mix and store performance measures like store sales and store profitability; and a U shaped relationship between flexible labor mix and store expenses.

The authors obtain data from 445 *RetailCo* stores for the period of July 2009 to August 2011. They collect data from three departments—finance, HR, and store operations—and obtain monthly financial statements data for each of the 445 stores. Statements contain stores' revenues and detailed information about expenses (e.g., fleet expenses, administrative expense, inventory shrink, etc.). They also received 30-day monthly forecasts of sales, labor hours, and payroll for each of the stores. HR data provided information on each employee who worked in the store for each

of the 26 months. The information included whether each employee was full-time, part-time, or temporary. The store operations data contained weekly information on the actual hours each employee worked aggregated to monthly level to match the financial data.

The authors measure store performance using sales, expenses, and profits. Monthly store sales are calculated as the total revenue net of returns. Monthly store expenses include all expenses in the store, including labor costs related to salaries and commissions paid, employee related costs connected to relocation and training, occupancy costs resulting from rent and property taxes, administrative expenses related to accidents and insurance, and inventory related costs including insurance shrink, and changes. Labor-related costs account for slightly over half the total expenses in the store.

The store profit represents the before-tax profit for each individual store for that month. It is a function of sales, expenses, and cost of materials. To normalize the performance measures for level of activity to enable comparison across stores and time, the authors divide each of the metrics by average monthly sales for that store. Part-time labor mix is defined as the ratio of part-time to full-time employees and temporary labor mix is defined as the ratio of temporary to full-time employees.

The authors regress the store performance measures (sales, expenses and profits) on the two types of labor mix. They include both linear and square terms of labor mix to capture the non-linear relationship. They also include store fixed effects, region-specific monthly indicator variables to control for seasonality in a year and region-specific time effects to control for seasonal effects that are common to all stores in a given region. The authors also include controls on sales forecast, employee turnover amongst part-time and full-time workers and actual labor hours.

The results show that both part-time and temporary labor-mixes demonstrate an inverted U-shaped relationship with sales. Temporary labor mix has a U-shaped relationship with expenses, while part-time labor mix has a decreasing, concave relationship with expenses. The authors also find an inverted-U shaped relationship between temporary and part-time labor mix and profitability. Based on counterfactual analysis, they show that temporary and part-time workers can increase store sales by 11.5 % over the monthly average during the peak demand period. Further, the volume flexibility offered by these flexible resources can increase profitability by 28 % over the monthly average during the same period.

3.3 Other Relevant Literature

Operations management literature has a long history of studies that use queueing theory-based staffing models to determine service requirements (Hassin and Haviv 2003). These include service settings, such as manned service-desks in retail stores, check-out counters, bank tellers, deli take outs, airport kiosks, theaters, etc., in which people form a queue in front of the counter to wait for service. Using information on arrival rates, service time, and abandonment rates, the models are

used to determine labor requirements to satisfy a service level constraint. In the context of a retail setting, most of the papers model retail stores as an Erlang C (or Erlang A) queue to determine the optimal number of retail workers (Berman and Larson 2004; Berman et al. 2005; Terekhov and Beck 2009). However, empirical research that examines staffing models with retail data is limited, as large-scale traffic data have only recently become available. Exceptions include Lu et al. (2013), Mani et al. (2014) and Tan and Netessine (2014). Lu et al. (2013) use queue data from a deli counter in a supermarket to show that customers focus on the length of the queue without adjusting sufficiently for the speed at which it is served. Mani et al. (2014), as explained earlier, use an empirical model motivated by square-root staffing model to determine the extent of understaffing and overstaffing in retail stores. Tan and Netessine (2014) use operational data from a restaurant chain to show an inverted U-shaped relationship between workload and performance and demonstrate how staffing capacity staffing capacity can be leveraged to optimize workload and increase sales.

As more granular data from traffic counters and other new technology become available, further research on the application of detailed staffing models to retail store operations may become possible.

4 Retail Technologies: Past, Present, and Future

For many decades now, the ubiquitous retail store has been identified with sales associates helping customers in their purchase decisions. Recent consumer and retailer research indicates two trends for the near future: first, that the store will continue to be the channel through which retailers receive the largest proportion of their revenue; and second, that, in general, consumers continue to prefer to shop and buy in the store (Gartner 2013a, b). While the storefront itself is slowly evolving from a simple brick-and-mortar presence to a hub of physical and virtual activity, optimization on store labor is beginning to take center stage in store operations.

In 2003, Gartner predicted that by 2013 retail stores will operate with 20 % less labor because of innovations in in-store retail technologies. Advances in customer-assistance technology (like automated merchandising solutions and self-checkout counters) and work management applications (including integration with real-time information on demand) were predicted to be two drivers of this transformation. At the time of this prediction, a high level of emphasis was placed on operational efficiency through both widespread deployment of point-of-sale terminals and electronic data interchange (EDI) linkages that helped retailers share demand information with supply-chain partners, and adoption of workforce management systems to help plan and schedule labor in store. While labor scheduling tools were not new to retail, they were largely deployed independent of other systems.

By 2007, most point-of-sale (POS) technologies had matured and been widely adopted by many mainstream retailers. Emerging technologies now focused primarily on improving store execution. In this context, end-to-end workforce

management applications were developed that would help retailers balance workload sent to stores, manage the tasks and activities within stores, monitor store compliance, and quickly collect and analyze store feedback. Alongside the development of workforce management solutions, many stores had also begun to experiment with traffic counters—to count customer traffic in stores—and in-store cameras—to prevent theft in their stores, in an effort to make store operations more efficient. Thermal imaging techniques, borrowed from the defense and manufacturing industries were just being commercialized for application in the retail industry. Non-intrusive intelligent sensors could be used to detect customer movement and hot spots in a retail store. The archival data is used to calculate store performance measures like conversion rate and basket values as well as customer service metrics like average queue length and average wait time. These sophisticated traffic-counting technologies were considered to be a technology trigger or breakthrough with huge potential for improving operational efficiency and customer service at the same time. While, over the last few years, quite a few success stories prove the usefulness of traffic counting and workforce management solutions, most retailers still tend to use these real-time systems as stand-alone applications (for example, focusing exclusively on implementing traffic counters or queue management systems). Examples of vendors providing traffic counters for retail applications include Shopper Trak, SMS, and Sensource. More advanced technologies that use GPS to track customers in store, identify behavior of new and repeat customers, and analyze impact of promotional displays in store are also available today from vendors like Euclid Analytics, Goliath Solutions, and Retail Solutions. During this same period, workforce management systems have also further evolved into labor standards systems, scheduling systems, and task management systems.

The advances in store technologies in the last decade have led to an explosion of data available to retailers; many retailers consequently lament that they are "drowning in numbers but have very little actionable insights" (Fisher and Raman 2010). At a typical retailer, real-time data is now available through multiple touch points, such as POS transaction log, customer traffic counters, video over IP (network video), radio frequency identification devices, location-aware applications, and remote monitoring of appliances, including heating, ventilation, and air conditioning. Understandably, there is now a clamor for analytic applications that would help retailers transform this massive data into useful knowledge. In fact, managing and optimizing on big data continually ranks, along with cost containment, among the top 10 priorities for retailers over the next 3–5 years. Industry research on state-of-the art analytical applications shows a similar trend. Gartner's 2013 analytics applications hype cycle classifies real-time store monitoring platforms as being close to the "peak of inflated expectations" phase of the technology life cycle, i.e., technologies that have shown promise in some early adopters and, if successful, can gain widespread adoption in the next 5–10 years.

Real-time store monitoring platforms deliver store activity monitoring on dashboards by bringing together signals and alerts from real-time data sources available in the retail stores. These include inputs from traffic counters, queue management sensors, point-of-sale transaction logs, remote sensors on in-store devices, and RFIDs.

The next challenge lies in combining these real-time feeds onto a single platform so that a store manager can have a comprehensive view of what is happening in their stores. Here, complex algorithms are used to analyze real-time signals, and exception reports or alerts are generated based on user-defined metrics.

One example of real-time store monitoring technology is use of information from infrared sensors above store checkout lanes to calculate average queue lengths and wait times in real time. When the queue length or wait time reach a particular threshold limit, store staff is either reassigned to open additional counters, or more self-checkout lanes are opened to help reduce congestion. At the back end, based on the enormous amount of data collected, the system can also both determine the optimum number of checkouts needed and project traffic congestion and service requirements in future. When combined with a labor management tool, customer wait-time information is linked to labor scheduling to improve labor efficiencies throughout the store. Another example of advancements in real-time store monitoring technology is the use of digital video surveillance systems, coupled with analytics software, to track customer behavior, such as dwell times. Dwell time—the time customers spend in different points in the store—allows retailers to gauge the effectiveness of displays, signage, and promotions. This system allows retailers to get valuable data and insights on every part of the store, from entrances and aisles to customer service, dressing rooms, and even bathrooms. Recent advancements allow retailers to conduct on-going traffic and conversion-rate analysis not only by store but also by aisle and display, on down to the SKU level. They can also use this information to decide which sections (or product categories) in a store might require more sales assistance and allocate labor accordingly. Some vendors in this space are Brikstream, BVI RetailNext, Irisys, Scopix, Retailigence, and Re Tel Technologies.

As with the adoption of any new technology, it is important that retailers consider not only the immediate costs and benefits but also how such a technology will aid various functional roles in their businesses. For example, although traditionally conversion has been the responsibility of the marketing (and merchandising) department, real-time conversion data can aid in judiciously allocating sales associates in the store. This emphasis on integrating demand information with staffing decisions is even more important today when store labor costs run between 10 and 13.5 % of the typical retailer's revenues and are set to rise dramatically as a result of changes in labor supply and the increasing volume of store tasks to be performed in the store (Forrester Research 2009). The biggest worry for many store managers is that tasks are loaded onto the store without visibility about the amount of labor required to execute them. Thus, technologies that help tie labor requirements with store activities and customer demand present retailers with a tremendous potential to streamline labor decisions while maintaining a high level of customer engagement in the stores. These technologies may have the potential to transform the retail store into a data-rich enterprise and thus pave the way for the use of more analytical models and decision support tools to improve store operations.

5 Future Research and Conclusions

Retail labor is an emerging area of research and holds exciting prospects for several reasons. First, the intense competition with online retailers has led many brick and mortar retailers to take a closer look at the in-store experience offered to their customers and find ways to distinguish themselves based on customer service. While retailers have traditionally cared about retail labor because of its huge impact on sales and expenses, surprisingly we find the penetration of analytical techniques for workforce management to be limited in the retail industry. It is unclear why retailers who invest millions of dollars to drive traffic into the stores through marketing activities would not invest sufficiently in labor planning to ensure that the incoming traffic is converted to sales. However, this situation is likely to change. We observe many new start-ups in the area of traffic counting and in-store technology that offer retailers new, hitherto unavailable, data. These data present an excellent opportunity for retailers to transform their store operations to enhance productivity and compete effectively with online retailers.

Second, the availability of new data could make it possible to answer questions that were difficult to do so earlier. For example, the availability of store traffic data now allows researchers to examine the impact of labor on conversion rate (Perdikaki et al. 2012), a metric that has been long been tracked in other settings such as online retailers. Other performance metrics, such as dwell time, the amount of time spent by customers in the store, and frequency of customer visits, could be examined in future research. In addition, the integration of online and brick-and-mortar operations raises new and interesting questions around the role of store labor and the design of its incentives. Prior research on store manager behavior has shown that change in incentives can have significant impact on store performance (DeHoratius and Raman 2007).

Third, enormous scope exists for applying analytical techniques to improve labor planning. A large body of research has addressed the labor planning issues in manufacturing, but such research is absent in retail. As explained in Sect. 3.1, the presence of significant differences between manufacturing and retailing necessitate studying retail labor as an independent problem. Fisher (2004) state that a retail store is an amalgam of a factory and sales office, so labor planning solutions in retail can potentially build on prior research in manufacturing but would need to additionally account for the differences that arise due to co-production. Another area of research would be to examine how to apply queueing theory to the retail setting. While queueing theory holds large prospects to improve retail store operations, the complexity of retail store operations offers new opportunities for extensions. For example, several aspects—customers being able to complete most activities without the help of sales associate; associates being able to multi-task by dealing with none, one, or multiple customers simultaneously; and associates performing different types of activities such as stocking, cashiering, helping customers, etc.—need to be accounted for appropriately before applying queueing theory to retail stores.

References

Anderson, E. G. (2001). The nonstationary staff-planning problem with business cycle and learning effects. *Management Science, 47*(6), 817–832.

Bard, J. F., & Wan, L. (2008). Workforce design with movement restrictions between workstation groups. *Manufacturing and Service Operations Management, 10*(1), 24–42.

Bechtold, S. E., & Jacobs, L. W. (1990). Implicit modeling of flexible break assignments in optimal shift scheduling. *Management Science, 36*(11), 1339–1351.

Berman, O., & Larson, R. C. (2004). A queueing control model for retail services having back room operations and cross trained workers. *Computers and Operations Research, 31* (2), 201–222.

Berman, O., Wiang, J., & Sapna, K. P. (2005). Optimal management of cross-trained workers in services with negligible switching costs. *European Journal of Operational Research, 167*(2), 349–369.

BLS. (2008). *Involuntary part-time work on the rise. Issues in labor statistics*. Washington, DC: U.S. Bureau of Labor Statistics.

Brownell, W. S., & Lowerre, J. M. (1976). Scheduling of work forces required in continuous operations under alternative labor policies. *Management Science, 22*(5), 597–605.

Daniels, R. L., Mazzolla, J. B., & Shi, D. (2004). Flow shop scheduling with partial resource flexibility. *Management Science, 50*(5), 658–669.

Dantzig, G. (1954). A comment on Edie's traffic delay at toll booths. *Operations Research, 2*(3), 339–341.

DeHoratius, N., & Raman, A. (2007). Store manager incentive design and retail performance: An exploratory investigation. *Manufacturing and Service Operations Management, 9*(4), 518–534.

Dill, W. R., Gawer, D. P., & Weber, W. L. (1966). Models and modeling for manpower planning. *Management Science, 13*(4), 142–167.

Fisher, M. (2004). To you it's a store. To me it's a factory. *ECR Journal, 4*(2), 9–18.

Fisher, M. L., & Krishnan, J. (2005). *Store level execution at Wawa. Case study*. The Wharton School, University of Pennsylvania, Philadelphia.

Fisher, M. L., Krishnan, J., & Netessine, S. (2006). Retail store execution. *Working paper*, University of Pennsylvania, Philadelphia, PA.

Fisher, M. L., & Raman, A. (2010). *The new science of retailing*. Boston, MA: Harvard Business School Press.

Forrester Research. (2009). *Filling the store labor productivity gap*. Cambridge, MA: Forrester Research.

Fryer, J. S. (1974). Labor flexibility in multi-echelon dual-constraint job shops. *Management Science, 20*, 1073–1080.

Gartner. (2003). *Emerging trends and technologies in the retail industry. IR COM-21—358*. Stamford, CT: Gartner.

Gartner. (2013a). *Agenda overview for retail, IR G00246684*. Stamford, CT: Gartner.

Gartner. (2013b). *Hype cycle for analytic applications, IR G00251137*. Stamford, CT: Gartner.

Hassin, R., & Haviv, M. (2003). *To queue or not to queue: Equilibrium behavior in queueing systems*. Norwell, MA: Kluwer.

Holt, C. C., Modigliani, F., & Simon, H. A. (1956). Linear decision rule for production and employment scheduling. *Management Science, 2*(2), 159–177.

Hopp, W. J., Tekin, E., & VanOyen, M. P. (2004). Benefits of skill chaining in production lines with cross-trained workers. *Management Science, 50*(1), 83–98.

Karmarkar, U., & Pitbladdo, R. (1995). Service markets and competition. *Journal of Operations Management, 12*(3), 397–411.

Kesavan, S., Staats, B. R., & Gilland, W. (2013). Volume flexibility in services: The costs and benefits of flexible labor resources. *Working paper*. University of North Carolina.

Kunreuther, H. C., & Morton, T. E. (1974). General planning horizons for production smoothing with deterministic demands. *Management Science, 20*(7), 1037–1046.

Lapré, M. A., & Nembhard, I. M. (2011). *Inside the organizational learning curve: Understanding the organizational learning process*. Norwell, MA: Now Publishers Inc.

Lippman, S. A., Rolfe, A. J., Wagner, H. M., & Yuan, J. S. C. (1967). Optimal production scheduling and employment smoothing with deterministic demands. *Management Science, 14*(3), 127–158.

Lockard, C. B., Wolf, M. (2012). Monthly labor review. *135* (1), 84–108.

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

Mani, V., et al. (2014). Estimating the impact of understaffing on sales and profitability in retail stores. *Production and Operations Management*. doi:10.1111/poms.12237.

Maxham, J. G., III, Netemeyer, R. G., & Lichenstein, D. R. (2008). The retail value chain: Linking employee perceptions to employee performance, customer evaluations, and store performance. *Marketing Science, 27*(2), 147–167.

Morris, J., & Showalter, M. (1983). Simple approaches to shift, days-off and tour scheduling problems. *Management Science, 29*(8), 942–950.

Mui, Y. Q. (2007). Circuit city cuts 3,400 'Overpaid' workers. Retrieved Dec 1, 2013, from http://www.washingtonpost.com/wp-dyn/content/article/2007/03/28/AR2007032802185.html

Netessine, S., Fisher, M. L., & Krishnan, J. (2010). Labor planning, execution, and retail store performance: An exploratory investigation. *Working paper*. The Wharton School, University of Pennsylvania.

Parasuraman, A., Zeithaml, V., & Berry, L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing, 64*(1), 12–40.

Perdikaki, O., Kesavan, S., & Swaminathan, J. M. (2012). Effect of retail store traffic on conversion rate and sales. *Manufacturing and Service Operations Management, 5*(2), 79–141.

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

Siebert, W. S., & Zubanov, N. (2009). Searching for the optimal level of employee turnover: A study of a large U.K. retail organization. *Academy of Management Journal, 52*(2), 294–313.

Tan, T. F., & Netessine, S. (2014). When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Science, 60*(6), 1574–1593. http://dx.doi.org/10.1287/mnsc.2014.1950.

Terekhov, D., & Beck, J. C. (2009). An extended queueing control model for facilities with front room and back room operations and mixed skilled workers. *European Journal of Operational Research, 198*(1), 223–231.

Thompson, G. M. (1995). Improved implicit optimal modeling of the labor shift scheduling problem. *Management Science, 41*(4), 595–607.

Ton, Z. (2009). The effect of labor on profitability: The role of quality. *Working paper*, Harvard University, Boston.

Ton, Z., & Huckman, R. S. (2008). Managing the impact of employee turnover on performance: The role of process conformance. *Organization Science, 19*(1), 56–68.

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

Zeithaml, V. A., Berry, L. L., & Parasuraman, A. (1996). The behavioral consequences of service quality. *Journal of Marketing, 60*(2), 31–46.

Zeithaml, V. A., Parasuraman, A., & Berry, L. L. (1990). *Delivering quality service: Balancing customer perceptions and expectations*. New York, NY: The Free Press.

Chapter 7
Category Captainship Practices in the Retail Industry

Mümin Kurtuluş and L. Beril Toktay

1 Introduction

A product category is defined as a group of products that consumers perceive to be interrelated and/or substitutable (Nielsen Marketing Research 1992). Soft drinks, baking products, and canned vegetables are some examples of retail categories. Categories can be viewed as the smallest strategic business unit within a retailer. Retailers implementing category management focus their efforts on managing the entire product category as a single business unit and maximize category profit as opposed to managing each product individually (i.e., either on a brand-by-brand or SKU-by-SKU basis). Category management emphasizes the management of product categories as a whole and allows the retailers to capture the synergies that may arise as a result of grouping the products together. Taking a holistic approach and focusing on category performance allows the retailers to capture synergies such as promotion coordination and store traffic driving strategies. Category management involves decisions such as merchandizing the product assortment, determining retail prices, and allocating shelf-space to each product on the basis of category goals. The category management approach requires the retailers to dedicate significant amount of resources to understanding the consumer trends and consumers' response to the assortment, pricing and shelf placement decisions of products within a category.

M. Kurtuluş (✉)
Owen Graduate School of Management, Vanderbilt University, 401 21st Avenue South, Nashville, TN 37203, USA
e-mail: mumin.kurtulus@owen.vanderbilt.edu

L.B. Toktay
College of Management, Georgia Institute of Technology, 800 West Peachtree Street NW, Atlanta, GA 30308-0520, USA

N. Agrawal, S.A. Smith (eds.), *Retail Supply Chain Management*,
International Series in Operations Research & Management Science 223,
DOI 10.1007/978-1-4899-7562-1_7

Prior research in marketing (e.g., Basuroy et al. 2001; Dhar et al. 2001; Gruen and Shah 2000) has shown that category management can result in significant benefits for the retailers.

Recently, many retailers have started to rely on their manufacturers for strategic recommendations and insights regarding category management decisions, a practice often referred to as *category captainship*. This approach has now become a common way to execute category management in certain product categories for many retailers. The increase in the number of product categories offered at retailers, combined with the scarcity of retailer resources required to manage each category effectively are some of the drivers of the widespread use of category captainship practices. Other factors are manufacturers' deep expertise in their own categories based on the market research they conduct for introducing new products and improving their existing products. The category captainship approach acknowledges that manufacturers can help retailers manage categories more effectively, and at a lower cost, by leveraging their existing consumer insights (Kurtuluş et al 2013). Even though the captains are not directly compensated for their services, the manufacturers view captainship as a source of competitive advantage over their competitors because the captain usually gains significant control over the key category management decisions (Kurtuluş and Toktay 2004).

In a typical captainship implementation, the retailer first selects a captain by soliciting proposals from the largest manufacturers in the category. The retailer selects the manufacturer that promises the largest improvement in category performance to serve as the captain. After the captain is selected, the retailer and the captain summarize the objectives for the captain and develop metrics to track the captain's performance (ACNielsen 2005; Kurtuluş et al. 2013). The performance metrics typically include measures such as target category profit and/or sales. The category captain then provides the retailer with a plan that includes recommendations about key category management decisions such as which brands to include or exclude from the category, how to display the products, how much space to allocate to each brand, and in some cases how to price the products in the category. The retailer is free to accept or reject any of the recommendations provided by the captain. The captain's performance is evaluated regularly based on the agreed metrics. If the captain's performance is unsatisfactory, the retailer might decide to assign the captainship role to another manufacturer. Retailers usually design the category captainship agreements to be short term (e.g., 1–2 years) in order to keep the flexibility to renegotiate the agreements or rotate the captainship position among different manufacturers.

Many retailers and manufacturers practice category captainship and report positive benefits. Retailers such as Wal-Mart, Metro, Safeway, and Kroger practice category captainship in some of their product categories and usually assign manufacturers such as Kraft Foods, P&G, Kellogg and Danone to serve as category captains because of their established brands in the market and their resource availability (Kurtuluş and Toktay 2004; Subramanian et al. 2010; Kurtuluş and Nakkas 2011; Kurtuluş et al. 2013; Progressive Grocer 2007, 2008). Below are some specific examples of category captainship implementations from practice.

Example 1: Carrefour, the second largest retailer in the world, has asked Colgate to serve as category captain and provide insights to improve the performance of the oral care category. Based on a number of consumer studies, Colgate suggested that Carrefour restructure the display in the oral care category so as to merchandise toothbrush products above toothpaste products, as opposed to merchandising them next to each other. As a result of the restructuring, Carrefour reported 6–16 % sales increase in the oral care categories in its retail markets (ECR Conference 2004). The sales increase in the oral care category came at a little cost to the entire channel because Colgate mostly utilized its already existing consumer studies and its expertise in the oral care category. If Carrefour had conducted the research necessary for such a restructuring, it would have been more expensive.

Example 2: Ross Products serves as category captain for Safeway in the infant formula category (Progressive Grocer 2004). Safeway asked Ross Products to examine the category and prescribe solutions to improve the profitability of the category. Ross' assessment of the category revealed that the category was undermerchandised: the infant formula subcategory was contributing 34 % of the baby care category's dollar volume, but was receiving only 11 % of the shelf-space. Ross recommended changes in shelf-space positioning, and also reviewed and revised the pricing to boost profitability. After implementing the recommendations, the category saw a 9.2 % sales growth benefiting both Safeway and Ross Products (Progressive Grocer 2004). One could argue that Safeway could have developed a similar prescription to improve the performance in the infant formula category without using Ross Products as a category captain, however, the cost of doing so would have been much higher as Safeway does not possess the expertise that Ross Products does.

Example 3: General Mills served as category captain for some of its retail partners in the Baking Ingredients and Mixes category (Progressive Grocer 2004, 2010). General Mills' recommendations are focused around SKU rationalization and variety-vs-duplication analysis. SKU rationalization is aimed at reducing the number of SKUs to reduce consumer confusion at the shelf and thus create growth. Similarly, excessive duplication does not add much in incremental volume. Removing duplications allows for expanded product variety, which in turn can generate more sales in the category and help it grow. One of the retailers for which General Mills serves as category captain has seen a 10.2 % increase in base dollar volume since General Mills' SKU rationalization efforts (Progressive Grocer 2004).

Although category captains are more common in the grocery and consumer products industries, category captainship practices are making an appearance in apparel retailing as well. VF Corp., the NC based manufacturer of brands such as Lee and Wrangler, serves as category captain for a number of its retail partners in the jeans category (Apparel Magazine 2005). VF Corp works with its retail partners to determine the product mix to be offered in each region, how products will be displayed on the sales floor, and how inventory levels will be managed in the category. Inspired by the success in the jeans category, VF Corp is looking forward to take on category captainship responsibility in other categories such as sports licensing and outdoor performance apparel categories.

The above examples illustrate that the scope of the recommendations in each category captainship implementation is different: While some retailers rely on their category captains for shelf and display management, others rely on their captain for assortment related decisions. In addition, category captainship practices vary in terms of the extent to which the retailer implements captain's recommendations, resulting in a continuum of practices. At one end of the spectrum, some retailers implement the category captain's recommendations as they are; at the other end, some retailers filter the recommendations provided by their captain and verify their appropriateness before implementing the recommendations (Steiner 2001).

The above examples, and many other successful category captainship implementations, demonstrate that by working together, retailers can considerably benefit from their manufacturers' expertise in managing their categories and deliver consumer value through supply chain collaboration. However, category captainship practices have also been controversial because the captains provide recommendations to the retailer regarding not only their own products, but their competitors' products too. In addition, conflict of interest between the retailer and the captain are inevitable because what is in the best interest of the category captain may not be the best for the retailer (Kurtuluş and Toktay 2004).

One of the key concerns with category captainship practices has been the captain's potential bias against their competitors' products (Steiner 2001; Desrochers et al. 2003; Greenberger 2003; Leary 2003; Klein and Wright 2006).[1] In this context, it is not surprising that there is an ongoing debate on whether or not category captainship is anti-competitive. The main concern expressed by anti-trust researchers has been that captainship practices might have negative impact on both the non-captain manufacturers and consumers. This is because the use of captains may result in lower variety and higher prices in the category, which may harm the consumers and exclude some of the manufacturers from the category. The term *competitive exclusion* has often been used to refer to situations where the captain takes advantage of its position and disadvantages the competitors' products in the category. Although competitive exclusion is a possible negative consequence of implementing captainship, it can be difficult to prove/detect because it can occur in many different forms.

Anti-trust researchers (e.g., Desrochers et al. 2003; Leary 2003; Klein and Wright 2006) and marketing researchers (Morgan et al. 2007; Gooner et al. 2011)

[1] While there are many cases under investigation due to claims of category captainship misconduct, one publicly known and well-documented case is the United States Tobacco Co. vs. Conwood Co. case. United States Tobacco Co. (UST), the biggest company in the smokeless-tobacco category, was recently ordered to pay a $1.05 billion antitrust award to Conwood, the second biggest competitor in the category (Greenberger 2003). Conwood had sued UST, the category captain, and had claimed that UST used its position as category captain to exclude competition and provide an advantage to its own brands. The court ruled that UST's practices resulted in unlawful monopolization, harming competition, and consequently, the consumers. Similarly, many other captainship arrangements in the tortillas, cranberries, and carbonated soft drinks categories are being investigated for potential category captainship misconduct (Desrochers et al. 2003).

have defined the competitive exclusion phenomenon broadly as the captain behaving opportunistically to favor its own product over competitors' products. Existing research on captainship has also defined some specific forms of exclusion. For example, Kurtuluş and Toktay (2011) point to the possibility of exclusion via a smaller shelf-space allocation to the non-captain manufacturers' products whereas Kurtuluş and Nakkas (2011) point out the possibility of exclusion via reduction in the number of products offered by the non-captain manufacturers after captainship is implemented.

To summarize, while many retailers and manufacturers claim positive benefits from implementing category captainship, there is also evidence regarding category captainship misconduct. Retailers planning to implement category captainship should develop an understanding of the pros and cons of such practices and should weigh potential advantages and disadvantages of using category captains for category management. The goal of this chapter is to provide an overview of the existing research on category captainship, and identify research directions that would improve our understanding of its impact.

The chapter is organized as follows. We start by reviewing the literature on category captainship in Sect. 2. In Sect. 3, we discuss the potential impact of category captainship practices on the retailing industry. Section 4 offers some future research directions.

2 Review of Existing Research on Category Captainship

Although category captainship practices have been very popular over the last decade, there is very little academic research regarding the category captainship practice and its consequences. The existing research on captainship can be grouped into four broad categories that aim to answer the following questions:

- What are the consequences of the retailer delegating the pricing decision to a category captain?
- What are the consequences of the retailer delegating the assortment selection decision to a category captain?
- When will category captainship emerge? What are the category characteristics that facilitate the emergence of category captainship?
- What are the antitrust concerns that may arise as a result of using category captains for category management? What can be done to mitigate these antitrust concerns?

The limited research about captainship is due to challenges such as the broad scope of captainship implementations and continuum of category captainship implementations. In general, the retailers rely on a category captain for recommendations about retail category management decisions such as pricing, assortment, shelf-space management, promotions, etc. However, researchers usually focus on recommendations in only one of these areas, limiting their research and findings to a

subset of captainship implementations. In addition, while some retailers implement their category captain's recommendations as they are, others use them only after modifying the recommendations. Researchers usually focus on one end of this spectrum where the retailer implements the recommendations as they are and ignore all other possibilities. In Sect. 4, we propose some avenues for future research that could potentially overcome these challenges and improve our understanding of category captainship practices. In what follows, we review the existing research on captainship by emphasizing the research questions addressed and the methodology used, and we describe how each paper contributes to a better understanding of captainship practices.

2.1 Consequences of Delegating the Pricing Decisions

The idea of an upstream party in a supply chain (such as a manufacturer) interfering with the retailer's pricing decisions is not new. There is a large amount of research in economics on so-called Resale Price Maintenance (RPM) practices where a manufacturer imposes a minimum or a maximum resale price on the retailers (e.g., Gilligian 1986; Overstreet 1983 and references therein). Research on RPM has mainly focused on offering explanations that shed light on the use of RPM practices. The most intuitive explanation is that manufacturers would use RPM and would limit retailers' flexibility in setting their retail prices optimally because there would be too much price competition between the retailers otherwise.

However, there are other alternative explanations. The traditional view has been that RPM can be used to prevent retailers from "free-riding" in providing services (Telser 1960). While one retailer may offer a service in how to use the product, another retailer might benefit or free ride by selling to a customer who has already learned about how to use the product from the other retailer. A more recent explanation offered by Deneckere et al. (1996) is that RPM can be used to respond optimally to demand uncertainty and to encourage retailers to hold inventories. Nevertheless, the literature remains inconclusive regarding the impact of RPM practices on consumer welfare; while some research indicates that RPM practices enhance consumer welfare, other work indicates the opposite (Ippolito and Overstreet 1996).

While the RPM and category captainship practices are similar in the sense that the manufacturer interferes with retailer's pricing decisions, there are significant differences between the two. RPM practices are manufacturer driven, while category captainship practices are usually driven by the retailers. In addition, while with RPM, the manufacturer imposes a retail price on its own products only, in category captainship, the manufacturer might recommend retail prices (and may interfere with prices) for all products in the category. In order to investigate the impact on stakeholders and consumer welfare, the RPM literature generally utilizes models where a single manufacturer sells to consumers through multiple competing retailers (e.g., Chen 1999; Deneckere et al. 1996). On the other hand, the category

captainship literature generally utilizes models where multiple manufacturers sell their products to the consumers through a common retailer (e.g., Wang et al. 2003; Subramanian et al. 2010; Kurtuluş and Toktay 2011; Kurtuluş and Nakkas 2011). To summarize, while RPM practices and category captainship practices differ significantly, the main research questions are similar: Both streams of research aim at providing justification for use of these practices by investigating the impact on involved parties and consumer welfare.

The two papers that focus on category captainship implementations where a retailer relies on a category captain for pricing decisions are Wang et al. (2003) and Kurtuluş and Toktay (2011). Both of these papers consider how each stakeholder in the supply chain is affected when the retailer delegates the pricing decisions to one of its leading manufacturers. Below we review both papers in detail.

Kurtuluş and Toktay (2011) consider a distribution channel where two manufacturers sell their products to consumers through a common shelf-space constrained retailer. The authors use a linear price-dependent demand model (Shubik and Levitan 1980) where consumer demand is given by

$$q_1 = a_1 - p_1 + \theta(p_2 - p_1) \quad q_2 = a_2 - p_2 + \theta(p_1 - p_2)$$

where p_1 and p_2 are the retail prices of the two products, and a_1 and a_2 can be interpreted as the relative *brand strength* of each product. For simplicity, the paper assumes that the manufacturers are symmetric, (i.e., $a_1 = a_2 = a$). The parameter $\theta \in [0, 1]$ is the cross-price sensitivity. As θ increases, the demand for product i, q_i, becomes more sensitive to competitor's price, p_j. The parameter θ can also be interpreted as the *degree of product differentiation* with $\theta = 0$ implying perfectly differentiated products and $\theta = 1$ implying substitutable products.[2]

Since retailers operate on very thin margins, every unit of shelf-space is scrutinized for profitability and allocating the total store space between categories has become a critical decision for retailers today. The authors capture the shelf-space allocation decision by assuming that the retailer determines the shelf-space for the category, which is denoted by S, based on the opportunity cost of the shelf-space, kS^2. This is consistent with current practice where retailers typically allocate category shelf-space based on the profitability of each category relative to the other categories (Corstjens and Doyle 1983; Chen et al. 1999) because space allocated to one category means profits foregone from another.

Once the retailer decides on the category shelf-space S, the pricing decisions are made subject to the constraint $q_1 + q_2 \leq S$ where q_1 and q_2 can be interpreted as demand rates for each product per replenishment period. In other words, the retailer prices the products so that the total demand rate does not exceed the shelf-space

[2] This type of linear demand system has been widely used in marketing (e.g., McGuire and Staelin 1983; Choi 1991) and economics (e.g., Vives 1999). These demand functions can be derived from an underlying consumer utility model where consumers maximize their utility.

availability. The quantities q_1 and q_2 can also be interpreted as the long-term volumes to be purchased and sold subject to a total volume target for the category.

The paper considers two scenarios that differ in who determines the retail prices. In the first scenario, retailer category management (RCM), the retailer first decides on the category shelf-space and announces this category shelf-space to the manufacturers. The manufacturers then simultaneously set their wholesale prices. Finally, given the wholesale prices, the retailer sets the retail prices for both products.

The model is solved by backward induction: In the third stage of the game, the retailer solves the following problem for given category shelf-space S and wholesale prices w_1 and w_2:

$$\begin{aligned} \max_{p_1,p_2} \ & (p_1 - w_1)q_1 + (p_2 - w_2)q_2 \\ \text{s.t.} \ & q_1 + q_2 \le S \\ & q_1 \ge 0, q_2 \ge 0 \end{aligned}$$

The authors fully characterize the quantity responses $\hat{q}_1(w_1, w_2)$ and $\hat{q}_2(w_1, w_2)$. Then at stage two, anticipating the retailer's demand responses, the manufacturers simultaneously set their wholesale prices. Each manufacturer maximizes

$$\Pi_i(w_i, w_j) = (w_i - c)\hat{q}_i(w_i, w_j) \quad \text{for } i,j = 1,2 \text{ and } i \neq j,$$

where c is manufacturer i's production cost. Finally, in the first stage of the game, the retailer determines the category shelf-space taking into account the sub-game starting in stage two, and the opportunity cost of shelf-space allocation, kS^2. Since manufacturers are symmetric, both manufacturers are allocated equal shelf-space in the RCM model.

In the second scenario, category captainship (CC), the retailer assigns one of the manufacturers as the captain and delegates the pricing decisions to that manufacturer. The paper models captainship by assuming that the retailer and the captain form an alliance. In making the category shelf-space decision, the retailer assumes that he will get a fraction ϕ of the alliance profit. The value of ϕ is either set at the beginning of the category captainship agreement, or it is the fraction of profits the retailer expects to obtain in *ex-post* negotiation with the captain. The sequence of events in the captainship model is as follows: (1) the retailer determines the amount of category shelf-space S and announces it; (2) the second manufacturer offers a wholesale price w_2 for its product to the alliance; (3) the captain sets the retail prices for both products to maximize the alliance profit subject to the shelf-space constraint.

Similar to the RCM model, the CC model is also solved by backward induction: In the third stage, the captain sets retail prices for both products to maximize the alliance profit for a given wholesale price w_2 and subject to the category shelf-space constraint S. The captain solves the following optimization problem:

$$\begin{aligned} \max_{p_1,p_2} \; & (p_1 - c)q_1 + (p_2 - w_2)q_2 \\ \text{s.t.} \quad & q_1 + q_2 \leq S \\ & q_1 \geq 0, q_2 \geq 0 \end{aligned}$$

The authors characterize the quantity responses $\hat{q}_1(w_2)$ and $\hat{q}_2(w_2)$ for all possible w_2. Then, the non-captain manufacturer sets the wholesale price w_2 in expectation of $\hat{q}_2(w_2)$ by maximizing its profit $(w_2 - c)\hat{q}_2(w_2)$. Finally, in the first stage, the retailer determines the category shelf-space based on its expected share ϕ of the profits in the sub-game starting in stage two, and the opportunity cost of shelf-space, kS^2. Even though the manufacturers are symmetric in terms of demand and cost parameters, in the captainship model the captain is allocated three quarters of the category shelf-space and the non-captain manufacturer is allocated only one quarter of the category shelf-space.

Kurtuluş and Toktay (2011) investigate the impact of switching from retailer category management (RCM) to category captainship (CC) on the category shelf-space and the profits of each party. The key-driving factor is the profitability of the category net of opportunity costs. The authors find that the switch to captainship can increase the profitability of the category for the retailer through the formation of the alliance via two effects: the elimination of double marginalization and the increased price pressure on the non-captain manufacturer. The authors find that the equilibrium category shelf-space under captainship may be higher if the retailer appropriates a significant share of the alliance profit.

The authors conclude that captainship practices should not immediately raise anti-trust concerns, or be viewed negatively by non-captain manufacturers as the resulting increase in the relative profitability of the category vis-a-vis the retailer's other categories can create value for non-captain manufacturers via an increase in the category shelf-space. In particular, the authors find that captainship does not result in competitive exclusion when the products are well differentiated and the retailer's share of alliance profits is high enough. With differentiated products, the gain from avoiding double marginalization and from the drop in the non-captain manufacturer's wholesale price is higher. Coupled with obtaining a high share of the alliance profit, these effects result in a large enough allocation to the category by the retailer that it offsets the non-captain's loss resulting from a smaller fraction of shelf-space allocation under captainship.

At the same time, the paper also provides support for competitive exclusion and shows that the non-captain manufacturers could be at a disadvantage when captainship is implemented in categories where either the products offered in a category are similar (i.e., substitutable) and/or the retailer is not powerful enough compared to the captain.

Similar to Kurtuluş and Toktay (2011), Wang et al. (2003) also consider the impact of captainship where the retailer relies on a captain for pricing decisions. Wang et al. (2003) consider a model with N manufacturers that sell their products through a retailer and investigate whether it is profitable for the retailer to delegate

pricing authority to a captain. The demand for product i in the model considered by Wang et al. (2003) is given by

$$q_i = \frac{1}{N}\left[a - p_i + \frac{1}{N-1}\sum_{i\neq j}^{N}\theta(p_j - p_i)\right]$$

where parameter a can be interpreted as the base level of category demand and parameter θ is the cross-price sensitivity.

In the absence of a category captain, the manufacturers act as Stackelberg leaders and offer wholesale prices $(w_1, w_2, \ldots, w_N)$ to the retailer at stage one of the game. Then at stage two, given the wholesale prices, the retailer sets the retail prices to maximize total category profit

$$\max_{p_1,\cdots,p_N} \sum_{i=1}^{N}(p_i - w_1)q_i.$$

The game is solved through backward induction. First, the retailer solves the above optimization problem for given wholesale prices and determines the quantity responses and then each manufacturer sets its own wholesale price in expectation of the quantity demanded of its own product, $\hat{q}_i(w_1, \ldots, w_N)$, to maximize profit. The production costs are assumed to be zero for all the products. At stage one of the game, each manufacturer solves

$$\max_{w_i} w_i\hat{q}_i(w_1, \ldots, w_N).$$

In the category captainship model, the authors assume that the manufacturer with index one (the first manufacturer) is assigned as the captain. Category captainship is modeled as an alliance between the retailer and the manufacturer of the first brand. In other words, under category captainship, the retailer and the category captain act as an integrated firm. In this model, after the $N-1$ manufacturers offer their wholesale prices $(w_2, w_3, \ldots, w_N)$, the alliance (where the captain and the retailer act as an integrated firm) sets the retail prices to maximize the alliance profit

$$\max_{p_1,\cdots,p_N} p_1q_1 + \sum_{i=2}^{N}(p_i - w_i)q_i.$$

Then, given the quantity responses $\hat{q}_i(w_2, \ldots, w_N)$, $i \geq 2$, the manufacturers set their wholesale prices.

The main result in Wang et al. is that using a category captain for category management is profitable for both the retailer and the category captain. The intuition is as follows: After the retailer and the category captain form an alliance, the alliance will gain from the category captain's brand (i.e., coordination between the retailer and the captain) and will lose from selling other brands in the category. It turns out that both the channel coordination effect and the competition effect have

a positive impact on the joint profit gain, therefore benefiting both the retailer and the category captain. On the other hand, category captainship generally does not benefit the non-captain manufacturers due to increased pressure from the channel. Furthermore, the paper identifies conditions under which category captainship can benefit all participating partners. Category captainship may benefit all parties in the supply chain if (1) the captain has the authority to choose the retail price for its own brand only (i.e., partial delegation); and (2) the non-captain manufacturer behaves strategically (i.e., adjusts its own wholesale price to the use of a captain in the supply chain).

In addition, the paper identifies conditions under which category captainship is more beneficial for the alliance members. The paper finds that the profitability of using a category captain is higher if the product category (1) has fewer products (lower N); (2) has higher price competition among products (higher cross-price sensitivity θ) and (3) has no store brand as opposed to having a store brand. The inclusion of a store brand modifies the demand system slightly and therefore the alliance profit. When there is a store brand, the alliance sets the retail prices to maximize the alliance profit

$$\max_{p_1,\ldots,p_N} p_1 q_1 + \sum_{i=2}^{N} (p_i - w_i) q_i + p_s q_s$$

where q_s and p_s are the demand and price for the store brand and q_i and q_s are given by

$$q_i = \frac{1}{N-1}\left[a - p_i + \frac{1}{N}\sum_{i \neq j}^{N} \theta(p_j - p_i) + \delta(p_s - p_i)\right]$$

$$q_s = \frac{1}{N-1}\left[a - p_s + \frac{1}{N}\sum_{j}^{N} \delta(p_j - p_s)\right]$$

The parameter δ in the above equations is the cross-price sensitivity between the manufacturers' brands and the store brand.

The model also offers some insights as to which manufacturer should be selected as a category captain. The ideal category captain is the manufacturer who has a higher brand strength (i.e., higher a) and a higher cross-price sensitivity. This finding is in line with the current practice where retailers assign their leading manufacturers as category captains.

To summarize, the contribution of both Wang et al. (2003) and Kurtuluş and Toktay (2011) is in pointing out that category captainship can be beneficial for not only the retailer and the captain but also for the non-captain manufacturer(s).

2.2 Consequences of Delegating the Assortment Selection Decision

In both Wang et al. (2003) and Kurtuluş and Toktay (2011), the retailer delegates the pricing authority to a leading manufacturer. However, in practice, the scope of category captainship is broader than making price recommendations. Retailers might rely on their category captains for assortment recommendations as well. Kurtuluş and Nakkas (2011) consider a model where the retailer delegates the assortment selection decision in the category to a leading manufacturer. The goal of this research is to study how the assortment offered to the consumers at the retailers will change if the captain is given an authority over the assortment decisions.

The existing literature on assortment planning in operations has mainly focused on assortment planning by the retailer (i.e., centralized assortment planning) (see Kok et al. (2008) for a review). While a number of papers consider assortment planning in the context of decentralized distribution channels (i.e., Villas-Boas 1998; Aydin and Hausman 2009),[3] Kurtuluş and Nakkas (2011) is the first paper that considers how captainship practices play a role on the assortment offered at a retailer.

Kurtuluş and Nakkas (2011) consider a two-stage supply chain with multiple manufacturers (where each manufacturer offers one product only) sell their products to the consumers through a retailer. A customer either purchases one of the products offered at the retailer or does not purchase anything. The paper uses a generic attraction market share type model (Bell et al. 1975; Gruca and Sudharshan 1991) to model demand for each product in the category. The multinomial logit (MNL), which has been extensively used in the operations literature to study assortment problems (e.g., van Ryzin and Mahajan 1999; Cachon and Kok 2007; Cachon et al. 2008), is one example of an attraction type market share model. Let A_i be the attraction of product $i = 1,2,\ldots, N$. For tractability, the paper focuses on a case where all products are equally attractive, that is $A_i = A$ for $i = 1,2,\ldots, N$. A_0 represent the attractiveness of the no-purchase option and A_0 is normalized to one. Given these assumptions, if the retailer decides to offer n products, the market share (or the purchase probability) for each product is given by

$$q(n) = \frac{A}{1 + nA}$$

[3] Villas-Boas (1998) considers a manufacturer's product line design in a setting where products are sold through an intermediary (i.e., retailer) and the intermediary does the ultimate targeting of products. Aydin and Hausman (2009) study the use of slotting fees by a manufacturer to coordinate the retailer's assortment decision in a setting where the manufacturer sells multiple products through a single retailer.

Let also λ denote the total category traffic. Thus, the average demand rate for each product is given by $\lambda q(n)$.

The paper assumes that all products have the same wholesale price w, retail prices p, and production costs are normalized to zero. The retailer's net profit margin is defined as $m = p - w$. In this setting, because all products have the same probability of being purchased by a consumer and the wholesale price is the same for all products, it is optimal for the retailer to choose the same price for all products (Shugan 1989; Cachon et al. 2008). Hence, the retailer adopts a constant margin policy. In addition, the authors assume that the retailer incurs an operational cost (e.g., cost of managing and executing the replenishment for each product), which is linear in the variety offered in the category, βn, with $\beta > 0$ (Honhon and Pan 2013).

The paper models category captainship by assuming that the category captain has better information about the consumers' preferences. This is in line with the main motivation of the retailers for using category captains. The authors capture the information asymmetry through the attraction parameter A in the demand model: While the retailer believes that the attraction parameter A is either high (A_H) or low (A_L) with probabilities α and $1-\alpha$, respectively, the captain knows the realization of A.

First, the paper considers a model where the retailer decides how many products to include in the assortment in the face of uncertainty regarding the attractiveness parameter A. The retailer selects the optimal variety n by solving

$$\max_n \alpha \frac{m\lambda n A_H}{1 + nA_H} + (1-\alpha)\frac{m\lambda n A_L}{1 + nA_L} - \beta n$$

where the first two terms are the expected revenue from sales and the last term captures the operational cost of managing variety. The authors show that there exists a unique variety level that maximizes the retailer's profit. The key insight derived from this model is that the retailer's imperfect knowledge about the consumers forces the retailer to act as an expected profit maximizer, and offer a suboptimal category variety. That is, if the retailer knew whether the consumers are L or H-type, the retailer would have offered a higher (when consumers are L-type) or lower (when consumers are H-type) variety compared to the case where the retailer does not know the consumers' type.

Second, the paper considers a model where the retailer delegates the assortment selection decision to a captain in return of a target category profit. The retailer delegates the assortment decision to a captain for two reasons. First, the category captain has better information about consumer preferences. The paper captures the captain's superior knowledge about consumers by assuming that the captain knows the realization of the attraction parameter A (i.e., whether consumers are H-type or L-type). Better information about the parameter A translates into an assortment that better matches consumers' needs. Second, the category captain can collaborate with

the retailer and increase traffic into the category through consumer education, promotions, improved in-store displays and merchandising plans. This benefit is captured by assuming that the captain increases the category traffic from λ to $\lambda + \Lambda$ where Λ denotes the traffic increase due to captainship and captures the captain's ability to stimulate demand at the retailer.

The sequence of events in the captainship scenario is as follows: At stage one, the retailer offers a category captainship contract, which includes a target profit. The captain either accepts or rejects the contract. At stage two, if the contract is accepted, the captain selects variety of the assortment at the retailer. If the captain rejects the contract, the retailer updates its beliefs about the consumers' preferences and decides on variety of the assortment. Essentially, the paper models the captainship as a two stage screening game in which the uninformed retailer makes a take-it-or-leave-it offer to the informed captain and characterizes the pure strategy perfect Bayesian equilibrium.

The category captainship scenario is solved by backward induction. First, the authors consider the captain's assortment selection problem. Then, the authors consider the retailer's target profit setting problem. For a given target profit level, denoted by K, the captain who faces type $i \in \{L,H\}$ consumers solves the following problem at the second stage:

$$\max_{n} (\lambda + \Lambda) \frac{wA_i}{1 + nA_i}$$

$$\text{s.t.} (\lambda + \Lambda) \frac{mnA_i}{1 + nA_i} - \beta n \geq K$$

The category captain's profit is strictly decreasing in the variety offered to the consumers because each additional product in the category cannibalizes the demand for the captain's product. However, the target profit constraint prevents the captain from offering its own product only. Therefore, the captain recommends an assortment where the target profit level is binding. The authors characterize the category captain's best response $n^i(K)$ for $i \in \{L,H\}$.

At stage one, the retailer sets the target profit level K in anticipation of the captain's behavior at the second stage. There are two types of equilibria in Bayesian games (Chu 1992): (1) separating equilibrium and (2) pooling equilibrium. In a separating equilibrium (SE), the uninformed retailer makes an offer such that the informed captain reveals its type. In other words, the retailer sets the target profit such that the captain accepts the offer only if the consumers are H-type. In a pooling equilibrium (PE), the informed captain does not reveal its type because both types accept the retailer's offer. The authors characterize the target profits K_{SE} and K_{PE} that lead to separating and pooling equilibria.

When setting the target profit, the retailer faces a tradeoff between the value of information (about consumer preferences) and the value of additional traffic into the category. If the value of information is greater than the value of additional

traffic, which is the case when Λ is small, the retailer prefers screening the captain. On the other hand, if the value of additional traffic is higher than the value of the captain's private information, which is the case when Λ is large, the retailer prefers the pooling equilibrium.

Comparing the variety levels in the two scenarios reveals that the transition from retail category management to category captainship can increase or decrease the variety offered to the consumers. This increase/decrease is due to two effects: (1) the *adjustment* effect and (2) the *competitive exclusion* effect. The adjustment effect can either increase or decrease the variety of the assortment and is due to the retailer's imperfect knowledge about consumers and the increased traffic into the category. In particular, the adjustment effect is a result of two forces: (1a) variety increase due to higher traffic, and (1b) variety increase or reduction due to better information about consumer preferences. When consumers are L-type, the adjustment effect increases the variety since both higher traffic and better information lead to increase in variety. However, when consumers are H-type, the adjustment effect is ambiguous since higher traffic leads to increase in variety but better information leads to reduction in variety. The adjustment effect suggests a reduced variety only if the possible variety reduction due to better information dominates the variety increase due to additional traffic. The competitive exclusion effect, on the other hand, always reduces the variety and is due to the captain taking advantage of its position and reducing the variety to increase its own profits.

The results in Kurtuluş and Nakkas (2011) have a number of implications regarding the implementation of captainship in practice. The first implication of the paper is that competitive exclusion via reduction in variety (i.e., exclusion of some brands) is possible. However, a reduction in variety under captainship is not always due to competitive exclusion but sometimes due to the adjustment effect. In particular, expected profit maximizing behavior forces the retailer to offer a suboptimal variety under retail category management. The category captain's additional consumer insights help the retailer to adjust its variety to better satisfy consumer's needs. While this adjustment takes place irrespective of the captain's traffic driving abilities, competitive exclusion takes place when the captain is capable of driving significant traffic into the category because the captain is in a stronger position against the retailer in this case. The authors suggest that the presence of these two effects could be one of the reasons for why competitive exclusion is difficult to detect in practice: a reduction in category variety could be due to either the competitive exclusion or the adjustment effect.

Second, Kurtuluş and Nakkas (2011) suggest that while the retailer and the category captain can benefit from captainship, contrary to the common belief, the non-captain manufacturers can also be better off under captainship. While competitive exclusion is a valid concern for the non-captain manufacturers in some instances, the authors find that the variety in the category might actually increase and the non-captain manufacturers can also benefit from captainship.

To summarize, Kurtuluş and Nakkas (2011) shed light on the consequences of captainship when the retailer relies on a captain for assortment decisions and show that category variety can increase or decrease. More importantly, however, this

paper shows (similar to Wang et al. (2003) and Kurtuluş and Toktay (2011)) that captainship could be beneficial for not only the retailer and the captain but also for the non-captain manufacturers.

2.3 Emergence of Category Captainship

Subramanian et al. (2010) examine when and why a retailer may engage one manufacturer exclusively as a category captain to provide category management services and the implications of doing so. Subramanian et al. (2010) consider a setting where two competing manufacturers sell to consumers through a retailer. Category captainship is modeled as follows: a category captain may undertake demand-enhancing services such as better shelf-space management, and design and management of displays within the stores. The paper uses a demand system similar to the one used by Wang et al. (2003) and Kurtuluş and Toktay (2011):

$$q_1 = a_1 - p_1 + \frac{\theta}{1-\theta}(p_2 - p_1) \quad q_2 = a_2 - p_2 + \frac{\theta}{1-\theta}(p_1 - p_2)$$

where the parameter θ is interpreted as the degree of cross-price sensitivity.

The retailer can assign one, both, or neither of the manufacturers to provide service to enhance demand. The sequence of events is as follows: (1) both manufacturers simultaneously propose the services that they would provide if selected as a captain; (2) the retailer can accept one of the proposals, reject both and engage both manufacturers, or decide not to have any retail service provided by the manufacturers. The retailer's category captaincy decision is denoted by $r \in \{0,1,2,J\}$ where $r = i \in \{1,2\}$ if manufacturer i's proposal is accepted, $r = J$ if the retailer decides for joint assignment, and $r = 0$ if the retailer rejects both proposals; (3) if the retailer accepts manufacturer i's proposal, then manufacturer i provides the proposed service. If the retailer chooses joint service, then the manufacturers simultaneously decide the service they will provide; (4) the manufacturers simultaneously set wholesale price w_i; and (5) the retailer sets retail prices p_i.

The authors assume that the service by manufacturers influences the consumers by shifting the base consumption level. When the retailer assigns neither of the manufacturers to provide service (i.e., $r = 0$), the base consumption levels are $a_i = \overline{a}_i$ where $\bar{a}_i$ denotes the consumer's default consumption level. When the retailer assigns only one of the manufacturers to provide demand-enhancing services (i.e., $r = 1$ or $r = 2$), it is assumed that service can increase the base level of demand. In this case, a manufacturer may provide a service that benefits both brands equally or may provide a service that is biased toward its own brand, which could be done at the expense of the competitor's brand. That is, a captain can provide: (1) category-expanding service; and (2) share-shifting service. The category-expanding and share-shifting services of manufacturer i are denoted by e_{ic} and e_{is}, respectively. The base consumption levels in this case are given by

$$a_1 = \overline{a}_i + \frac{e_{ic} + e_{is}}{2} \quad a_j = \overline{a}_j + \frac{e_{ic} - e_{is}}{2} \quad \text{for } i,j = \{1,2\},\ j \neq i$$

In this model, category-expanding service boosts the base consumption level for both brands, whereas share-shifting service increases the base consumption level for the category captain's brand at the expense of the competitor's brand. When $e_{ic} < e_{is}$, the captain's service enhances its own demand and decreases the rival's demand and is the service is mainly share-shifting. On the other hand, when $e_{ic} > e_{is}$, the captain's service enhances demand for all brands and the service is mainly category-expanding. The cost of providing service (e_c,e_s) is given by

$$C(e_c, e_s) = \frac{1}{2}\left[4\frac{k}{1-k}e_c^2 + (e_c + e_s)^2\right]$$

where $k \in [1,1/3]$ is a cost parameter that indicates how much more costly category-expanding service is relative to share-shifting service.

The authors also consider an alternative to the category captain arrangement where the retailer involves both manufacturers simultaneously, which the authors refer to as the joint service provision, for retail service (i.e., $r = J$). Let e_i^J denote the service provided by manufacturer i in the joint service model. The base consumption levels in this case are given by

$$a_1 = \overline{a}_1 + \frac{e_{1c}^J + e_{1s}^J}{2} + \frac{e_{2c}^J - e_{2s}^J}{2} \qquad a_2 = \overline{a}_2 + \frac{e_{1c}^J + e_{1s}^J}{2} + \frac{e_{2c}^J - e_{2s}^J}{2}$$

The cost of service in the joint service model is given by $(1/\mu)C(e_{ic}^J, e_{is}^J)$ where $\mu \in [0,1]$ captures the relative efficiency of joint service provision as compared to providing service exclusively as the captain. When $\mu = 1$, the service under the joint service model is as efficient as under the captain arrangement. As μ decreases, joint service becomes relatively less efficient. When $\mu \to 0$, joint service is inefficient and becomes infeasible.

Given these assumptions, the retailer and manufacturers' profits can be written as

$$\Pi_R = (p_1 - w_1)q_1 + (p_2 - w_2)q_2$$

$$\Pi_1 = w_1 q_1 - \delta(r = 1)C(e_{1c}, e_{1s}) - \delta(r = J)\frac{1}{\mu}C(e_{1c}, e_{1s})$$

$$\Pi_2 = w_2 q_2 - \delta(r = 2)C(e_{2c}, e_{2s}) - \delta(r = J)\frac{1}{\mu}C(e_{2c}, e_{2s})$$

where $\delta(x)$ is the indicator function and is equal to one if x is true and zero if x is false.

The authors find that a captain may provide a service that enhances demand for all brands in a category despite doing so is more costly for the captain. However, the non-captain manufacturer may benefit from the captainship arrangement even

if the captain's service depletes its demand. This is more likely to happen in categories where cross-price sensitivity between the competing brands is high. The authors find a negative relation between the degree of manufacturers' price competition (cross-price sensitivity) in a category and the extent of their competition to become category captain. Consequently, the authors conclude that captainship can be beneficial for manufacturers in product categories where cross-price sensitivity is high. Furthermore, the authors identify conditions under which the manufacturers may even be worse off than they would be without the captainship implying that captainship is not always beneficial for the manufacturers.

The retailer, on the other hand, benefits from category captainship when the cross-price sensitivity is low because when the cross-price sensitivity is low, the competition for category captainship stimulates service to such an extent that the retailer prefers to appoint one of the manufacturers as a captain rather than engaging both manufacturers jointly. The findings in Subramanian et al. (2010) may help explain why, despite concerns regarding competitive exclusion, the practice of captainship where the retailer relies on a single manufacturer has become increasingly popular over the recent years, and why there is limited evidence of harm to non-captain manufacturers.

While Subramanian et al. (2010) consider the emergence of category captainship in a context where the retailer relies on a captain for demand enhancing service only, Kurtuluş et al. (2014) consider the emergence of captainship in a setting where the retailer relies on a captain for both demand enhancing service and assortment decisions. Kurtuluş et al. (2014) observe that the prevalence of captainship practices varies significantly from one category to another. Based on a number of cases from trade publication Progressive Grocer and their interviews with several category managers, they observe that many successful implementations have taken place in certain categories (e.g., Canned and Packaged Foods, Frozen Foods, and Health and Beauty Care). They also observe that there are no successful implementations in categories such as Dairy Milk and Fresh Produce. The authors conjecture that this is presumably because captainship delivers higher value to the involved parties in some categories and lower in others.

Motivated by these observations, Kurtuluş et al. (2014) investigate the environments where captainship is more valuable for both the retailer and the captain, and identify the conditions under which captainship benefits all parties involved. This is the first paper that models the competition among manufacturers for captainship and the retailer's captain selection process via an auction where the manufacturers bid for the captainship role.

To this end, Kurtuluş et al. (2014) consider a two-stage supply chain with multiple manufacturers that sell their products to consumers through a retailer. The scope of category management in this paper is assortment decisions and demand-enhancing activities. The paper models demand enhancing as follows: It is assumed that the total category demand is a function of the effort that the retailer (or the captain) exerts into marketing activities such as consumer education programs, advertisement campaigns, and designing efficient planograms. The base rate

category traffic is normalized to one. By exerting marketing effort x, the retailer (or the captain) can increase the category traffic to $(1 + x)$. In order to capture the decreasing returns to marketing effort, the model assumes a convex cost function of the form $x^2/(2c)$ where c is the traffic driving capability of the party exerting the effort.

Similar to Kurtuluş and Nakkas (2011), this paper uses a generic attraction market share model (Bell et al. 1975; Gruca and Sudharshan 1991) to model demand for each product in the category where all products are equally attractive, that is $A_i = A$ for $i = 1,2,\ldots,N$ and the no-purchase option's attractiveness is set to $A_0 = 1$. The market share of each product when the retailer offers n products is given by $q(n) = A/(1 + nA)$. Thus, the average demand rate for each product is given by $(1 + x)q(n)$.

Similar to Kurtuluş and Nakkas (2011), this paper also assumes that all products have the same wholesale price w and retail prices p, and production costs are normalized to zero. The retailer's net profit margin is $m = p - w$. Similar to the model in Kurtuluş and Nakkas (2011), the authors assume that the retailer incurs an operational cost, which is linear in the variety offered in the category, βn with $\beta > 0$ (Honhon and Pan 2013; Kurtuluş and Nakkas 2011).

The authors first consider the benchmark scenario which is in line with the traditional approach where the retailer manages the category internally and decides on the marketing effort, x, and the number of products in the assortment, n, to maximize its profit; that is,

$$\max_{x,n} (1 + x)\frac{mnA}{1 + nA} - \beta n - \frac{x^2}{2c_R}$$

where the first term in the retailer's profit is the revenue from sales, the second term is the operational cost of managing variety, and the last term is the cost of effort (with c_R denoting the retailer's capability to drive traffic). Solving the retailer's problem, the authors characterize the retailer's optimal effort and variety as well as the profits of the retailer and manufacturers that are included in the assortment in the benchmark scenario.

The authors then consider the category captainship scenario where the retailer selects a captain and outsources the category management activities (marketing effort and assortment) to the captain. To capture the heterogeneity in manufacturers' abilities to drive traffic, the authors assume that the cost of increasing category traffic by x_i (for manufacturer i) is given by $x_i^2/(2c_i)$ where c_i is the privately known capability of manufacturer i. The retailer believes that manufacturers' capabilities c_i are independent and drawn from a uniform distribution on the interval $[0, \overline{c}]$.

In practice, retailers select their captains by soliciting proposals from multiple manufacturers for category captainship. The retailer usually selects the manufacturer that promises to deliver the highest performance improvement. The authors model the process of captain selection and the competition among manufacturers

for captainship as a first-price auction where the retailer invites K of the N manufacturers to submit proposals for the captainship role.

The sequence of events is as follows: First, the retailer announces the captainship auction and K manufacturers simultaneously bid their promised total category sales to the retailer. The highest bidder is selected to serve as a captain. The captain exerts marketing effort and decides on the variety to be offered at the retailer. The captainship scenario is solved by backward induction by first deriving the captain's variety and effort decisions assuming that the captain has been selected. If the manufacturer with capability c has been selected as a captain by bidding S, the captain selects variety to maximize profit subject to meeting the target S; that is,

$$\max_{x,n} (1+x)\frac{wA}{1+nA} - \frac{x^2}{2c}$$
$$\text{s.t.} \quad (1+x)\frac{nA}{1+nA} \geq S$$

The authors characterize the category captain's effort and variety response for given target sales level S. Then the authors consider the bidding behavior in the captain selection auction where the manufacturers bid for the captainship role. In the bidding for captainship, each manufacturer faces the following trade-off: If a manufacturer wins the auction, the manufacturer is assured that his product will be included in the assortment but incurs the cost of exerting effort. On the other hand, if the manufacturer loses the auction, then he benefits from the captain's effort (without incurring cost) but there is a possibility that his product will be excluded from the assortment. The auction for captainship is not a standard sealed-bid first-price auction since the manufacturers benefit from captainship even if they lose the auction but are included in the assortment. Thus, the captainship auction creates positive externalities that are endogenously determined by the captain's post-auction marketing effort and variety decisions. These positive externalities create a free-riding incentive for the bidders. The strength of the externalities is determined by the probability of exclusion for the non-captain manufacturers, which is an increasing function of the number of manufacturers N.

In this context, the authors find that the most capable manufacturer wins the auction and characterize the equilibrium effort and variety set by the captain. They also characterize the resulting expected *ex-ante* profits for the retailer, the captain, and the non-captain manufacturers who are included in the assortment. The authors proceed to study the value of category captainship by comparing the *ex-ante* expected profits of the involved parties in the benchmark and captainship scenarios and derive a number of insights, which are summarized below.

Emergence of category captainship: Captainship is valuable for both the retailer and the captain (therefore more likely to emerge) when the captain is more cost effective (more capable) in exerting marketing effort compared to the retailer, and the cost of managing variety, retail margins (relative to manufacturers' margins), and competition for captainship are moderate.

One factor contributing to the emergence of captainship in categories such as Canned Fruits and Vegetables and Frozen Pizza is the capability differential between the manufacturers and retailers in these categories. Most manufacturers in these categories have a national presence and dedicate significant resources into category management (e.g., Heinz, Kraft, and Dole in Canned Fruits and Vegetables; Kraft and General Mills in Frozen Pizza). The rate of new product introductions in these categories is high because of frequently changing consumer needs. Manufacturers closely follow consumer trends; hence they are more capable of developing strategies to grow these categories compared to the retailers. In addition, a number of manufacturers with significant capabilities compete for captainship, which is another factor that contributes to the successful captainship implementations in these categories.

On the other hand, the authors point out that the lack of successful captainship implementations in categories such as the dairy milk can be attributed to limited competition for captainship and lower supplier capability. Consumer preferences in such categories are well understood and stable and there are only a few smaller manufacturers that have limited resources to dedicate into category management.

Impact of captainship on non-captain manufacturers: When a manufacturer is assigned to serve as a captain, this usually results in frustration for the non-captain manufacturers because of the fear of exclusion. The authors demonstrate that this is a valid concern in some cases but also point that captainship can benefit not only the retailer and the captain, but also the non-captain manufacturers. Whether the non-captain manufacturers benefit from captainship is determined by whether the benefits of the increased traffic dominate the possibility of being excluded from the category.

Impact of captainship on marketing effort and variety: When the retailer performs category management, an increase in marketing effort leads to an increase in variety. When these decisions are delegated to a captain, a higher marketing effort allows the captain to reduce variety to increase its market share. Hence, when the effort and variety levels are compared across the two scenarios, the effort is usually higher but variety is lower under captainship.

2.4 Antitrust Concerns

Some economists have voiced antitrust concerns related to category captainship (Steiner 2001; Desrochers et al. 2003; Leary 2003; Klein and Wright 2006). In the US, the Antitrust Institute has voiced reservations about category captainship. In Europe, ECR has taken the lead to ensure that category captainship is implemented in compliance with European Union competition rules.

Desrochers et al. (2003) states that antitrust concerns related to category captainship practices focus around two issues: (1) competitive exclusion and (2) competitive collusion. The exclusion-based concern is that smaller competitors are denied the right to compete for category captainship because they do not have the

necessary resources (Desrochers et al. 2003). Retailers usually assign one of their leading manufacturers to serve as a category captain because only those manufacturers have the necessary resources that can benefit the retailer. Big manufacturers already invest a great deal in consumer research and can use these resources toward helping retailers manage their categories better. The concern is that category captain manufacturers' power will be further enhanced and smaller manufacturers will be put at a disadvantage.

Prior research on captainship has provided some evidence supporting and some evidence refuting the competitive exclusion hypothesis and is inconclusive. For example, Morgan et al. (2007) argue that the category captains will engage in opportunistic behavior. However, Gooner et al. (2011) show that category captains can improve category management at the retailer without engaging in opportunistic behavior. Subramanian et al. (2010), Kurtuluş and Toktay (2011), and Kurtuluş and Nakkas (2011) offer some theoretical evidence that competitive exclusion exists but also point to the possibility that captainship can benefit all involved parties including the non-captain manufacturers.

Competitive collusion concerns include the possibility that a category captain can use its role to facilitate collusion and limit the competition among rivals in the category (Desrochers et al. 2003). First, the category captain may transfer sensitive information such as pricing, merchandising, and promotion plans from one manufacturer to another. When manufacturers in the category know about their rivals' pricing, they might price more or less aggressively, or if they know about their rivals' promotion plans, they may promote their brands more selectively. Second, the category captain can coordinate its recommendations across the retailers for which it serves as category captain. Desrochers et al. (2003) suggest that if retailers are more selective in sharing sensitive data with their category captains, some forms of competitive collusion scenarios can be avoided.

To summarize, while category captainship practices in the retailing sector present a very valuable opportunity for the retailers to benefit from their captain manufacturers' expertise and resources, these practices also open up an opportunity for the captain manufacturers to take advantage of their positions as captains and exclude competitors and restrict competition in the categories. While research shows that category captainship may have significant positive impact on the retailer's and the captain's and in some instances on the non-captain manufacturers' performances, existing research also identifies circumstances under which captainship practices result in competitive exclusion.

3 Impact of Category Captainship Practices on the Retail Industry

In this section, we consider how category captainship practices could potentially change the nature of the manufacturer-retailer relationships and the landscape in the retail industry. Practices such as category captainship delegate considerable power

to the category captain manufacturers because in most cases they can effectively control outcomes in the category (Desrochers et al. 2003). While some retailers continue to work with their category captains and verify their recommendations, other retailers prefer to implement their captain's recommendations 'as presented by the captain' mainly due to lack of resources. While private information on the category captain's part makes it easier for the category captain to provide biased recommendations and control the outcomes in the category, it also makes it more difficult for the retailers to detect bias in a category captain's recommendations. The category captain's influence over the retailer also depends on the size of the retailer. Small retailers are more likely to accept and implement the captain's recommendations in 'as is' manner, whereas larger retailers have more control over the process and are more likely to implement their category captain's recommendations after verifying them.

In order to decrease the amount of control given to the captains, some retailers assign a second manufacturer in the category to serve as a co-captain and use them as consultants to verify the category captain's recommendations. In addition, the retailers renegotiate the captainship agreements by reviewing the captain's performance frequently to balance the power in the supply chain (Kurtuluş and Toktay 2004).

A potential adverse effect of category captainship on retailers is the loss of capability to manage the categories internally. Retailers should be aware that category management requires a thorough understanding of consumer preferences and purchase patterns, a knowledge base that is hard to build once that expertise is lost (Kurtuluş and Toktay 2004).

Traditionally, manufacturers such as Procter&Gamble and Unilever were the main players in the consumer goods industry and retailers were primarily a means of reaching consumers. The early 1990s saw an increase in the number of high quality new product introductions and the emergence of other strong manufacturers, which led to higher competition for shelf-space. This, combined with the retailers' awareness of the importance to be in contact with end consumers, provided the basis for a shift in power from manufacturers to retailers. Many retailers such as Wal-Mart and Carrefour owe their rapid growth to these developments (Corstjens and Corstjens 1995).

As Corstjens and Corstjens describe in their influential book Store Wars, "...the giant retailers, now, stand as an obstacle between the manufacturers and the end consumers, about as welcome as a row of high-rise hotels between the manufacturer's villa and the beach." Their book describes the contemporary national brand manufacturers over the past two decades as being in a continuous battle for shelf-space and mind-space at the retailers. It is therefore not surprising that manufacturers would advocate any initiative that can increase their influence over retail decisions, and category captainship is one such practice. But by outsourcing retail category management to their leading manufacturers, retailers may in the long run lose their capabilities in managing their product categories and their knowledge about consumers. This loss of capability may prepare the basis for a shift in power back from the retailers to the manufacturers (Kurtuluş and Toktay 2004).

Given this changing landscape in the consumer goods supply chains over the past few decades; an intriguing question is what will happen to the retailer-manufacturer relationships and power balance in the consumer goods supply chains in the near future. With the growing popularity of category captainship practices (and other similar practices such as vendor managed inventory and direct store delivery) in the retail industry, the number of manufacturer-retailer partnerships (e.g., Wal-Mart and P&G, Carrefour and Colgate) is increasing. While such partnerships will positively influence the partner manufacturers, they will also place the non-partnering manufacturers at a disadvantage, forcing them to become a partner to a leading retailer. Manufacturers' battle for shelf-space and mind-space over the past decade has started to transform into a battle for being a partner (e.g., category captain) for a major retailer (Kurtuluş and Toktay 2004).

4 Future Research Directions

Although category captainship practices became widespread in the retail industry over the past decade, the consequences of using captains for category management are not fully understood by either academics or practitioners. Therefore, we believe that there is room for more original research in this field. We have identified five directions for future research that would help both academics and practitioners to better understand the consequences of category captainship practices.

First, existing research on category captainship assumes that the retailers either delegate the pricing, or the assortment or retail service decisions such as shelf-space management to a captain. However, in practice, the scope of category captainship implementations is broader: retailers rely on their captain's for a combination of these decisions. Therefore, exiting models cannot fully capture the category captainship phenomenon. The question of how different category captainship arrangements impact the retailer and the manufacturers needs to be answered when the retailer relies on its category captain for a combination of assortment, pricing, shelf-space management, and promotion planning recommendations. Future research can take advantage of the existing research on joint inventory and pricing decisions in operations (see Petruzzi and Dada (1999), Elmaghraby and Keskinocak (2003), and Yano and Gilbert (2003) for literature reviews on different aspects of the joint pricing and inventory decisions) that could be used as the basis for investigating the impact of jointly delegating the shelf-space allocation and pricing decisions to a leading manufacturer. In addition, there is a literature on trade promotions in marketing (e.g., Lal and Villas-Boas 1998; Kim and Staelin 1999) and operations (e.g., Iyer and Ye 2000; Huchzermeier et al. 2002) that could be used as the basis for research to understand the impact of recommendations made by captains to their retailers about different aspects of promotion planning.

Second, existing research on category captainship is mainly based on mathematical models. However, answering broader questions would require empirical research. In particular, empirically testing the impact of category captainship practices on the

financial performance of the retailers and understanding when such practices would benefit the retailers would be a good starting point. Empirical research is also needed to test the hypothesis that category captainship may result in competitive exclusion. Such empirical research would provide a basis for the antitrust cases that are under investigation regarding category captainship misconduct.

Third, existing research on category captainship exclusively focuses on categories where products are substitutes. However, a product category sometimes can consist of complementary products such as toothpaste and toothbrush products in the oral care category. Future research should be conducted to understand the differences in category captainship implementations where the products are substitutes versus complements, and whether categories where the retailer offers complementary products are more suitable for category captainship.

Fourth, future research should explore the value of having an independent third party (i.e., intermediary) providing category management services for retailers. Companies such as ACNielsen collect and sell syndicated data and software that can be used for category management; however, they do not provide category management recommendations. Research is needed to understand the advantages and disadvantages of using a third party for category captainship. On one hand, retailers could take advantage of the expertise and resources of the third party providers without worrying about bias in the recommendations provided. On the other hand, the retailers should be concerned about losing their internal category management capabilities. Another source of concern for the retailers is that these third party providers would provide recommendations to many retailers that compete for the same consumers, potentially causing the retailer to lose its competitive edge.

Finally, future research should consider if and how information leakages as a result of captainship implementations play a role on the value of captainship for the retailers. Category captainship requires that the retailer share significant amount of confidential information with its captain manufacturers. Given that a manufacturer often serves as a category captain for many retailers that compete for the same consumers, the captain manufacturer serves as an information hub by collecting valuable consumer information from multiple retailers. As a result, the captain manufacturers gain significant power in making the category decisions such as pricing for not only their own brands but for all brands in a category. Retailers, on the other hand, may abstain from sharing proprietary information because the leakage of proprietary information to competitors via the category captain can result in loss of competitiveness. It would be valuable to investigate if and how such leakages can play a role on the value of category captainship for the retailers. Future research in this area can take advantage of and build on the existing research on Resale Price Maintenance (e.g., Chen 1999; Deneckere et al. 1996) discussed in Sect. 2.1, which utilizes models where a single manufacturer sells to consumers through multiple competing retailers.

References

ACNielsen. (2005). *Step four: Set performance targets and measure progress with a category scorecard in consumer-centric category management: How to increase profits by managing categories based on consumer needs*. New York, NY: Wiley.

Apparel Magazine. (2005, November 10). VF on 'Three Cs' of category captainship. *Apparel Magazine*.

Aydin, G., & Hausman, W. H. (2009). The role of slotting fees in the coordination of assortment decisions. *Production and Operations Management, 18*(6), 635–652.

Basuroy, S., Mantrala, M., & Walters, R. G. (2001). The impact of category management on retailer prices and performance: Theory and evidence. *Journal of Marketing, 65*(4), 16–32.

Bell, D. E., Keeney, R. L., & Little, J. D. C. (1975). A market share theorem. *Journal of Marketing Research, 12*(2), 136–141.

Cachon, G. P., & Kok, A. G. (2007). Category management and coordination of categories in retail assortment planning in the presence of basket shoppers. *Management Science, 53*(6), 934–951.

Cachon, G. P., Terwiesch, C., & Xu, Y. (2008). On the effects of consumer search and market entry in a multi-product competitive market. *Marketing Science, 27*(3), 461–473.

Chen, Y. (1999). Oligopoly price discrimination and resale price maintenance. *RAND Journal of Economics, 30*(3), 441.

Chen, Y., Hess, J. D., Wilcox, R. T., & Zhang, Z. J. (1999). Accounting profits versus marketing profits: A relevant metric for category management. *Marketing Science, 18*, 208–229.

Choi, S. C. (1991). Price competition in a channel structure with a common retailer. *Marketing Science, 10*(4), 271–296.

Chu, W. (1992). Demand signalling and screening in channels of distribution. *Marketing Science, 11*(4), 327–347.

ECR Conference. (2004). Category management is here to stay, Brussels, 2004. *A category with tremendous potential*. http://www.ecrnet.org/05-projects/catman/Bxl%202004_category%20management.ppt#622,21,1.

Corstjens, J., & Corstjens, M. (1995). *Store wars: The battle for mindspace and shelf-space*. New York, NY: Wiley.

Corstjens, M., & Doyle, P. (1983). A dynamic model for strategically allocating retail space. *Journal of the Operational Research Society, 34*(10), 943–951.

Deneckere, R., Marvel, H. P., & Peck, J. (1996). Demand uncertainty, inventories, and resale price maintenance. *The Quarterly Journal of Economics, 111*, 885–913.

Desrochers, D. M., Gundlach, G. T., & Foer, A. A. (2003). Analysis of antitrust challenges to category captain arrangements. *Journal of Public Policy and Marketing, 22*(2), 201–215.

Dhar, S. K., Hoch, S. J., & Kumar, N. (2001). Effective category management depends on the role of the category. *Journal of Retailing, 77*(2), 165–184.

Elmaghraby, W., & Keskinocak, P. (2003). Dynamic pricing in the presence of inventory considerations: Research overview, current practices and future directions. *Management Science, 49*, 1287–1309.

Gilligian, T. (1986). The competitive effects of resale price maintenance. *RAND Journal of Economics, 17*, 544–556.

Gooner, R. A., Morgan, N. A., & Perreault, W. D. (2011). Is retail category management worth the effort (and does a category captain help or hinder)? *Journal of Marketing, 75*, 18–33.

Greenberger, R. S. (2003, January 15). UST must pay $ 1.05 billion to a big tobacco competitor. *Asian Wall Street Journal*, New York, NY, p.A.8.

Gruca, T. S., & Sudharshan, D. (1991). Equilibrium characteristics of multinomial logit market share models. *Journal of Marketing Research, 28*(November 1991), 480–482.

Gruen, T. W., & Shah, R. H. (2000). Determinants and outcomes of plan objectivity and implementation in category management relationships. *Journal of Retailing, 76*(4), 483–510.

Honhon, D., & Pan, A. (2013). Assortment planning with vertically differentiated products. *Production and Operations Management, 21*(2), 253–275.

Huchzermeier, A., Iyer, A., & Freheit, J. (2002). The supply chain impact of smart customers in a promotional environment. *Manufacturing and Service Operations Management, 4*(3), 228.
Ippolito, P. M., & Overstreet, T. R. (1996). Resale price maintenance: An economic assessment of the federal commission's case against the corning glass works. *Journal of Law and Economics, 39*, 285.
Iyer, A. V., & Ye, J. (2000). Assessing the value of information sharing in a promotional retail environment. *Manufacturing and Service Operations Management, 2*(2), 128–143.
Kim, S. Y., & Staelin, R. (1999). Manufacturer allowances and retail pass-through rates in a competitive environment. *Marketing Science, 18*(1), 59–76.
Kök, A. G., Fisher, M. L., & Vaidyanathan, R. (2008). Assortment planning: Review of literature and industry practice. In N. Agrawal & S. Smith (Eds.), *Retail supply chain management*. New York, NY: Springer.
Kurtuluş, M., & Nakkas, A. (2011). Retail assortment planning under category captainship. *Manufacturing and Service Operations Management, 13*(1), 124–142.
Kurtuluş, M., & Toktay, L. B. (2004). Category captainship: Who wins, who loses? *ECR Journal, 4*(2), 2004.
Kurtuluş, M., & Toktay, L. B. (2011). Category captainship vs. retailer category management under limited retail shelf-space. *Production and Operations Management, 20*(1), 47–56.
Kurtuluş, M., Nakkas, A., & Ulku, S. (2014). The value of category captainship in the presence of manufacturer competition. *Production and Operations Management, 23*(3), 420–430.
Lal, R., & Villas-Boas, M. (1998). Price promotions and trade deals with multiproduct retailers. *Management Science, 44*(7), 935–949.
Leary, T. B. (2003). *A second look at category management*. Federal Trade Commision report. Retrieved, from www.ftc.gov.
McGuire, T. W., & Staelin, R. (1983). An industry equilibrium analysis of downstream vertical integration. *Marketing Science, 2*(2), 161–191.
Morgan, N. A., Kaleka, A., & Gooner, R. A. (2007). Focal supplier opportunism in supermarket retailer category management. *Journal of Operations Management, 25*, 512–527.
Nielsen Marketing Research. (1992). *Category management: Positioning your organization to win*. Lincolnwood, IL: NTC Business Books.
Overstreet, T. (1983). *Resale price maintenance: Economic theories and empirical evidence*. Washington, DC: Federal Trade Commission.
Petruzzi, N. C., & Dada, M. (1999). Pricing and the newsvendor problem: A review with extensions. *Operations Research, 47*, 183–194.
Progressive Grocer. (2004, November 15). Category captains 2004: Captains of the industry. *Progressive Grocer*.
Progressive Grocer. (2007, November 15). Category captains 2007: Armed and ready. *Progressive Grocer*.
Progressive Grocer. (2008, November 15). Category captains 2008. *Progressive Grocer*.
Progressive Grocer. (2010, November 15). Category captains 2010. *Progressive Grocer*.
Shubik, M. J., & Levitan, R. E. (1980). *Market structure and behavior*. Cambridge, MA: Harvard University Press.
Shugan, S. M. (1989). Product assortment in a triopoly. *Management Science, 35*(3), 304–321.
Steiner, R. L. (2001). Category management – A pervasive, new vertical/horizontal format. *Antitrust, 15*(Spring), 77–81.
Subramanian, U., Raju, J. S., Dhar, S. K., & Wang, Y. (2010). Competitive consequences of using a category captain. *Management Science, 56*(10), 1739–1765.
Telser, L. G. (1960). Why should manufacturers want free trade? *Journal of Law and Economics, 3*, 86.
van Ryzin, G., & Mahajan, S. (1999). On the relationship between inventory costs and variety benefits in retail assortment. *Management Science, 45*(11), 1496–1509.
Villas-Boas, J. M. (1998). Product line design for a distribution channel. *Marketing Science, 17*(2), 156–169.
Vives, X. (1999). *Oligopoly pricing: Old ideas and new tools*. Cambridge, MA: The MIT Press.

Wang, Y., Raju, J. S., & Dhar, S. K. (2003). *The choice and consequences of using a category captain for category management*. Pennsylvania, PA: The Wharton School, University of Pennsylvania.

Klein, B., & Wright, J. D. (2006). *Antitrust analysis of category management: Conwood v. United States Tobacco*. George Mason University Law and Economics Research Paper Series.

Yano, C., & Gilbert, S. M. (2003). Coordinated pricing and production/procurement decisions: A review. In A. Chakravarty & J. Eliashberg (Eds.), *Managing business interfaces: Marketing, engineering and manufacturing perspectives*. Berlin: Kluwer Academic Publishing.

Chapter 8
Assortment Planning: Review of Literature and Industry Practice

A. Gürhan Kök, Marshall L. Fisher, and Ramnath Vaidyanathan

1 Introduction

A retailer's assortment is defined by the set of products carried in each store at each point in time. The goal of assortment planning is to specify an assortment that maximizes sales or gross margin subject to various constraints, such as a limited budget for purchase of products, limited shelf space for displaying products, and a variety of miscellaneous constraints such as a desire to have at least two vendors for each type of product.

Clearly the assortment a retailer carries has an enormous impact on sales and gross margin, and hence assortment planning has received high priority from retailers, consultants and software providers. However, no dominant solution has yet emerged for assortment planning, so assortment planning represents a wonderful opportunity for academia to contribute to enhancing retail practice. Moreover, an academic literature on assortment planning is beginning to emerge. The purpose of this chapter is to review the academic literature on assortment planning, to overview the approaches to assortment planning used by several

This paper is an invited chapter to appear in *Retail Supply Chain Management*, Eds. N. Agrawal and S. A. Smith, Kluwer Publishers.

A.G. Kök
College of Administrative Sciences and Economics, Koç University, Istanbul, Turkey
e-mail: gkok@ku.edu.tr; http://home.ku.edu.tr/gkok/

M.L. Fisher
The Wharton School, University of Pennsylvania, Philadelphia, PA, USA
e-mail: fisher@wharton.upenn.edu

R. Vaidyanathan (✉)
Desautels Faculty of Management, McGill University, Montréal, QC, Canada
e-mail: ramnath.vaidyanathan@mcgill.ca

N. Agrawal, S.A. Smith (eds.), *Retail Supply Chain Management*,
International Series in Operations Research & Management Science 223,
DOI 10.1007/978-1-4899-7562-1_8

retailers so as to provide some examples of practice, and to suggest directions for future research.

Retailers engage in assortment planning because they need to periodically revise their assortment. Several factors require a retailer to change their assortment, including seasons (the fall assortment for an apparel retailer will be different from the spring assortment), the introduction of new products and changes in consumer tastes.

Most retailers segment the stock keeping units (SKU) they carry into groups called categories. For example, for a consumer electronics retailer, a category might be personal computers. Within categories, they will usually define subcategories, such as laptops and desktops within the computer category. (The terminology used varies across retailers e.g. department, class and subclass may be used instead of category and subcategory, but the practice of grouping SKUs with similar attributes for planning purposes is universal.) Retailers focus most of their energy on deciding what fraction of their shelf space and product purchase budget to devote to each category and subcategory. For example, a consumer electronics retailer would worry more about how to divide their resources between laptops and desktops than about which specific models of each to carry, a decision that is usually left to a more junior buyer. The resource allocation decisions are based on their own historical sales in each subcategory, especially whether sales in a subcategory have been trending up or down, together with external information from a variety of sources such as industry shows, vendors and competitor moves.

Given fixed store space and financial resources, assortment planning requires a tradeoff between three elements: how many different categories does the retailer carry (called a retailer's breadth), how many SKUs do they carry in each category (called depth), and how much inventory do they stock of each SKU, which obviously affects their in-stock rate. The breadth vs. depth tradeoff is a fundamental strategic choice faced by all retailers. Some, like department stores, will elect to carry a large number of different categories. Others, such as category killers like Toys 'R Us and Best Buy, will specialize in a smaller number of categories, but have great depth in each category.

We have all had the experience of going into a store looking for a particular product, not finding it, and settling for another similar product instead. This is called substitution, and the willingness of customers to substitute within a particular category is an important parameter in assortment planning. If customers have a high propensity to substitute in a category, then providing great depth and a high in-stock rate is less critical. The reverse is also true.

We can delineate three patterns with respect to customer substitution: (1) the customer shops a store repeatedly for a daily consumable and one day she finds it stocked out so she buys another. This is called stock-out based substitution. (2) a customer identifies a favorite product based on ads or what she has seen in other stores, but when she tries to find it in a particular store, she can't because they don't carry it, so see buys another product. This is called assortment based substitution. (3) the consumer chooses her favorite product from the ones she sees on the shelf in a store when she is shopping and buys it if it has higher utility than her no purchase option.

In this case, there may be other products she would have preferred, (but she didn't see them either because the retailer didn't carry them or because they were stocked out), and in this sense we can say she substituted, although she may not be aware that these other products exist and hence doesn't herself think of her purchase decision as involving substitution. The first two patterns are common with daily consumables like food and the later with consumer durables like apparel or consumer electronics.

Assortment planning is a relatively new but quickly growing field of academic study. The academic approach to the assortment planning problem rests on the formulation of an optimization problem with which to choose the optimal set of products to be carried and the inventory level of each product. Decisions for each product are interdependent because products are linked in considerations such as shelf space availability, substitutability between products, common vendors (brands), joint replenishment policies and so forth. Most of the literature focuses on a single category or subcategory of products at a given point in time. While a retailer might have a different assortment at each store, the academic literature has focused on determining a single assortment for a retailer, which could be viewed as either a common assortment to be carried at all stores or the solution to the assortment planning problem for a single store.

This chapter begins in Sect. 2 by briefly reviewing four streams of literature that assortment planning models build on: product variety and product line design, shelf space allocation, multi-product inventory systems and a consumer's perception of variety.

In Sect. 3, we discuss empirical results on consumer substitution behavior and present three demand models used in assortment planning: the multinomial logit, exogenous demand and locational choice models.

In Sect. 4, we describe optimization based assortment planning studies. Sections 4.1–4.3 review optimization approaches for the basic assortment planning problem. The models and solution methodologies in these papers vary because of differences in the underlying demand model and the application context. We then review variations on the basic assortment planning problem, including assortment planning with supply chain considerations in Sect. 4.4, assortment planning with demand learning and assortment changes during the selling season in Sect. 4.5, and multi-category assortment planning that considers the interactions between different categories due to existence of basket shopping consumers in Sect. 4.7.

In Sect. 5, we discuss demand and substitution estimation methodologies. The methods depend on the demand model and the type of data that is available.

In Sect. 6, we present industry approaches to assortment planning. We describe the assortment planning process at four prominent retailers: Electronics retailer Best Buy, book and music retailer Borders, Indian jewelry retailer Tanishq, and Dutch supermarket chain Albert Heijn. As will be seen, these companies take significantly different approaches and emphasize different aspects of the assortment problem.

In Sect. 7, we provide a critical comparison of the academic and industry approaches and use this to identify research opportunities to bridge the gap between the two approaches.

For an earlier overview of the assortment planning literature, see Mahajan and van Ryzin (1999).

2 Related Literature

In this section, we briefly review the literature on topics related to assortment planning.

2.1 Product Variety and Product Line Design

Product selection and the availability of products has a high impact on the retailer's sales, and as a result gross profits and assortment planning has been the focus of numerous industry studies, mostly concerned with whether assortments were too broad or narrow. Retailers have increased product selection in all merchandise categories for a number of reasons, including heterogeneous customer preferences, consumers seeking variety and competition between brands: Quelch and Kenny (1994) report that the number of products in the market place increased by 16 % per year between 1985 and 1992 while shelf space expanded only by 1.5 % per year during the same period. This has raised questions as to whether rapid growth in variety is excessive. For example, many retailers are adopting an "efficient assortment" strategy, which primarily seeks to find the profit maximizing level of variety by eliminating low-selling products (Kurt Salmon Associates 1993), and "category management," which attempts to maximize profits within a category (AC Nielsen 1998). There is empirical evidence that variety levels have become so excessive that reducing variety does not decrease sales (Dreze et al. 1994; Broniarczyk et al. 1998; Boatwright and Nunes 2001). And from the perspective of operations within the store and across the supply chain, it is clear that variety is costly: a broader assortment implies less demand and inventory per product, which can lead to slow selling inventory, poor product availability, higher handling costs and greater markdown costs.

The literature that studies the economics of product variety is vast. The main model in this field is the oligopoly competition between single product firms based on Hotelling (1929). In the Hotelling model, consumers are distributed uniformly on a line segment and firms choose their positions on the line segment and their prices to maximize profits. Consumers' utility from each firm is decreasing in the firm's price and their physical distance to the firm. Each

consumer chooses the firm that provides her the maximum utility. The objective is to find the number of firms, their locations and their prices in equilibrium and the resulting consumer welfare. Extensions of this model are used to study product differentiation. There are two types of product differentiation. In a horizontally differentiated market, products are different in features that can't be ordered. In that case, each of the products is ranked first for some of the consumers. A typical example is shirts of different color. In a vertically differentiated market, products can be ordered according to their objective quality from the highest to the lowest. A higher quality product is more desirable than a lower quality product for any consumer. Anderson et al. (1992) and Lancaster (1990) provide excellent reviews of this literature.

One of the outgrowths of the literature on the economics of product variety is the product line design problem pioneered by Mussa and Rosen (1978) and Moorthy (1984). A monopolist chooses a subset of products from a continuum of vertically differentiated products and their prices to be sold in a market to a variegated set of customer classes in order to maximize total profit. Consider cars as a product with a single attribute, say engine size. The monopolist's problem is to choose what size engines to put in the cars and how to price the final product. These papers assume convex production costs and do not consider operational issues such as fixed costs, changeover costs, and inventories. Joint consideration of marketing and production decisions in product line design is reviewed by Eliashberg and Steinberg (1993). Dobson and Kalish (1993) propose a mathematical programming solution for this problem in the presence of fixed costs for each product included in the assortment. Desai et al. (2001) study the product line design problem with component commonality. Netessine and Taylor (2007) extend Moorthy's (1984) work by using the Economic Order Quantity (EOQ) model to incorporate economies of scale. de Groote (1994) also considers concave production costs and analyzes the product line design problem in a horizontally differentiated market. He shows that the firm chooses a product line to cover the whole market and the product locations are equally spaced. Alptekinoğlu (2004) extends this work to two competing firms, one offering infinite variety through mass customization and the other limited variety under mass production. He shows that the mass producer needs to reduce variety in order to mitigate the price competition. Chen et al. (1998) is the only paper that considers product positioning and pricing with inventory considerations. They show that the optimal solution for this model under stochastic demand can be constructed using dynamic programming.

These models were early treatments of assortment planning from the manufacturer's view that were precursors of similar models developed for retailing. The manufacturer's problem is one of product positioning in an attribute space (quality or some other attribute) and pricing. The retailer's problem is to select products from the product lines of several manufacturers. A more careful consideration of inventories at product level is needed in retail assortment planning, since inventories have a direct impact on both sales and costs for the retailer.

2.2 Multi-Item Inventory Models

Multi-item inventory problems are also highly relevant to the assortment planning problem. The inventory management of multiple products under a single a shelf space or budget constraint is studied extensively in the operations literature and solutions using Lagrangian multipliers is presented in various textbooks, e.g., Hadley and Whitin (1963). Downs et al. (2002) describe a heuristic approximation to the multi-period version of this problem with lost sales. In these models, the demand of products are not dependent on others' inventory levels (i.e., there is no substitution between products).

The other group of inventory models with multiple products consider stock-out based substitution, focusing on the stocking decisions given a selection, but not the selection of the products. These models are based on an exogenous model of demand which we shall describe in the next section. Briefly, the total demand of a product is the sum of its own initial demand and the substitution demand from other products. Substitution demand from product k to j is a fixed proportion α_{kj} of the unsatisfied demand of product j. McGillivray and Silver (1978) first introduced the problem with two products. Parlar and Goyal (1984) study the decentralized version of the problem. Noonan (1995) and Rajaram and Tang (2001) present heuristic algorithms for the solution of the case with n products. Netessine and Rudi (2003) investigate the case with n products under centralized and decentralized management regimes. The complexity of the problem is prohibitive and it is not possible to obtain an explicit solution to the problem. Netessine and Rudi (2003) find that a decentralized regime carries more inventory than the centralized regime because of the competition effects. Mahajan and van Ryzin (2001b) establish similar results under dynamic customer substitution with the multinomial logit choice model. Parlar (1985) and Avsar and Baykal-Gursoy (2002) study the infinite horizon version of this problem under centralized and competitive scenarios respectively. Lippman and McCardle (1997) consider a single period model under decentralized management, where aggregate demand is a random variable and demand for each firm is a result of different rules of initial allocation and reallocation of excess demand. Bassok et al. (1999) consider an alternative substitution model, in which the retailer observes the entire demand before allocating the inventory to products. In this retailer controlled substitution model, the retailer may upgrade a customer to a higher quality product. The reallocation solution is obtained by solving a transportation problem.

The literature on assemble-to-order systems is also related. The demand for individual components are linked through the demand for finished goods. See Song and Zipkin (2003) for a review. An online retailer's order fulfillment problem when customers can order multiple products can be viewed as an assemble-to-order systems. Song (1998) estimates the order fill rate in such systems and discusses other examples from retailing.

2.3 Shelf Space Allocation Models

In some product segments such as grocery and pharmaceuticals, how much shelf space is allocated to a given product category is an important component of the assortment planning process. This view seems especially relevant for fast moving products whose demand is sufficiently high that a significant amount of inventory is carried on the shelf. This contrasts with other categories e.g., shoes, music, books where only one or two units are carried for most SKUs, hence amount of inventory and shelf space are not critical decisions at product level. As one example, Transworld Entertainment carries 50,000 SKUs in an average store but stock more than one of only the 300 best sellers.

In an influential paper Corstjens and Doyle (1981) suggest a method for allocating shelf space to categories. They perform store experiments to estimate sales of product i as $\alpha_i s_i^{\beta_i} \prod_j s_j^{\delta_{ij}}$, where s_i is the space allocated to product i, β_i is own space elasticity, and δ_{ij}s are the cross-space elasticities. Cost functions of the form $\gamma_i s_i^{\tau_i}$, are also estimated from the experiments. The problem of profit maximization with a shelf space constraint is solved within a geometric programming framework. Their results are significantly better than commercial algorithms that allocate space proportional to sales or to gross profit by ignoring interdependencies between product groups. The estimation and optimization procedures can not be applied to large problems, hence they elect to work with product groups rather than SKUs. Bultez and Naert (1988) apply the Corstjens and Doyle (1981) model at the brand level assuming symmetric cross elasticities (i.e., $\delta_{ij} = \delta$ for all i, j) within product groups. Their model is tested at four different Belgian supermarket chains, leading to encouraging results.

An interesting paper by Borin and Farris (1995) reports the sensitivity of the shelf space allocation models to forecast accuracy. They compare the solution with correct parameters to that with incorrect parameter estimates. Even when the error in parameter estimates are 24 %, the net loss in category return on inventory is just over 5 % compared to the optimal allocation based on true estimates. This proves the robustness of these models to estimation errors. Similar to these shelf space allocation papers, but using an inventory theoretic perspective, Urban (1998) models the own and cross product effects of displayed inventory on demand rate in a mathematical program and solves for shelf space allocation and optimal order-up-to quantities. He reports that on average a greedy heuristic yields solutions that are within 1 % of a solution obtained by genetic programming.

Irion et al. (2012) extend the Corstjens and Doyle model to study the shelf space allocation problem at the product level. Demand for each product is a function of its own and other products' shelf space through own and cross shelf space elasticities. The cost for each product consists of linear purchasing costs, inventory costs from an economic order quantity model, and a fixed cost of being included in the assortment. The objective is to allocate (integer) number of facings to each product in order to maximize profits under a total shelf space availability constraint and lower and upper bounds on the number of facings for each product. The problem is

transformed into a mixed integer program (MIP) with linear constraints and objective function through a series of linearization steps. The linearization framework is general enough to accommodate several extensions. However, there is no empirical evidence that product level demand can be modeled as a function of the shelf space allocated to the product itself and competing products via own and cross space elasticities.

Shelf space allocation papers do not explicitly address assortment selection and inventory decisions and ignore the stochastic nature of demand.

2.4 Perception of Variety

Consumer choice models often assume that customers are perfectly knowledgeable about their preferences and the product offerings. Therefore, consumers are always better off when they choose from a broader set of products. However, empirical studies show that consumer choice is affected by their perception of the variety level rather than the real variety level. This perception can be influenced by the space devoted to a category, the presence or absence of a favorite item (Broniarczyk et al. 1998), or the arrangement of the assortment (Simonson 1999). Hoch et al. (1999) define a measure of the dissimilarity between product pairs as the count of attributes on which a product pair differs. They show that this measure is critical to the perception of variety of an assortment and that consumers are more satisfied with stores carrying those assortments perceived as offering high variety. van Herpen and Pieters (2002) find the impact of two attribute-based measures that significantly impact the perception of variety. These measures are entropy (whether all products have the same color or different colors) and dissociation between attributes (whether color and fabric choice across products are uncorrelated). The perception of variety at a store is especially important for variety-seeking consumers. Variety seeking consumers tend to switch away from the product consumed on the last occasion. Variety-seeking literature demonstrated that consumers adopt this behavior when purchasing food or choosing among hedonic products such as restaurants and music. See Kahn (1995) for a review. Intrapersonal factors (e.g., satiation and the need for stimulation), external factors (e.g., price change, new product introduction), and uncertainty about future preferences promote variety-seeking behavior. On a final note, variety can even negatively affect consumers experience: confusion or complexity due to higher variety may cause dissatisfaction of consumers and decrease sales (Huffman and Kahn 1998).

3 Demand Models

This section provides a review of demand models as background for assortment planning models. We first present the empirical evidence for consumer driven substitution which is a fundamental assumption in many assortment planning

models. The Multinomial Logit model is a discrete consumer choice model, which assumes that consumers are rational utility maximizers and derive customer choice behavior from first principles. Exogenous demand models directly specify the demand for each product and what an individual does when the product he or she demands is not available. The locational choice model is also a utility-based model. Before proceeding, we will define the notation for assortment planning in a single subcategory at a single store. This notation is common throughout this chapter and additional time or store subscripts are introduced when necessary.

N The set of products in a subcategory, $N = \{1, 2, \ldots, n\}$,
S The subset of products carried by the retailer, $S \subset N$,
r_j Selling price of product j,
c_j Purchasing cost of product j,
λ Mean number of customers visiting the store per period.

3.1 Consumer Driven Substitution

We define two types of substitution with a supply side view of the causes of substitution: *Stockout-based* substitution is the switch to an available variant by a consumer when her favorite product is carried in the store, but is stocked-out at the time of her shopping. *Assortment-based* substitution is the switch to an available variant by a consumer when her favorite product is not carried in the store.

The substitution possibilities in retailing can be classified into three groups. (a) Consumer shops a store repeatedly for a daily consumable, and one day she finds it stocked out so she buys another. This is an example of stockout-based substitution. (b) Consumer has a favorite product based on ads or her past purchases at other stores, but the particular store she visited on a given day may not carry that product. This is an example of assortment-based substitution. (c) Consumer chooses her favorite from what she sees on the shelf and buys it if it is better than her no purchase option. In this case, there may be other products she may have preferred, but she didn't see them either because the retailer didn't carry them or they are stocked out. This could be an example for either substitution type depending on whether the first choice product is temporarily stocked out or not carried at that store. First two cases fit repeat purchases like food and the third fits one time purchases like apparel.

Let's focus on the options of a consumer who can not find her favorite product in a store, because it is either temporarily stocked out or not carried at all. She can (a) buy one of the available items from that category (substitute), (b) decide to come back later for that product (delay), (c) decide to shop at another store (lost customer). If the consumer chooses to substitute, the sale is lost from the perspective of the first favorite product. Table 8.1 summarizes the findings of empirical studies on the consumer response to stockouts. The most recent one, Gruen et al. (2002) examine consumer response to stockouts across eight categories at

Table 8.1 Consumer response to stockouts in six studies of substitute-delay-leave behavior

	Substitute (%)	Delay (%)	Leave (%)
Progressive Grocer (1968a,b)	48	24	28
Walter and Grabner (1975)	83	3	14
Schary and Christopher (1979)	22	30	48
Emmelhainz et al. (1991)	36	25	39
Zinn and Liu (2001)	62	15	23
Gruen et al. (2002)	45	15	40

retailers worldwide and report that 45 % of customers substitute, i.e., buy one of the available items from that category, 15 % delay purchase, 31 % switch to another store, and 9 % never buy that item.

The above mentioned papers study the consumer response to stockouts, i.e. stockout based substitution, although none of them explicitly excludes assortment-based substitution. Campo et al. (2004) investigate the consumer response to out-of-stocks (OOS) as opposed to permanent assortment reductions (PAR). They report that although the retailer losses in case of a PAR may be larger than those in case of an OOS, there are also significant similarities in consumer reactions in the two cases and OOS reactions for an item can be indicative of PAR responses for that item.

3.2 Multinomial Logit

The Multinomial Logit (MNL) model is a utility-based model that is commonly used in economics and marketing literatures. We create product 0 to represent the no-purchase option, i.e., a customer that chooses 0 does not purchase any products. Each customer visiting the store associates a utility U_j with each option $j \in S \cup \{0\}$. The utility is decomposed into two parts, the deterministic component of the utility u_j and a random component ε_j.

$$U_j = u_j + \varepsilon_j.$$

The random component is modeled as a Gumbel random variable. Also known as Double Exponential or Extreme value Type-I, it is characterized by the distribution

$$Pr\{X \leq \varepsilon\} = \exp(-\exp - (\varepsilon/\mu + \gamma))),$$

where γ is Euler's constant (0.57722). Its mean is zero, and variance is $\mu^2\pi^2/6$. A higher μ implies a higher degree of heterogeneity among the customers. The realizations of ε_j are independent across consumers. Therefore, while each consumer has the same expected utility for each product, realized utility may be

different. This can be due to the heterogeneity of preferences across customers or unobservable factors in the utility of the product to the individual.

An individual chooses the product with the highest utility among the set of available choices. Hence, the probability that an individual chooses product j from $S \cup \{0\}$ is

$$p_j(S) = \Pr\left\{U_j = \max_{k \in S \cup \{0\}} (U_k)\right\}.$$

The Gumbel distribution is closed under maximization. Using this property, we can show that the probability that a customer chooses product j from $S \cup \{0\}$ is

$$p_j(S) = \frac{e^{u_j/\mu}}{\sum_{k \in S \cup \{0\}} e^{u_k/\mu}}. \tag{8.1}$$

See Anderson et al. (1992) for a proof. This closed form expression makes the MNL model an ideal candidate to model consumer choice in analytical studies. See Ben-Akiva and Lerman (1985) for applications to the travel industry, Anderson et al. (1992) for MNL based models of product differentiation, Basuroy and Nguyen (1998) for equilibrium analysis of market share games and industry structure. Moreover, starting with Guadagni and Little (1983), marketing researchers found that MNL model is very useful in estimating demand for a group of products. We will briefly discuss the parameter estimation of MNL model in Sect. 5.1. For more details on the MNL model and its relation to other choice models, see Anderson et al. (1992) or Mahajan and van Ryzin (1999).

The major criticism of the MNL model stems from its Independence of Irrelevant Alternatives (IIA) property. This property holds if the ratio of choice probabilities of two alternatives is independent of the other alternatives in the choice process. Formally, this property is

for all $R \subset N, T \subset N, R \subset T$, *for all* $j \in R, k \in R$,

$$\frac{p_j(R)}{p_k(R)} = \frac{p_j(T)}{p_k(T)}.$$

IIA property would not hold in cases where there are subgroups of products in the choice set such that the products within the subgroup are more similar with each other than across subgroups. Consider an assortment with two products from different brands. If brand loyalty is high, adding a new product from the first brand can cannibalize the sales of its sister product more than the rival product. IIA does not capture this important aspect of consumer choice. Another example that illustrates this property is the "blue bus/red bus paradox": Consider an individual going to work and has the same probability of using his or her car or of taking the bus: $\Pr\{car\} = \Pr\{bus\} = 1/2$. Suppose now that there are

two buses available that are identical except for their color, red or blue. Assume that the individual is indifferent about the color of the bus he or she takes. The choice set is {car, red bus, blue bus}. One would intuitively expect that $\Pr\{car\} = 1/2$ and $\Pr\{red\ bus\} = \Pr\{blue\ bus\} = 1/4$. However, the MNL model implies that $\Pr\{car\} = \Pr\{red\ bus\} = \Pr\{blue\ bus\} = 1/3$.

The Nested Logit Model introduced by Ben-Akiva and Lerman (1985) is one way to deal with the IIA property. A two-stage nested process is used for modeling choice, e.g., first brand choice then SKU choice. The choice set N is partitioned into subsets N_l, $l = 1, \ldots, m$ such that $\cup_{l=1}^{m} N_l = N$ and $N_l \cap N_k = \emptyset$ for any l and k. The individual chooses with a certain probability one of the subsets, from which he or she chooses a variant from that subset. The utility from the choice within subset N_l is also Gumbel distributed with mean $\mu \ln \sum_{j \in N_l} e^{u_j/\mu}$ and scale parameter μ. As a result, the choice process between the subsets follows the MNL model as well and the probability that a consumer chooses variant j in subset N_l is

$$P_j(N) = P_{N_l}(N) * P_j(N_l).$$

Chapter 2 in Anderson et al. describes the Nested Logit in great detail. In the Nested Logit Model, the IIA property no longer holds when two alternatives are not in the same subgroup. However, the use of the Nested Logit requires the knowledge of key attributes and their hierarchy for consumers and makes estimation problems more difficult. Nested Logit model is used in modeling the competition between two-multiproduct firms in several studies (Anderson et al. 1992; Cachon et al. 2008).

Another related shortcoming of the MNL model is related to substitution between different products. The MNL model in its simplest form is unable to capture an important characteristic of the substitution behavior. The utility of the no-purchase option with respect to the utility of the products in S determines the rate of substitution. Consider the following example, where $S = \{1, 2\}, \mu = 1$, and $u_0 = u_1 = u_2$. The share of each option is determined by the implication of MNL that the probability of choosing option i is $\exp(u_i)/(\exp(u_0) + \exp(u_1) + \exp(u_2)) = 1/3$ for $i = 0, 1, 2$. Hence, two thirds of the customers are willing to make a purchase from the category. If the second product is unavailable, the probability of her choosing the first product is $\exp(u_1)/(\exp(u_0) + \exp(u_1)) = 1/2$. That is, half of the consumers whose favorite is stocked out will switch to the other product as a substitute and the other half will prefer no-purchase alternative to the other product. In this example, the penetration to the category (purchase incidence) is 2/3 and the average substitution rate is 1/2. These two quantities are linked via u_i's. We can control the substitution rate by varying u_0, but that also determines the initial penetration rate to the category. Hence, it is not possible with this model to have two categories with the same penetration rate but different substitution rates, which we have found severely limits the applicability of this model.

Miranda Bront et al. (2009) show that the CDLP model of the assortment problem with multiple segments is NP-hard and propose a column generation algorithm. Rusmevichientong and Topaloglu (2012) propose a robust formulation of the assortment optimization problem.

3.3 Exogenous Demand Model

Exogenous demand models directly specify the demand for each product and what an individual does when the product he or she demands is not available. There is no underlying consumer behavior such as a utility model that generates the demand levels or that explains why consumers behave as described in the model. As mentioned before, this is the most commonly used demand model in the literature on inventory management for substitutable products. The following assumptions fully characterize the choice behavior of customers.

(A1) Every customer chooses her favorite variant from the set N. The probability that a customer chooses product j is denoted by p_j. $\sum_{j \in N \cup \{0\}} p_j = 1$.

(A2) If the favorite product is not available for any reason, with probability δ she chooses a second favorite and with probability $1 - \delta$ she elects not to purchase. The probability of substituting product j for k is α_{kj}.

When the substitute item is unavailable, consumers repeat the same procedure: decide whether or not to purchase and choose a substitute. The lost sales probability $(1 - \delta)$ and the substitution probabilities could remain the same for each repeated attempt or specified differently for each round.

As a result of (A1) average demand rate for product j is $d_j = \lambda\, p_j$, and total demand to the category is $\sum_{j \in N} d_j = \lambda(1 - p_0)$.

α_{kj} is specified by a substitution probability matrix that can take different forms to represent different probabilistic mechanisms. Consider the following examples for a four-product category.

Random substitution matrix

$$\begin{bmatrix} 0 & \frac{\delta}{n-1} & \frac{\delta}{n-1} & \frac{\delta}{n-1} \\ \frac{\delta}{n-1} & 0 & \frac{\delta}{n-1} & \frac{\delta}{n-1} \\ \frac{\delta}{n-1} & \frac{\delta}{n-1} & 0 & \frac{\delta}{n-1} \\ \frac{\delta}{n-1} & \frac{\delta}{n-1} & \frac{\delta}{n-1} & 0 \end{bmatrix}$$

Adjacent substitution matrix

$$\begin{bmatrix} 0 & \delta & 0 & 0 \\ \delta/2 & 0 & \delta/2 & 0 \\ 0 & \delta/2 & 0 & \delta/2 \\ 0 & 0 & \delta & 0 \end{bmatrix}$$

Within subgroups substitution matrix

$$\begin{bmatrix} 0 & \delta & 0 & 0 \\ \delta & 0 & 0 & 0 \\ 0 & 0 & 0 & \delta \\ 0 & 0 & \delta & 0 \end{bmatrix}$$

Proportional substitution matrix

$$\begin{bmatrix} 0 & \delta d_2/(\lambda - d_1) & \delta d_3/(\lambda - d_1) & \delta d_4/(\lambda - d_1) \\ \delta d_1/(\lambda - d_2) & 0 & \delta d_3/(\lambda - d_2) & \delta d_4/(\lambda - d_2) \\ \delta d_1/(\lambda - d_3) & \delta d_2/(\lambda - d_3) & 0 & \delta d_4/(\lambda - d_3) \\ \delta d_1/(\lambda - d_4) & \delta d_2/(\lambda - d_4) & \delta d_3/(\lambda - d_4) & 0 \end{bmatrix}$$

The single parameter δ enables us to differentiate between product categories with low and high substitution rates. The adjacent substitution matrix assumes that products are ordered along an attribute space and allows for substitution between neighboring products only. For example, if a customer can't find 1 % milk in stock, she may be willing to accept either 2 % or skim, but not whole milk. Subgroups substitution matrix allows for substitution within the subgroups only. For example, in the coffee category, consumers may treat decaffeinated coffee and regular coffee as subgroups and not substitute between subgroups.

In the proportional substitution model, the general expression for α_{kj} is

$$\alpha_{kj} = \delta \frac{d_j}{\sum_{l \in N \setminus \{k\}} d_l}. \tag{8.2}$$

The proportional substitution matrix has properties that are consistent with what would happen in a utility-based framework such as the MNL model. $\alpha_{kj} > \alpha_{kl}$ if $d_j > d_l$. Suppose that a store doesn't carry the whole assortment, i.e., $N \setminus S \neq \emptyset$. Since only one round of substitution is allowed, the realized substitution rate from variant k to other products is $\sum_{j \in S} \alpha_{kj} = \delta \sum_{j \in S} d_j / \sum_{l \in N \setminus \{k\}} d_l$, which is increasing in the set S. This means that a consumer who can not find her favorite variant in the store is more likely to buy a substitute, as the set of potential substitutes grows.

We next state an assumption commonly made in assortment planning models for tractability.

(A3) No more attempts to substitute occur. Either the substitute product is available and the sale is made, or the sale is lost.

Limiting the number of substitution attempts (A3) is not too restrictive. Smith and Agrawal (2000) show that number of attempts allowed has a smaller effect as more items are stocked, because the probability of finding a satisfactory item by the second try quickly approaches one. Kök (2003) presents an example where

effective demands under a three-attempts substitution model with rate $\delta = 0.5$ can be approximated almost perfectly with a single-attempt-substitution model with rate $\delta = 0.58$.

The exogenous demand model has more degrees of freedom than the MNL model. Since the options in the choice set are assumed to be homogenous, MNL model is unable to capture the types of adjacent substitution, one-product substitution, or within subgroup substitution. In the MNL model the substitution rates depend on the relative utility of the options in $N \cup \{0\}$. This is both an advantage and a disadvantage for the MNL model. The advantage is that it allows one to easily incorporate marketing variables such as prices and promotions into the choice model. The disadvantage is that it cannot differentiate between the initial choice and substitution behavior. Unlike the MNL model, the exogenous demand model can differentiate between categories that have same initial demand for the category but different substitution rates through the choice of p_0 and δ. Therefore, the MNL model cannot treat assortment-based and stockout-based substitutions differently. In contrast, it is certainly possible to use a different δ or different substitution probability matrices for assortment-based and stockout-based substitutions in the exogenous demand model.

3.4 Locational Choice Model

Also known as the address or the characteristics approach, the locational choice model was originally developed by Hotelling (1929) to study the pricing and location decisions of competing firms. Extending Hotelling's work, Lancaster (1966, 1975) proposed a locational model of consumer choice behavior. In this model, products are viewed as a bundle of their characteristics (attributes) and each product can be represented as a vector in the characteristics space, whose components indicate how much of each characteristic is embodied in that product. For example, defining characteristics of a car include its engine size, gas consumption, and reliability. Each individual is characterized by an ideal point in the characteristics space, which corresponds to his or her most preferred combination of characteristics.

Suppose that there are m characteristics of a product. Let z_j denote the location of variant j in R^m. Consider a consumer whose ideal product is defined by $y \in R^m$. The utility of variant j to the consumer is

$$U_j = k - r_j - g(y, z_j),$$

where k is a positive constant, r_j is the price, and $g : R^m \to R$ is a distance function, representing the disutility associated with the distance from the consumer's ideal point, e.g., Euclidean distance or the rectilinear distance. The consumer chooses the variant that gives him or her the maximum utility. For an extensive discussion of the

address approach and its relation to stochastic utility models such as the MNL model, the reader is referred to Chap. 4 in Anderson et al. (1992).

There is one major difference between the locational choice model and the MNL model. In the MNL model, substitution can happen between any two products. In the locational choice model however, IIA property does not hold and substitution between products is localized to products with specifications that are close to each other in the characteristics space. Hence, the firm can control the rate of substitution between products by selecting their locations to be far apart or close to each other.

4 Assortment Selection and Inventory Planning

The majority of the papers focus on assortment decisions at a single store. Most papers take a static view of the assortment planning problem, that is the assortment decisions are made once and inventory costs are computed either from a single period model or the steady-state average of a multi-period model. In Sects. 4.1–4.3, we review four such papers categorized according to the demand model that they are based on. The papers based on the choice models are more stylized but are able to obtain structural properties of the optimal solution. The papers based on the exogenous demand model are more flexible and have more applicability because they allow for more realistic details in modeling, such as nonidentical prices and case packs. In Sect. 4.4, we review assortment planning papers with supply chain considerations. Section 4.5 discusses a dynamic assortment planning model in which the retailer has a chance to update its assortment throughout the season as it updates its demand estimates every period for products in the assortment. A recent development in the assortment planning literature is the consideration of multiple categories, where consumers are basket shoppers and the assortment decisions across categories are interdependent. In Sect. 4.7, we discuss two such papers. The first presents an optimization method and the second discusses the long-run impact of variety by considering store choice decisions of consumers.

4.1 Assortment Planning with Multinomial Logit: The van Ryzin and Mahajan Model

van Ryzin and Mahajan (1999) formulate the assortment planning problem by using a MNL model of consumer choice. Assume $r_j = r$ and $c_j = c$ for all j. Products are indexed in descending order of their popularity, i.e., such that $u_1 \geq u_2 \geq .. \geq u_n$. Define $v_j = e^{u_j/\mu}$. By the MNL share formula, the probability that a customer demands product j is

$$p_j(S) = \frac{v_j}{\sum_{k \in S \cup \{0\}} v_j}. \tag{8.3}$$

We assume consumers make their product choice (if any) when they observe the assortment, and they do not look for a substitute if the product of their choice is stocked out. Hence, $p_j(S)$ is independent of the inventory status of the products in S. Note that the demand increase in product j due to the decision $S \subseteq N$ is

$$p_j(S) - p_j(N).$$

This demand increase is due to what is termed assortment-based substitution and is comprised of demand from consumer who would have preferred a product in $N - S$ but had to substitute to product j. van Ryzin and Mahajan (1999) also calls this static substitution.

In contrast, in dynamic substitution, consumers observe the inventory levels of all products at the time of their arrival and make their product choice among the products that are available. Hence, dynamic substitution includes both assortment- and stockout-based substitution.

The expected profit of a variant $j \in S$ is

$$\pi_j(S) = (r - c)\lambda p_j(S) - C(\lambda p_j(S)),$$

where $C(\cdot)$ is the operational costs. The cost function is assumed to be concave and increasing to reflect the economies of scale in inventory models such as the EOQ or the newsvendor models.

The objective is to maximize the total category profits by solving

$$\max_{S \subset N} \sum_{j \in S} \pi_j(S).$$

The optimal assortment finds a balance between including a new product and increasing the total demand to the category and cannibalizing the demand of other products' sales and increasing their average cost.

Consider the net profit impact of adding a variant j to assortment S. Define $S_j = S \cup \{j\}$.

$$h(v_j) = \pi_j(S_j) - \left(\sum_{k \in S} \pi_k(S) - \sum_{k \in S} \pi_k(S_j) \right)$$

If the profit of product j is more than the sum of the profit losses of the products in S, then adding j improves profits.

Theorem 1 *The function $h(v_j)$ is quasi-convex in v_j in the interval $[0,\infty)$.*

Since a quasi-convex function achieves its maximum at the end points of the interval, the profit is maximized either by not adding a product to the assortment or by adding the product with the highest v (i.e., the most popular product). This observation leads to the following result that characterizes the structure of the optimal assortment. Define the popular assortment set:

$$P = \{\{\}, \{1\}, \{1,2\}, .., \{1,2,..,n\}\}.$$

Theorem 2 *The optimal assortment is always in the popular assortment set.*

This result is intuitive and powerful: it reduces the number of assortments to be considered from 2^n to n. Since only assortment-based substitution is considered, the demand for each product, the optimal inventory level and the resulting profit can be computed for each of the n assortments in the popular assortment set. The above theorems as stated are from Cachon et al. (2005). van Ryzin and Mahajan (1999) originally proved this result for a cost function from the newsvendor model. Specifically, they use the expected costs of a newsvendor model assuming that D is distributed according to a Normal distribution with mean λ and standard deviation σ. The optimal stocking level of product j is the newsvendor stocking quantity:

$$x_j = \lambda p_j(S) + z\sigma\left(\lambda p_j(S)\right)^{\beta},$$

where $z = \Phi^{-1}(1 - c/r)$ and $\beta \in [0,1)$ controls the coefficient of variation of the demand to product j as a function of its mean. The resulting cost function is

$$C(\lambda p_j(S)) = r\sigma\frac{e^{-z^2}}{\sqrt{2\pi}}\left(\lambda p_j(S)\right)^{\beta}.$$

The authors show that a deeper assortment is more profitable with a sufficiently high price, and a sufficiently high no-purchase preference. In order to compare different merchandising categories, the authors define the fashion of a category using majorization arguments. In a more fashionable category, the utility across products are more balanced, therefore in expectation the market shares of all products are evenly distributed. The paper shows that everything else being equal, the profit of a more fashionable category is lower due to the fragmentation of demand.

This model captures the main trade-off between variety and the increased average inventory costs. The analysis leads to the elegant results that establish the structural properties of the optimal assortment. However, not all assortment planning problems fit the assumption of homogenous group of products with identical prices and costs. The style/color/size combination of shirts in a clothing retailer may be a good example. Even then, the substitutions would occur across styles/

colors but not sizes. The assumption that there is a single opportunity to make assortment and inventory decisions can be defended in products with short life cycles, where the season is too short to make changes in the assortment and bring the new products to market before the season is over. Clearly, the main result (Theorem 2) does not hold when products have nonidentical price, cost parameters, or different operational characteristics such as demand variance, case pack, and minimum order quantity.

4.1.1 Extensions

Mahajan and van Ryzin (2001a) study the same problem under dynamic substitution. That is, the retailer faces the problem of finding the optimal product selection and stocking levels where customers dynamically substitute among products when inventory is depleted. Consider a customer with the following realization of the utilities: $u_6 > u_4 > u_3 > u_5 > u_0 > u_1 > u_2$. Suppose that the store carries assortment $S = \{1, 2, 3, 4\}$. In the static substitution model, this consumer would choose product 4, buy it if it is available and leave the store if it is not. In the dynamic substitution model, products 4, 3, and 5 are all acceptable to the customer, in that order of preference. Depending on the inventory levels of those products, she will buy the one that is available in the store at the time she visited the store, and won't buy anything only if none of those three products is available. Using a sample path analysis, the authors show that the problem is not even quasi-concave. By comparing the results of a stochastic gradient algorithm with two newsvendor heuristics, they conclude that the retailer should stock more of the more popular variants and less of the less popular variants than a traditional newsvendor analysis suggests. Also, the numerical results support the theoretical insight (Theorem 2) obtained under static substitution. Maddah and Bish (2004) extend the van Ryzin Mahajan model by considering the pricing decisions as well.

Cachon et al. (2005) study the van Ryzin and Mahajan (1999) model in the presence of consumer search, motivated by the following observation: Even when a consumer finds an acceptable product at the retail store, the consumer still faces an uncertainty about the products outside the store's assortment. Therefore, she may be willing to go to another store and explore other alternatives with the hope of finding a better product. In the independent search model, consumers expect each retailer's assortment to be unique, and hence utility of search is independent of the assortment. Examples for this setting include jewelry stores and antique dealers. In the overlapping assortment search model, products across retailers overlap, hence the value of search decreases with the assortment size at the retailer. For example, all retailers choose their digital camera assortments from the product lines of a few manufacturers. In contrast to the no-search model, in the presence of consumer search it may be optimal to include an unprofitable product in the assortment. Therefore, failing to incorporate consumer search in assortment planning results in narrower assortments and lower profits.

Vaidyanathan and Fisher (2012) study the assortment planning problem under a setup similar to van Ryzin and Mahajan (1999), but in the presence of more general demand distributions. They approximate the expected profit function and evaluate simple heuristics to select the optimal assortment and set inventory levels, in the presence of stock-out substitution. They also present analytical bounds on the error due to optimizing an approximate profit function instead of the exact one.

Miller et al. (2010) consider the retailer's assortment selection problem with heterogeneous customers and test the impact of different consumer choice models on the optimal assortment. They develop a sequential choice model in which customers first form Consideration Sets and then make product choices based on the MNL model.

Li (2007) extend the van Ryzin and Mahajan (1999) and show that under continuous traffic, the optimal assortment consists of a set of products with the highest profit rates, even when product margins are unequal.

Kök and Xu (2011) use a nested logit model to study assortment decisions for a product category with heterogeneous product types from two brands. They consider two different hierarchical structures for the nests: a brand-primary model in which consumers choose a brand first, then a product type in the chosen brand, and a type-primary model in which consumers choose a product type first, and then a brand within that product type. They extend the structural properties of assortment decisions characterized by van Ryzin and Mahajan (1999) to the case of Nested Logit. A more detailed discussion of this paper can be found in Sect. 4.6.

Alptekinoğlu and Grasas (2014) apply the nested logit model to study assortment decisions under consumer returns, for a set of horizontally differentiated products. They show that when refund amounts are sufficiently high, or when returns are disallowed, the optimal assortment consists of only the most popular products, a result consistent with van Ryzin and Mahajan (1999). However, when return policies are relatively strict, and refund amounts are low, they find that it might be optimal for the retailer to offer a mix of the most popular and eccentric products. They support their findings with some empirical evidence that eccentric products are usually associated with higher return probabilities.

Davis et al. (2014) show that the assortment optimization problem under the nested logit model can be solved in polynomial time, when customers are assumed to always make their purchase from the selected nest, and the nest dissimilarity parameters satisfy certain conditions. In the absence of either of these assumptions, they demonstrate that the problem is NP-hard.

Alptekinoğlu and Semple (2013) propose a new discrete choice model, termed the Exponomial Choice Model, which modifies the MNL model, by assuming exponentially distributed random errors. They obtain closed form expressions for the choice probabilities and find that unlike the MNL model, the exponomial model does not suffer from the independence of irrelevant alternatives (IIA) property. Additionally, they show that the exponomial choice model is easy to estimate, since the loglikelihood function is concave in the unknown parameters. They derive structural properties of the optimal assortment and prices, under a number of

scenarios. Finally, they estimate the exponomial choice model on two sets of choice data and compare the results with the MNL model.

4.1.2 Preference Ordering Models

Honhon et al. (2010) study a single-period joint assortment and inventory planning problem when customers are classified based on their preference ordering of products. They assume that total customer demand is random, and the market is comprised of fixed proportions of different customer types, based on preference ordering. They develop efficient, pseudopolynomial time algorithms to solve the resulting assortment optimization problem.

Honhon et al. (2012) study the optimal assortment problem under the assumption that (a) customers can be characterized into types based on a rank-ordered list of products they are willing to purchase, (b) proportion of consumers of each type is random and (c) purchases are dynamic, consumer-driven and stockout based. Following Honhon et al. (2010), the authors relax the assumption of random proportions to show that the expected profits for the resulting fixed proportions model (FP) can be used to construct tight bounds on the expected profits for the random proportions model. Finally, they use these bounds and numerical simulations to (a) study optimality gap as a function of problem parameters and (b) conclude that the FP heuristic performs favorably to other previously known heuristics in literature.

Honhon et al. (2012) study the optimal assortment selection problem under four different ranking-based consumer choice models, the one-way substitution, the locational choice, the outtree, and the intree preference model. They model the problem assuming that the retailer incurs a fixed carrying cost for every product offered, a goodwill penalty when a customer is unable to find his first choice, and lost sales penalty when a customer is not able to find any acceptable product. Under these assumptions, they find that the first three models can be solved efficiently using a shortest path algorithm or dynamic program. For the intree preference model, they construct an algorithm that is efficient and performs better than enumeration based methods in numerical experiments.

Pan and Honhon (2012) study the assortment planning problem for a category of vertically differentiated products. There is a fixed cost to include a product in the assortment and additional variable costs are incurred per unit sold. Customers are utility maximizers and differ in their valuation of quality, which is exogenously determined. They find that under fixed selling prices, the optimal assortment might include strictly dominated products, that are less attractive on every possible dimension, as compared to at least one other product not carried in the assortment. In the scenario where the retailer can set the selling prices, they find that this counter-intuitive feature of the optimal assortment disappears. They propose several efficient algorithms to determine the optimal assortment and pricing structure, and test them on real data for two product categories.

4.2 Assortment Planning Under Exogenous Demand Models

In this subsection, we review two closely related assortment planning models that consider both assortment-based and stockout-based substitution. Smith and Agrawal (2000) focus on constructing lower and upper bounds to the problem in order to formulate a mathematical program. Kök and Fisher (2007) formulate the problem in the context of an application at a supermarket chain and proposes a heuristic solution to a similar mathematical program. They also provide structural results on the assortments that generate new insights and guidelines for practitioners and researchers.

4.2.1 Smith and Agrawal Model

Smith and Agrawal (2000) (hereafter SA) study the assortment planning problem with the exogenous demand model. SA models the arrival process of customers carefully and updates the inventory levels after each customer arrival. Given assortment S, SA sets the stocking level of each product to achieve exogenously determined service levels f_j. Let $g_j(S,m)$ denote the probability that m^{th} customer chooses product j and $A_k(S,m)$ a binary variable indicating the availability of product k when the m^{th} customer arrived. Both clearly depend on the choice of previous customers and the number of substitution attempts made by the customer. For one substitution-attempt-only model,

$$g_j(S,m) = d_j + \sum_{k \notin S} d_k \alpha_{kj} + \sum_{k \in S \setminus \{j\}} d_k \alpha_{kj} (1 - A_k(S,m))$$

The first term is the original demand for product j, the second term is the demand from assortment substitution and the third from stockout substitution. Since exactly determining $g_j(S,m)$ is complex, SA develops lower and upper bounds. The lower bound is achieved by considering only assortment-based substitution and the upper bound by assuming that products achieve f_j in-stock probability even for the first customer, hence overestimating stockout substitution. Specifically,

$$\begin{aligned} &h_j(S) \le g_j(S,m) \le H_j(S) \quad \text{for all} \quad m, \text{where} \\ &h_j(S) = d_j + \sum_{k \notin S} d_k \alpha_{kj}, \\ &H_j(S) = d_j + \sum_{k \notin S} d_k \alpha_{kj} + \sum_{k \in S \setminus \{j\}} d_k \alpha_{kj} f_k. \end{aligned} \tag{8.4}$$

SA shows that these bounds are tight and uses the lower bound $h_j(S)$ to approximate the demand rate. That is, effective demand for product j given assortment S follows a distribution with mean $h_j(S)$. SA provides similar bounds to the demand rate under the repeated-attempts substitution model. Agrawal and Smith (1996) found that

Negative Binomial distribution (NBD) fits retail sales data very well. SA shows that when the total number of customers that visit a store is distributed with NBD, the demand for each product would also follow NBD.

The optimization problem is to maximize the total category profits:

$$\max_{S \subset N} Z = \sum_{j \in S} \pi_j(S)$$

where the profit function for each product j is the newsvendor profit minus the fixed cost of stocking an item V_j.

$$\pi_j(S) = \left(r_j - c_j\right) h_j(S) - c_j E[x_j - D_j | h_j(S)]^+ - \left(r_j - c_j\right) E[D_j - x_j | h_j(S)]^+ - V_j,$$

where D_j is the random variable representing the demand for product j, x_j is the optimal newsvendor stocking quantity to achieve the target stocking level $f_j = 1 - c_j/r_j$, e.g., $\Pr\{D_j \geq x_j \,|\, h_j(S)\} = f_j$ for a continuous demand distribution. Incorporating salvage value, or holding costs to the newsvendor profit function above is trivial.

This optimization problem is a nonlinear integer programming problem. SA proposes solving the problem via enumeration for small n and a linearization approximation for large n. A single constraint such as a shelf space or a budget constraint can be incorporated into the optimization model. SA proposes a Lagrangian Relaxation approach followed by a one-dimensional search on the dual variable for the resulting mathematical program.

Several insights are obtained from illustrative examples. Substitution effects reduce the optimal assortment size when fixed costs are present. However, even when there are no fixed costs present, substitution effects can reduce the optimal assortment size, because products have different margins. Contrary to the main result of van Ryzin and Mahajan (1999), it may not be optimal to stock the most popular item—a result of the adjacent substitution matrix or the one-item substitution matrix.

4.2.2 Kök and Fisher Model

The methodology described in Kök and Fisher (2007) is applied at Albert Heijn, BV, a leading supermarket chain in the Netherlands with 1,187 stores and about $10 billion in sales. The replenishment system at Albert Heijn is typical in the grocery industry. All the products in a category are subject to the same delivery schedule and fixed leadtime. There is no backroom, therefore orders are directly delivered to the shelves. Shelves are divided into *facings*. SKUs in a category share the same shelf area but not the same facing, i.e., only one kind of SKU can be put in a facing. Capacity of a facing depends on the depth of the shelf and the physical size of a unit of the SKU. The inventory model is a periodic review model with stochastic

demand, lost sales and positive constant delivery lead-time. The number of facings allocated to product j, f_j, determines its maximum level of inventory, $k_j f_j$, where k_j is the capacity of a facing. At the beginning of each period, an integral number of case packs (batches) of size b_j is ordered to take the inventory position as close as possible to the maximum inventory level without exceeding it. Case sizes vary significantly across products and significantly affect returns from inventory. The performance measure is gross profit, which is per-unit margin times sales minus selling price times disposed inventory.

We focus on a single subcategory of products initially for expositional simplicity and then explain how to incorporate the interactions between multiple subcategories. The decision process involves allocating a discrete number of facings to each product in order to maximize total expected gross profits subject to a shelf space constraint:

$$\begin{aligned} \max_{f_j, j\in N} Z(\mathbf{f}) &= \sum_j G_j(f_j, D_j(\mathbf{f}, \mathbf{d})) \\ s.t. \quad & \sum_j f_j w_j \le ShelfSpace_{AP} \\ & f_j \in \{0, 1, 2, ..\}, \quad \text{for all } j \end{aligned} \tag{8.5}$$

where f_j is the number of facings allocated to product j, and w_j is the width of a facing of product j. G_j is the (long run) average gross profit from product j given f_j and demand rate D_j. Due to substitution, effective demand for a product includes the original demand for the product and substitution demand from other products. Hence, $D_j(\mathbf{f}, \mathbf{d})$, the effective demand rate of product j, depends on the facing allocation and the demand rates of all products in the subcategory, i.e., $\mathbf{f} = (f_1, f_2, \ldots, f_n)$ and $\mathbf{d} = (d_1, d_2, \ldots, d_n)$, where d_j is the original demand rate of product j (i.e., number of customers who would select j as their first choice if presented with all products in N). The store's assortment is denoted S and is determined by the facing allocation, i.e., $S = \{j \in N : f_j > 0\}$.

Similar to SA, the effective demand rate function under this substitution model is

$$D_j(\mathbf{f}, \mathbf{d}) = d_j + \left(\sum_{k:f_k=0} \alpha_{kj} d_k + \sum_{k:f_k>0} \alpha_{kj} L_k(f_k, d_k) \right) \tag{8.6}$$

where the L_k function is the lost sales (average unmet demand) of product k. In our application we estimate $L_k(f_k, d_k)$ via simulation. In (8.6), $\sum_{k:f_k=0} \alpha_{kj} d_k$ is the demand for j due to assortment-based substitution and $\sum_{k:f_k>0} \alpha_{kj} L_k(d_k, f_k)$ is the demand for j due to stockout-based substitution.

In a stochastic inventory model as described above, G_j is a nonlinear function of the allocated facings to product j. It is a function of the facings of product j (f_j), and the facings of all other SKUs in a subcategory through the D_j function. Hence, AP is a knapsack problem with a nonlinear and nonseparable objective function, whose coefficients need to be calculated for every combination of the decision variables. Even if we rule out stockout-based substitution, we need to consider 'in' and 'out' of the assortment values for all products leading to 2^n combinations.

We propose the following iterative heuristic that solves a series of separable problems. The details of the algorithm can be found in Kök and Fisher (2007). We set $D_j(\mathbf{f}, \mathbf{d}) = d_j$ for all j and solve (AP) with the original demand rates resulting in a particular facings allocation $\mathbf{f}^0$. At iteration t, we recompute $D_j(\mathbf{f}^{t-1}, \mathbf{d})$ given δ for all j according to Eq. (8.6). Note that $\sum_j G_j\left(f_j^t, D_j(\mathbf{f}^{t-1}, \mathbf{d})\right)$ is separable now, because $D_j(\mathbf{f}^{t-1}, \mathbf{d})$ are computed a priori. We then solve (AP) with $Z(\mathbf{f}^t) = \sum_j G_j\left(f_j^t, D_j(\mathbf{f}^{t-1}, \mathbf{d})\right)$ via a Greedy Heuristic. We keep iterating until f_j^t converges for all j. In a computational study, the Iterative Heuristic performs very well with an average optimality gap of 0.5 %.

(AP) can be generalized to multiple subcategories of products that share the same shelf space by including several subcategories in the summations in the objective function and the shelf space constraint. Let subscript $i = 1, \ldots, I$ be the subcategory index. The objective function in the multiple subcategory case would be $Z(\mathbf{f}) = \sum_i \sum_j G_{ij}(f_{ij}, D_{ij}(\mathbf{f}_i, \mathbf{d}_i))$, the shelf space constraint can be modified similarly.

Structural Properties of the Iterative Heuristic

The Iterative Heuristic is based on a Greedy Heuristic. Therefore we can find properties of the resulting solution by exploiting the way the Greedy Heuristic works. First we note that the gross profit function for a product depends on demand, margin and operational constraints. Demand level and per-unit margin affect the maximum gross profit a product can generate if sufficient inventory is held. Operational constraints, such as case-pack sizes and delivery leadtime affect the curvature of the gross profit function. For example, a product with a smaller case-pack (batch size) has a higher slope of the gross profit curve for low inventory levels, and therefore can achieve the maximum gross profit with less inventory. These observations lead to the following theorems taken from Kök and Fisher (2007).

Products A and B belong to a subcategory with substitution rate $\delta \geq 0$. They are nonperishable. They are subject to the replenishment system described at the beginning of this subsection. The leadtime is zero. Demand for both products follow the same family of probability distributions. Effective demand for product A (B) has a mean D_A (D_B) and coefficient of variation ρ_A (ρ_B). Unless otherwise stated, $d_A = d_B$, $\rho_A = \rho_B$, $r_A = r_B$, $c_A = c_B$, and $b_A = b_B = 1$.

Theorem 3 *Consider products A and B. Let* $\tilde{\mathbf{f}}$ *denote the vector of facing allocations for all products in the subcategory other than A and B. If exactly one of the following conditions is met,*

(i) *All else is equal and* $d_A > d_B$. *The demand distribution is one of Poisson, Exponential or Normal distribution.*
(ii) *All else is equal and* $r_A - c_A \geq r_B - c_B$.
(iii) $w_A \leq w_B$,

then $f_A \geq f_B$ *in the final solution of the Iterative Heuristic.*

The implications of the first part of this theorem is clear: an allocation algorithm based on demand rates should work fairly well when products are differentiated by demand rates only. This is similar to the property of optimal assortments in the unconstrained problem in van Ryzin and Mahajan (1999). However, the above theorem proves additional results that the product with higher margin, or lower space requirement should be given priority in the assortment.

Theorem 4 *Consider products A and B. Let* $\tilde{\mathbf{f}}$ *denote the vector of facing allocations for all products in the subcategory other than A and B. If exactly one of the following conditions is met,*

(i) *All else is equal and* $\rho_A < \rho_B$,
(ii) *All else is equal,* $b_A \geq 1$, *and* b_B *is an integer multiple of* b_A,

then, the following holds. In the final solution of the Iterative Heuristic, if product B is included in the assortment then so is A (i.e., $f_B > 0 \Rightarrow f_A > 0$).

Theorem 4 characterizes the impact of the operational characteristics of a product on the assortment choice. When one of the conditions of the Theorem 4 holds, i.e., when B has either a larger batch size or higher demand variability, due to limited shelf space, if A is not included in the assortment, neither is B. Since the maximum value of G_A is higher and the slope is higher for low inventory levels, the profit impact of first facing is higher for A, resulting in a higher rank in the ordered input list to the Greedy Heuristic. However, if both products are in the assortment, it is possible to have $f_B > f_A$ in the solution. The reason for this is that G_A reaches its maximum level quickly with the early facing allocations, whereas it takes more facings for B to reach its maximum. In such cases, allocation heuristics based on demand rates perform poorly. A reasonable rule of thumb based on these observations would be the following. First high demand rate products shall be included in the assortment, then more facings shall be allocated to the products that have more restrictive operational constraints.

We applied our estimation methodology (to be described in Sect. 5.2.2) and optimization methodology to the data from 37 stores and two categories. The categories include 34 subcategories or 234 SKUs. (AP) is solved for each category for a given category shelf space. The facing allocations for SKUs also determine the space allocation between subcategories. We compare the category gross profit of the recommended assortments with that of the current assortments at Albert Heijn.

The gross profits of the recommended system is 13.8 % higher than that of the current assortment. The financial impact of our methodology is a 52 % increase in pretax profits of Albert Heijn.

Other work on assortment planning with exogenous demand include Rajaram (2001). He develops a heuristic based on Lagrangian relaxation for the single period assortment planning problem in fashion retailing without consideration of substitution between products.

4.3 Assortment Planning Under Locational Choice

Gaur and Honhon (2006) study the assortment planning model under the locational choice demand model. The products in the category differ by a single characteristic that does not affect quality or price such as yogurt with different amounts of fat-content. The assortment carried by the retailer is represented by a vector of product specifications $(b_1, .., b_s)$ where s is the assortment size and $b_j \in [0, 1]$ denotes the location of product j. Each consumer is characterized by an ideal point in [0,1] and chooses the product that is closest to him or her. The coverage interval of product j is defined as the subinterval that contains the most preferred good of all consumers for whom the product yields a nonnegative utility. The first choice interval of product j is defined as the subinterval that contains the most preferred goods of all consumers who choose j as a first choice. To extend Lancaster's model to stochastic demand, the authors assume that customers arrive to the store according to a Poisson process and that the ideal points of consumers are independent and identically distributed with a continuous probability distribution on finite support [0,1]. Only unimodal distributions are considered, implying that there exists a unique most popular product, and that the density of consumers decreases as we move away from the most popular product.

The operational aspects of the problem are similar to the van Ryzin and Mahajan model reviewed in Sect. 4.1: all products are assumed to have identical costs and selling prices, there is a single selling period, inventory costs are derived from a newsvendor model: excess demand at the end of the period is lost and excess inventory is salvaged. The only difference is that there is a fixed cost associated with including a product in the assortment. This model is closely related to the marketing product line design models in the marketing literature and operations-marketing papers such as de Groote (1994).

Under static substitution (assortment-based substitution), a consumer chooses a first choice product given the assortment but without observing inventory levels and does not make a second choice if the first choice is not available. Under dynamic substitution, the consumer chooses a product (if any) among the available products. This is equivalent to choosing a first choice product from the assortment and then looking for the next best alternative (if any) if the first choice is not available. This is equivalent to stock-out based substitution with repeated attempts.

The paper characterizes the properties of the optimal solution under static substitution and develops approximations under dynamic substitution. We skip the details of the analysis and briefly discuss the results from this paper. The authors show that, under static substitution, the distance between products in the optimal assortment are large enough so that there is no substitution between them. The most popular product, the one that would be located at the mode of the distribution is not included in the assortment when the economies of scale enjoyed by the most popular product is overcome by the diseconomies of scale it created for the other products. This property contrasts with the property of the optimal assortments under the MNL model (Theorem 2). We believe that the difference is not because of the different choice model, but because the problem considered here is a product line design problem at its heart. The authors find that the retailer may choose not to cover the entire market due to fixed costs. An analogous result is obtained under the MNL model as well, but that is purely due to economies of scale created for more popular products by not including some products in the assortment. Whereas in this model, it is optimal to cover the entire market when fixed costs are not present.

The problem is more complex under the dynamic substitution problem, as it is under other demand models. The profits computed under the static substitution assumption provides a lower bound to the dynamic problem, since it does not capture the profits from repeated attempts of the stock-out based substitution. An upper bound is obtained by solving a relaxation of the problem. Namely, the retailer gets to observe the ideal points of all arriving customers before allocating inventory to customers to maximize the profits. This is similar to Bassok et al. (1999) where consumers do not directly choose a product, but they are assigned a product (if any) either according to an exogenous rule or the retailer's decisions. Clearly, the retailer can generate more profits by doing the allocation itself rather than following the choices of the customers arriving in a random process. The solutions to these bounds are also proposed as heuristic approaches. In a numerical study, the authors make the following observations. Both heuristics generate solutions that are 1.5 % within the optimal solution on average. This suggests that the static substitution solution, which is easier to obtain, would serve as a good approximation in most cases. Dynamic substitution has the greatest impact when demand is low, customer distribution in the attribute space is heterogenous, and consumers are willing to substitute more. The retailer provides higher variety under dynamic substitution than under static substitution and locates products closer to each other so that a consumer can derive positive utility from more than one product. The firm offers more acceptable alternatives to the customers whose ideal product is located in areas where consumer density is high.

There are other papers that formulate mathematical models for selecting optimal assortments when customer heterogeneity is represented by locational choice. McBride and Zufryden (1988) deal with manufacturer's product line selection which require specification of product attributes and Kohli and Sukumar (1990) deal with the retailer's problem of choosing an assortment from a set of products.

Alptekinoğlu et al. (2012) extend the Hotelling-Lancaster locational choice model for studying the assortment planning problem for a category of horizontally differentiated products. They assume that consumer preferences are distributed along a straight line, and the disutility costs due to substitution are asymmetric and convex with respect to distance. They show that when preferences follow a unimodal distribution, the prices and market share of the products drop with distance in respect to the product that covers the mode (or the most popular product). They show that their approach leads to exact solutions when consumer tastes are distributed discretely. For continuous distributions, they propose a shortest path formulation, which can be computed efficiently.

4.4 Assortment Planning in Decentralized Supply Chains

The assortment planning papers reviewed until this section are single location models. There has been some recent work exploring assortment planning issues in two-tier supply chains. Aydin and Hausman (2003) consider the assortment planning problem with MNL (i.e. the van Ryzin and Mahajan model) in a decentralized supply chain with one supplier and one retailer. They find that the retailer chooses a narrower assortment than the supply chain optimal assortment since her profit margins are lower than that of the centralized (vertically integrated) supply chain. The manufacturer can induce coordination by paying the retailer a per-product fee, resembling the slotting fees in the grocery industry, while making both parties more profitable.

Singh et al. (2005) study the effect of product variety on supply chain structures, building on the van Ryzin and Mahajan model. In the traditional channel, the retailers stock and own the inventory, whereas in the drop-shipping channel, the wholesaler stocks and owns the inventory and ships the products directly to customers after the customers place an order at a retailer. Drop-shipping is a common practice in internet retailing: it offers the benefits of risk pooling when there are multiple retailers, but retailers have to pay a per unit fee for drop-shipping. As a result, product variety in the drop-shipping channel is higher than the traditional channel when drop-shipping fees are low and number of retailers is large. The authors derive conditions on the parameters under which the retailers or the wholesaler or both prefer the drop-shipping channel. They also study a vertically integrated firm with multiple retailers and find that a hybrid supply chain structure may be optimal for some parameter combinations: the popular products are stocked at the retailer while the less popular products are stocked at the warehouse and drop-shipped to the customers. The assortment size at the retailer gets smaller as the number of retailers increase or the drop-shipping costs decrease.

Kurtulus and Toktay (2007) compare the traditional category management and category captainship in a setting with two products and deterministic demand under a shelf space constraint. In category captainship, one of the vendors is assigned as the category captain and the pricing and assortment decisions are delegated to her.

The argument for category captainship is that the leading manufacturer in a category may have more experience with the category and resources than the retailer. They find that the assortment may be narrower under category captainship, because the noncaptain brand may be priced out of the assortment. Kurtulus (2005) considers the impact of category captainship under three types of contracts in a setting similar to the van Ryzin and Mahajan model. While the resulting assortment is still in the popular assortment set under the target profit and target sales contracts, it is in the least popular assortment set under the target variety contract.

4.5 Dynamic Assortment Planning

All of the assortment planning papers reviewed in the previous sections consider static assortment planning problems and do not consider revising or changing assortment selection as time elapses. This makes sense for fashion and apparel retailers, because long development, procurement and production lead times constrain retailers to make assortment decisions in advance of the selling season. With limited ability to revise product assortments, academics and industry practitioners focused on optimizing the production quantities in order to delay the production of those products that have high demand uncertainty (e.g., Fisher and Raman 1996). However, innovative firms such as Zara (Spain), Mango (Spain), and World Co. (Japan) created highly responsive and flexible supply chains and cut the design-to-shelf lead time down to 2–5 weeks, as opposed to 6–9 months for a traditional retailer, which enabled them to make design and assortment selection decisions during the selling season. Raman et al. (2001) describes how such short response times are achieved at World Co. through process and organizational changes in the supply chain. Learning the fashion trends and responding with an updated product selection is most critical for these high fashion companies.

Allowing changes in the assortment during a single selling season introduces several new issues. The products put in the store this week can't be removed next week and hence condition the decisions this week; there may be costs associated with adding new products or dropping products from the assortment; it may be optimal to put products in the stores to learn about the demand, even if it isn't optimal to do so given the current knowledge.

Caro and Gallien (2007) formulate the dynamic assortment problem faced by these retailers: At the beginning of each period, the retailer decides which assortment should be offered and gathers demand data for the products carried in the assortment in each period. There is a budget constraint that limits the number of products offered in each period to K. Due to design-to-shelf lead time, an assortment decision can be implemented only after l periods. This problem relates to the classical exploration versus exploitation trade-off. The firm must decide whether to optimize revenues based on the current information (exploitation), or try to learn more about the demand of products not in the assortment with the hope of identifying popular products (exploration).

The authors make several assumptions for tractability. The demand for a product is independent of the demand or the availability of the other products (i.e., there is no substitution between products or correlation in demand). The demand rate for each product is constant throughout the season. There is a perfect inventory replenishment process, therefore there are no lost sales or economies of scale in the operating costs. More importantly, no products carry over from period to period, therefore it is feasible to change the assortment independent of the previous assortment. There are no switching costs. Some of these assumptions are relaxed later.

The demand for product $j \in N$ is from a stationary Poisson process throughout the season. The rate of arrival λ_j is unknown and actual demand is observed only when the product is included in the assortment. The retailer uses a Bayesian learning mechanism: he starts each period with a prior belief that λ_j is distributed according to a Gamma distribution with shape parameter m_j and scale parameter α_j. Suppose that product j is included in the assortment and observed demand is d_j. The prior distribution of λ_j is updated as $Gamma(m_j + d_j, \alpha_j + 1)$. The mean of this distribution is the average sales of product j throughout the periods it is carried. Let $\mathbf{f} = (f_1, \ldots, f_n)$ be a vector of binary variables indicating whether the product is in the assortment and F the set of feasible assortments, $F = \{\mathbf{f} : \sum_{j \in N} f_j \leq K\}$. Similarly, let $\mathbf{m}, \boldsymbol{\alpha}$, and $\mathbf{d}$ denote the vectors of m_j, α_j, d_j, respectively. Assume that assortment implementation leadtime l is zero.

The dynamic programming formulation is

$$J_t^*(\mathbf{m}, \alpha) = \max_{\mathbf{f} \in F} \sum_{j \in N} f_j r_j E[\lambda_j] + E J_{t+1}^*(\mathbf{m} + \mathbf{d} \cdot \mathbf{f}, \alpha + \mathbf{f}).$$

Since solution of this dynamic program can be computationally overwhelming, the authors propose a Lagrangian relaxation (of the constraint on the number of products in the assortment) and the decomposition of weakly coupled dynamic programs to develop an upper bound. Performance of two heuristics are compared. The index policy balances exploration by including high expected profit products and exploitation by including products with high demand variance in a single-period look ahead policy. The greedy heuristic selects in each period the K products with the highest expected profits. The index policy is near optimal when there is some prior data on demand available and outperforms the greedy heuristic especially with little prior information about demand or the leadtime. The paper then demonstrates that the heuristics perform well when there are assortment switching costs, demand substitution, and a positive implementation lag.

Another learning method that Zara and other high-fashion companies employ is learning the attributes of the high selling products. That is, if a certain color is hot this season, and products with a special fabric are selling relatively well, the prior distribution of the demand for a product with that fabric-color combination can be updated, even if the product were never included in the assortment before. The attribute-based estimation method by Fader and Hardie (1996) mentioned in

Sect. 5.1 can be instrumental in estimating the demand for new products in this setting.

Rusmevichientong et al. (2010) develop algorithms to compute the optimal assortment under multinomial logit demand and capacity constraints. They derive structural insights on the optimal assortment for the static case, and utilize it to develop an adaptive policy for the dynamic problem, where the algorithm learns demand parameters form past data and chooses the optimal assortment based on that. They find that their algorithm performs well on being applied to sales data from an online retailer.

Ulu et al. (2012) study the dynamic assortment problem under horizontal differentiation, when consumer preferences are distributed according to the locational choice model. They assume that the firm knows where customers are located, but is unaware of their probability distribution. They model the problem using a discrete-time dynamic program, where in each period the retailer chooses an assortment and set of prices to maximize expected profits over the entire horizon, and customers choose the utility maximizing product from the assortment. The retailer updates beliefs on the distribution of customers in a Bayesian fashion. Under this scenario, they show that it is possible to partially order assortments based on their information content. They demonstrate that it might be optimal for the retailer to alternate between exploration and exploitation, and sometimes offer sub optimal loss producing assortments in a bid to learn valuable information about consumer preferences.

Bernstein et al. (2011) present a novel model exploring dynamic assortment decisions in a setting with multiple heterogeneous customer segments. They show that rationing products to some customer segments may be optimal. This insight is different from those obtained in the revenue management literature, as the rationing outcome is not due to differences in costs or prices, but due to the interplay between heterogeneity in customer segments and limited inventories. They demonstrate the potential impact of assortment customization based on a real data set obtained from a large fashion retailer. They find that the revenue impact of assortment customization can be significant indicating its potential as another lever for revenue maximization in addition to pricing.

Saure and Zeevi (2013) consider the interesting case where a retailer tries to learn about consumer preferences by strategically offering different assortments. The main tradeoff facing the retailer is to balance the value of learning with the goal of maximizing revenues. They study a family of stylized assortment planning problems under this scenario, and develop a family of policies that balance this tradeoff. Their major finding is that the optimal policy limits experimentation with suboptimal products, thereby reducing the impact of experimentation on revenues.

4.6 Competitive Assortment Models

Cachon and Kök (2007) study the assortment planning problem with multiple merchandise categories and basket shopping consumers (i.e., consumers who desire to purchase from multiple categories). They present a duopoly model in which retailers choose prices and variety level in each category and consumers make their store choice between retail stores and a no-purchase alternative based on their utilities from each category. The common practice of category management (CM) is an example of a decentralized regime for controlling assortment because each category manager is responsible for maximizing his or her assigned category's profit. Alternatively, a retailer can make category decisions across the store with a centralized regime. They show that CM never finds the optimal solution and provides both less variety and higher prices than optimal. In a numerical study, they demonstrate that profit loss due to CM can be significant. Finally, they propose a decentralized regime that uses basket profits, a new metric, rather than accounting profits. Basket profits are easily evaluated using point-of-sale data, and the proposed method produces near-optimal solutions.

Hopp and Xu (2008) consider a static approximation of the assortment planning problem under stock-out substitution. They model demand using fluid networks and obtain a mapping between service and inventory, which allows them to analyze the previously intractable, joint assortment, inventory and pricing problem in both competitive and non-competitive scenarios. They show that the static approximation models the dynamic scenario very closely, and obtain several interesting structural insights under duopolistic competition. First, they find that under joint price and inventory competition, prices are lower, while demand and inventory levels are higher. Second, they observe that under joint price and assortment competition, prices and variety offered by each retailer are both lower. However, the total number of products and the aggregate inventory levels in a duopoly market and both higher than in a monopolistic market.

Kök and Xu (2011) study assortment planning and pricing for a product category with heterogeneous product types from two brands. They model consumer choice using the Nested Multinomial Logit framework with two different hierarchical structures: a brand-primary model in which consumers choose a brand first, then a product type in the chosen brand, and a type-primary model in which consumers choose a product type first, then a brand within that product type. They find that optimal (centralized) and competitive (decentralized between brands) assortments and prices have quite distinctive properties across different models. Specifically, with the brand-primary model, both the optimal and the competitive assortments for each brand consist of the most popular product types from the brand. They extend the structural properties of assortment decisions characterized by van Ryzin and Mahajan (1999) to the case of Nested Logit. Under the brand primary model, structure remains the same under competitive and centralized regimes. The type-primary choice model, however, leads to a structural difference: The optimal and the competitive assortments for each brand may not always consist of the most

popular product types of the brand. Instead, the overall assortment in the category consists of a set of most popular product types. Further, due to the combinatorial nature of the type-primary model, the existence of equilibrium may not be guaranteed. This paper also characterizes the optimal pricing of products. They find that a lower price should be charged for more popular product types due to economies of scale. Under competition, the brand with the higher market share would charge higher prices.

Besbes and Saure (2011) study the assortment problem under a duopoly, when consumers make their purchase decisions with full knowledge of the retailers' assortments. They show that when prices are exogenous, and the products carried by the retailers are exclusive, the number of equilibria are bounded, and the retailers always prefer the same equilibrium. When the assortments overlap, they show that an equilibrium may or may not exist, and the number of equilibria might increase exponentially with the number of products. Under the scenario of joint assortment and price competition, they show that at most one equilibrium exists. Finally, they demonstrate that competition leads to lower prices and expanded variety, as compared to a monopolistic setting.

Martínez-de-Albéniz and Roels (2011) consider shelf-space competition in a multi-supplier retail outlet. They find that when retailers allocate shelf space between products based on sales velocity and margins, and suppliers set wholesale prices to maximize the shelf space they are allocated, they tend to keep margins high. Moreover, the incentives of the two parties are misaligned, leading to suboptimal prices and shelf space allocations. Additionally, they find that the impact of suboptimal pricing far outweighs the effect of suboptimal shelf space allocation.

Kök and Martínez-de-Albéniz (2013) study the impact of quick response capabilities of supply chains on product variety in a competitive environment. In industries where customer needs quickly change, retailers such as Zara can postpone their assortment decisions (amount of variety, balance across categories) to close-to-season or in-season due to shorter design-to-shelf lead times. The authors study how assortment competition depends on the postponement capabilities of retailers. They develop a stylized model where two retailers choose their assortment breadth either before or after market characteristics are revealed. They find that slower retailers provide a higher variety and being fast is equivalent to offering 30–50 % more variety.

4.7 Assortment Planning Models with Multiple Categories/Stores

Although research has primarily focused on single category choice decisions, there is recent research that examines multiple category purchases in a single shopping occasion by modeling the dependency across multi-category items explicitly (see

Russell et al. 1997 for a review). Manchanda et al. (1999) find that two categories may co-occur in a consumer basket either due to their complementary nature (e.g., cake mix and frosting) or due to coincidence (e.g., similar purchase cycles or other unobserved factors). Bell and Lattin (1998) show that consumers make their store choice based on the total basket utility. Fixed costs for each store visit (e.g., search and travel costs) provide an intuitive explanation for why consumers basket shop. Bell et al. (1998) use market basket data to analyze consumer store choices and explicitly consider the roles of fixed and variable costs of shopping.

Baumol and Ide (1956) study the notion of right level of variety in a very stylized model. The retailer chooses N, the number of different product categories to offer. Consumer utility is increasing in variety, but decreasing in in-store search costs (which increases with N). Therefore for each consumer there is a range of N that makes the store attractive for shopping. The operating cost is the sum of inventory costs per category from an EOQ model and handling costs that is concave increasing in N. The resulting retailer profit function is not well-behaved, therefore profit maximizing level of variety is difficult to characterize and the insights from this model are fairly limited.

There are two papers that consider assortment planning with multiple categories in more detail. Agrawal and Smith (2003) extend the Smith and Agrawal (2000) model and the analysis described in Sect. 4.2.1 to the case where customers demand sets of products. Cachon and Kök (2007) compare the prices and variety levels in multiple categories under category management to the optimal variety levels in the presence of basket shopping consumers.

The modeling and solution approach in Agrawal and Smith (2003) is very similar to their earlier work. Each arriving customer demands a purchase set. If the initially preferred purchase set is not available, the customer may do one of the following: (a) substitute a smaller set that does not contain the missing item, (b) substitute a completely different purchase set, (c) not purchase anything. This behavior is governed by substitution probability matrices. The demand for each set considering the substitution demand from other sets is characterized as in Eq. (8.4). The profit maximization problem is formulated as a mathematical program. For a customer to purchase any set, all the items in the set have to be available. Therefore, the expected profit is much more sensitive to percentage of customers who purchase in sets, the average size of a purchase set, and the substitution structure and parameters. The following observations from numerical examples are quite interesting.

Profits under adjacent substitution structure is much higher than that under random substitution, because under adjacent substitution stocking every other set in the list would result in lower lost sales than that under random substitution. As the percentage of customers who purchases in sets increases (while keeping the total demand constant), the optimal assortment size increases (decreases) if the fixed cost of including a product is low (high). Profits increase with substitution rate δ. Finally, optimizing the category by disregarding the substitution and the purchase sets can result in considerably lower profits than optimal.

Cachon and Kök (2007) work with a stylized model to develop managerial insights regarding the assortment planning process in an environment with multiple categories. Consider two retailers X and Y that carry two categories of goods. Retailer r offers n_{rj} products and sets its margin p_{rj} in category j. The consumer choice model is based on a nested Multinomial Logit (MNL) framework. A consumer's utility from purchasing product i in category j at retailer r is $u_{rji} = v_{rji} - p_{rj} + \varepsilon$ where v_{rji} is the expected utility from the product less the unit cost of the product and ε is i.i.d with Gumbel distribution with zero mean. There are three types of consumers in the market that are characterized by the contents of their shopping baskets: type 1 consumers would like to buy a product in category 1 only, type 2 consumers would like to buy a product in category 2 only, type b consumers are basket shoppers and would like to buy a product from both categories. Consumers buy exactly one unit of one product in every category included in their basket.

The authors show that the choice probability of a non-basket shopper between retailers X, Y and a no-purchase alternative can be written using the nested MNL model as follows:

$$s_{rj} = \frac{A_{rj}}{A_{xj} + A_{yj} + Z_j} \quad \text{for} \quad r = x, y, \quad \text{and} \quad j = 1, 2,$$

where A_{rj} is the attractiveness function for each alternative (an aggregate function of price and variety level). Using the nested MNL results of Ben-Akiva and Lerman (1985), as described in Sect. 3.2, it can be expressed as

$$A_{rj} = e^{-p_{rj}} \sum_{i=1}^{n_{rj}} e^{v_{rji}}, \quad \text{for} \quad r = x, y.$$

Now, consider a basket-shopping consumer. A basket-shopping consumer chooses retailer r only if she prefers the assortment at r for both categories. As a result, the probability that a basket shopper chooses retailer r is

$$s_{rb} = s_{r1} s_{r2} \quad \text{for} \quad r = x, y. \tag{8.7}$$

This is a multiplicative basket-shopping model, as a retailer's share of basket shoppers is multiplicative in its share in each category. An additive model for this problem has been discussed in Kök (2003).

The common practice of category management (CM) is an example of a decentralized regime for controlling assortment because each category manager is charged with maximizing profit for his or her assigned category. Since basket shoppers' store choice decision depends on the prices and variety levels of other categories, one category's optimal decisions depends on the decisions of the other categories. Hence, a game theoretic situation arises. CM can be interpreted as an explicit non-cooperative game between the category managers, since each category manager is responsible exclusively for the profits of her own category.

Alternatively, it can be interpreted as an iterative application of single category planning where each category's variety level is optimized assuming all other assortment decisions for the retailer are fixed. Decentralized regimes such as CM are analytically manageable but they ignore (in their pure form) the impact of cross-category interactions. Centralized regimes account for these effects but it is extremely difficult, in practice, to design a model to account for all cross-category effects, to estimate its parameters with available data and solve it.

The authors show that if there are any basket shoppers, CM provides less variety and higher prices than centralized store management. CM can lead to poor decisions because the category manager does not sufficiently account for how his or her decisions influences total store traffic. These results hold both for a single retailer and in duopoly competition. Numerical examples demonstrate that the profit loss due to CM can be significant. The dominant strategy for each retailer is to switch to centralized management.

To address the potential problem with a decentralized approach to assortment planning, we propose a simple heuristic that retains decentralized decision making (category managers optimize their own categories' profit) but adjusts how profits are measured. To be specific, instead of using an accounting measure of a category's profit, the authors define a new measure called *basket profits*. Basket profits can be estimated using point-of-sale data. It enables CM to approximately measure the true marginal benefits of merchandising decisions and lead to near-optimal profits. This analytical approach is an attractive alternative relative to ad-hoc coordination across category managers.

Fisher and Vaidyanathan (2014), consider the assortment localization problem, of choosing assortments that can vary by store, subject to a maximum number of different assortments. They model a SKU as a set of attributes and also model possible substitutions when a customer's first choice is not in the assortment. estimate demand and substitution probabilities from sales history using maximum likelihood estimation. They apply maximum likelihood estimation to sales history of the SKUs currently carried by the retailer to estimate the demand for attribute levels and substitution probabilities, and from this, the demand for any potential SKU, including those not currently carried by the retailer. They develop several heuristics for choosing SKUs to be carried in an assortment, and apply this approach to optimize assortments for three real examples: snack cakes, tires and automotive appearance chemicals. A portion of their recommendations for tires and appearance chemicals were implemented and produced sales increases of 5.8 % and 3.6 % respectively, which are significant improvements relative to typical retailer annual comparable store revenue increases.

5 Demand Estimation

In this section, we briefly discuss the estimation of the demand models specified in Sect. 3. The estimation method depends on the type of data that is available.

5.1 Estimation of the MNL

5.1.1 With Panel Data

Starting with the seminal work of Guadagni and Little (1983), an enormous number of marketing papers estimated the parameters of the MNL model to understand the impact of marketing mix variables on demand. These papers use panel data in which the purchasing behavior of households over time are tracked by the use of store loyalty cards. Consider the purchase decision of the household that visited the store in time t. The systematic component of the utility u_{jt} is specified as a linear function of m independent variables including product specific intercepts, price, an attribute of product j, loyalty of the household to the brand of product j (measured as exponentially weighted average of binary variables indicating whether or not the household purchased this brand). Let $x_{jt} = (x_{jt1}, x_{jt2}, \ldots, x_{jtm})$ denote the vector of these attributes for the household's shopping trip at time t, S_t denote the assortment at time t including the no-purchase option, and $\beta = (\beta_1, .., \beta_m)$ denote the vector of common coefficients.

$$u_{jt} = \beta^T x_{jt}, \qquad j = 0, 1, .., n.$$

The outcome of the choice experiment by a household in time t is

$$y_{jt} = \begin{cases} 1, & \text{if product } j \text{ is chosen in time } t \\ 0, & \text{otherwise} \end{cases}$$

Given u_{jt} it is possible to compute the choice probabilities according to MNL formula (8.1). To obtain the maximum likelihood estimates (MLE) for the coefficients, we can write log of the likelihood function by multiplying the probability of observing the choice outcome across all t:

$$L(\beta) = \sum_t \sum_j y_{jt} \left(\beta^T x_{jt} - \ln \sum_{k \in S_t} e^{\beta^T x_{kt}} \right).$$

McFadden (1974) shows that the log-likelihood function is concave, therefore any nonlinear optimization technique can be used to find the MLE estimate of β. Fader and Hardie (1996) suggest the use of more of the product's attributes and dropping product-specific dummy variables in x_j in the estimation. They argue that this results in a more parsimonious estimation method as the number of coefficients to be estimated would not grow with number of products but with number of significant characteristics. Moreover, this approach enables estimation of the demand for new products.

Extensions of this model such as Chiang (1991), Bucklin and Gupta (1992), and Chintagunta (1993) also investigate whether to buy, and how much to buy decisions of households. In these papers, the whether-to-buy decision is modeled as a binary

choice between the no-purchase alternative and the resulting utility from the product choice and quantity decisions in a nested way. Chong et al. (2001) extend the classical Guadagni and Little (1983) model using a nested MNL model, including three new brand-width measures that capture the similarities and the differences among products within and across brands.

Multiplicative Competitive Interactions (MCI) model offers a viable alternative to MNL. Although less popular than MNL, it is used in the marketing area to study market share games (e.g. Gruca and Sudharshan 1991) and it has empirical support. See Cooper and Nakanishi (1988) for a detailed discussion and estimation methods.

5.1.2 With Sales Transaction Data

Consider the demand process in the van Ryzin and Mahajan model, where consumer arrivals follow a Poisson process with rate λ and consumers select an alternative based on the MNL model. Our goal is to estimate λ and β from sales data. Sales transactions are the records of the purchasing time and the product choice for each customer who made a purchase. This is an incomplete data set in the sense that only the arrivals of customers who made a purchase are recorded. Define a period as a very small time interval such that the probability of having more than one customer arrival in a period is zero. Let t denote the index of periods. There is a sales record for a period only if a purchase is made in that period. It is impossible to distinguish a period without an arrival, from a period in which there was an arrival but the customer did not purchase anything. Therefore, the approach described above cannot be used.

The *Expectation-Maximization* (EM) algorithm is the most widely used method to correct for missing data. Proposed by Dempster et al. (1977), the EM method uses the complete-likelihood function in an iterative algorithm. Talluri and van Ryzin (2004) describe an estimation approach based on this method in the context of airline revenue management, but the algorithm is applicable to the retail setting described in Sect. 4.1. Let P denote the set of periods that there has not been a purchase made and $a_t = 1$ if there has been a customer arrival in period t. The unknown data is $(a_t)_{t \in P}$. We start with arbitrary (λ, β). The E-step replaces the incomplete data with their estimates. That is, we find the expectation of a_t for all $t \in P$ given the current estimates (λ, β). The M-step maximizes the complete-data likelihood function to obtain new estimates. The likelihood function is similar to that in the previous subsection, but includes the arrival probabilities λ. The procedure is repeated until the parameter estimates converge. Greene (1997) shows that the procedure converges under fairly weak conditions. If the expected log-likelihood function is continuous in the parameters, Wu (1983) shows that the limiting value of the procedure would be a stationary point of the incomplete-data log-likelihood function. The advantage of the procedure is that maximizing the complete-data likelihood function is much easier than maximizing an incomplete-data likelihood function.

Musalem et al. (2010) use store-level data and partial information on product availability to estimate consumer demand under stock-out based substitution. They develop a structural demand model that simulates the effect of stock-outs using a time-varying set of available alternatives, and is able to capture very flexible substitution patterns. They demonstrate how their model can be used to quantify lost sales and provide insights on the financial consequences of stock-outs. Finally, they suggest how price promotions can be used effectively to counter some of the negative economic impact.

Vulcano et al. (2012) focus on the problem of estimating demand model when only sales transaction data are available. They model demand by combining a poisson arrival process with a multinomial choice process. Instead of estimating the arrival and choice parameters simultaneously by maximizing an intractable likelihood function, they treat observed demand as incomplete realizations of primary demand, and utilize an Expectation-Maximization approach to develop simple and efficient algorithms to estimate the model parameters. They test the utility of their approach on one simulated and two industry data sets.

Jain et al. (2014) consider how sales transaction timing data can lead to better demand estimates. They find that the optimal order quantity is higher when the retailer takes into account actual stock-out times, as compared to the case where demand is fully observed. However, in most cases, where the demand uncertainty is high, and the margins are low, the extent of over-ordering with timing data tends to be lower than that with only stock-out event data. They demonstrate using numerical simulations, that the use of stock-out timing data reduces the loss in expected profits by 74.8 % as compared to the case where only stock-out events are observed.

5.1.3 With Sales Summary Data

The information available in sales data is different from the panel data in several ways, hence requires a different approach. One possibility is the approach in Kök and Fisher (2007), which will be described here. The data typically available for estimating the parameters of a demand model includes the number of customers visiting each store on a given day, sales for each product-store-day, as well as the values of variables that influence demand such as weather, holidays, and marketing variables like price and promotion. At Albert Heijn, the data set included SKU-day-store level sales data through a period of 20 weeks for seven merchandise categories from 37 Albert Heijn stores. For each store-day, the number of customers visiting the store is recorded. For each SKU-day-store, sales data comprised of the number of units sold, the number of customers that bought that product, selling price, and whether the product is on promotion or not. In addition, we have daily weather data and a calendar of holidays (e.g., Christmas week, Easter, etc.). The categories are cereals, bread spreads, butter and margarine, canned fruits, canned vegetables, cookies, and banquet sweets. There were 114 subcategories in these seven categories. The size of subcategories varies from 1 to 29 SKUs, with an average of 7.7 and a standard deviation of 5.7.

The model of consumer purchase behavior is based on three decisions: (1) whether or not to buy from a subcategory (*purchase-incidence*), (2) which variant to buy (*choice*) given purchase incidence, and (3) how many units to buy (*quantity*).[1] This hierarchical model is quite standard in the marketing literature and commonly used with panel data.

The demand for product j is

$$D_j = K(PQ)_j = K\pi p_j q_j \tag{8.8}$$

where K is the number of customers that visit the store at a given day, $(PQ)_j$ is the average demand for product j per customer, π is the probability of purchase incidence (i.e., the probability that a customer visiting the store buys anything from the subcategory of interest), p_j is the choice probability (i.e., the probability that variant j is chosen by a customer given purchase incidence), and q_j is the average quantity of units that a customer buys given purchase incidence and choice of product j.

The purchase incidence is modeled as a binary choice:

$$\pi = \frac{e^v}{1+e^v} \tag{8.9}$$

where v is the expected utility from the subcategory that depends on the demand drivers in the subcategory.

The product choice is modeled with the Multinomial Logit framework, where p_j are given by (8.1). The average utility of product j to a customer, u_j, is assumed to be a function of product characteristics, marketing and environmental variables.

Let subscript h denote store index, and t denote time index (i.e., day of the observation).

We compute p_{jht} from the sales data as the ratio of number of customers that bought product j to number of the customers that bought any product in the subcategory at store h on day t. At Albert Heijn, price and promotion are the variables influencing u_j. We fit an ordinary linear regression to the log-centered transformation of (8.1) (see Cooper and Nakanishi 1988 for details) to estimate $\delta_j^C, \alpha_1^C, \alpha_2^C,$ and $\theta_k^C, k = 1, \ldots, n$.

$$\ln\left(\frac{p_{jht}}{\overline{p}_{ht}}\right) = u_j = \delta_j^C + \sum_{k \in N} \theta_k^C I_{jk} + \alpha_1^C \left(R_{jht} - \overline{R}_{ht}\right) + \alpha_2^C \left(A_{jht} - \overline{A}_{ht}\right), \quad \text{for all } j \in S \tag{8.10}$$

[1] This hierarchical model of choice is similar to Bucklin and Gupta (1992) that models the first two decisions with an additional focus on the segmentation of customers and Chintagunta (1993) that models all three decisions. Both papers work with household panel data, whereas we work with daily sales data.

where $\overline{p}_{ht} = \left(\prod_{j \in S} p_{jht}\right)^{1/|S|}$, $I_{jk} = \{1, \text{if } j = k; 0 \text{ otherwise}\}$, R is price, $\overline{R}$ is average price in the subcategory, $A_{jht} = \{1, \text{if product } j \text{ is on promotion on day } t \text{ at store } h; 0, \text{otherwise}\}$, and $\overline{A}$ is average promotion level in the subcategory. It is straightforward to incorporate variables other than price and promotion into this approach.

We compute π_{ht}, the probability of purchase-incidence for the subcategory, from sales data as the ratio of number of customers who bought any product in S to the number of customers visited the store h on day t. We use the following logistic regression equation to estimate $\alpha_0^\pi, \alpha_1^\pi, \alpha_2^\pi, \alpha_{4t}^\pi, \gamma_k^\pi,\ k = 1, ..6,$ and $\beta_l^\pi, l = 1, .., 14$ in (8.11).

$$\ln\left(\frac{\pi_{ht}}{1-\pi_{ht}}\right) = v = \alpha_0^\pi + \alpha_1^\pi T_t + \alpha_2^\pi HDI_t + \sum_{k=1}^{6} \gamma_k^\pi D_t^k + \alpha_{4t}^\pi \overline{A}_{ht} + \sum_{l=1}^{14} \beta_l^\pi E_t^l \quad (8.11)$$

where T is the weather temperature, HDI (Human Discomfort Index) is a combination of hours of sunshine and humidity, D^k are day of the week 0–1 dummies and E^l are holiday 0–1 dummies for Christmas, Easter, etc. Other variables could be used appropriately in a different context.

We compute q_{jht} from sales data as the number of units of product j sold divided by the number of customers who bought product j at store h on day t and use linear regression to estimate $\alpha_{0j}^Q, \alpha_{1j}^Q, \alpha_{2j}^Q,$ and $\beta_{jl}^Q, l = 1, \ldots, 14$ in (8.12).

$$q_{jht} = \alpha_{0j}^Q + \alpha_{1j}^Q A_{jht} + \alpha_{2j}^Q HDI_t + \sum_{l=1}^{14} \beta_{jl}^Q E_t^l, \quad \text{for all } j \in S \quad (8.12)$$

In the grocery industry, K_{ht}, the daily number of customers who made transactions in store h on day t is a good proxy for the number of customers who visited the store. We use log-linear regression to estimate $\alpha_{0h}^K, \alpha_{1h}^K, \alpha_{2h}^K, \gamma_k^K, k = 1, .., 6,$ and $\beta_{1l}^K, l = 1, \ldots, 14$ in (8.13).

$$\ln(K_{ht}) = \alpha_{0h}^K + \alpha_{1h}^K T_t + \alpha_{2h}^K HDI_t + \sum_{k=1}^{6} \gamma_k^K D_t^k + \sum_{l=1}^{14} \beta_{1l}^K E_t^l \quad (8.13)$$

This four stage model of demand estimation has been tested for quality of fit and prediction for multiple stores and subcategories. The average of mean absolute deviation (MAD) across all products, subcategories and stores is 67 % in the fit sample and 74 % in the test sample. Average bias of our approach is 0 % and −9 % in fit and test samples, respectively. The current method used at Albert Heijn is estimating $(PQ)_j$ for each SKU directly via logistic regression with similar explanatory variables. The MAD of this method is 72 % and 94 % and average bias is −43 % and −30 % in the fit and test samples, respectively.

5.2 *Estimation of Substitution Rates in Exogenous Demand Models*

5.2.1 Estimation of Stockout-Based Substitution

Anupindi et al. (1998) estimate the demand for two products and the substitution rates between them using data from vending machines. They assume that consumers arrive according to a Poisson process with rate λ and choose product A (B) as their first choice product with probability p_A (p_B) and substitute according to an asymmetric substitution matrix $\begin{bmatrix} 0 & \alpha_{AB} \\ \alpha_{BA} & 0 \end{bmatrix}$. The demand for product A when B is not available is Poisson with rate $\lambda(p_A + p_B\alpha_{BA})$.

They consider two information scenarios. In the first one, so-called perpetual inventory data, each sales transaction and the exact time that each product runs out of stock (if they do) is observed. In this case, it is not difficult to write down the log-likelihood function and maximize it to obtain the MLE estimates. They show that the timing of the stockouts and the sales volume before and after those times are sufficient statistics. Therefore, it is not necessary to trace each sales transaction. This result of course would not hold if the arrival process were a nonstationary process.

In the second information scenario, so-called periodic review data, the stockout times of the products are not observed, but whether or not they are in-stock at the time of replenishment is known. We encounter an incomplete data problem, and again we can use the EM algorithm briefly discussed in Sect. 5.1.2 to correct for the missing data (i.e., the stockout times). To be able to generalize the methodology to more than two products, it is necessary to make further assumptions. The authors restrict the substitution behavior to a single-attempt model, i.e., no repeated attempts are allowed and they estimate the parameters for a problem with six products. Their results show that naive demand estimation based on sales data is biased, even for items that rarely stockout. They also find significant differences in the substitution rates of the six brands.

Anupindi et al. (1998) estimate stationary demand rates (i.e., do not consider a choice process) and a substitution matrix. Talluri and van Ryzin (2004) estimate demand rate and the parameters of the MNL choice model (λ, β) but do not consider a substitution matrix. Kök and Fisher (2007) generalize these two approaches and propose a procedure that simultaneously estimates the parameters of the MNL model, on which the consumer's original choice is based, and a general substitution probability matrix.

5.2.2 Estimation of Assortment-Based Substitution

Some retailers do not track inventory data. Some others do, but there is empirical evidence that the inventory data may not be accurate (e.g. DeHoratius and Raman 2008). Hence, sales data may be the only source of information in some cases. Here we review the methodology proposed by Kök and Fisher (2007) to estimate substitution rates using sales data. We assume that substitution structure (i.e., the type of the matrix) is known, and we only need to estimate the substitution rate δ. We demonstrate the method for the proportional substitution matrix, that is assume α_{kj} is given by (8.2).

The methodology can be explained briefly as follows. Suppose that a store carries assortment $S \subset N$ with 100 % service rate (i.e., no stockout-based substitution takes place). We observe D_j for products $j \in S$ from sales data. Notice that at a store that has full assortment (i.e., $S = N$), no substitution takes place, hence $D_j = d_j$ for all j. We can therefore estimate d_j for $j \in N$ from sales data of a similar store that carries a full assortment. We can conclude that the substitution rate is positive for this subcategory if $\Sigma_{j \in S} D_j > \Sigma_{j \in S} d_j$. Let $y(S) = \Sigma_{j \in S} D_j$. Given $\mathbf{d}$, substitution rate δ, and assortment S, we compute what each product in S would have sold at this store using Eq. (8.6), and the total subcategory sales denoted (S, δ). The error associated with a given δ is the difference between the observed and theoretical subcategory sales at a store [i.e., $y(S) - \widehat{y}(S, \delta)$]. We find the substitution rate δ that minimizes the total error across all available data from multiple stores and different time periods. The details of the procedure can be found in the paper.

As Campo et al. (2004) point out, there are significant similarities in consumer reactions to a permanent assortment reduction and to stockouts. Therefore, the substitution rate estimated for assortment based substitution can be also used for stockout-based substitution if that cannot be estimated. Another advantage of this methodology is that it enables us to estimate the demand rates of products in a store including those that have never been carried in that particular store.

The next step after the estimation of the substitution rate is the computation of the true demand rates. This involves two tasks. (a) deflating the demand rate of the variants already in the assortment S_h, and (b) estimating a positive demand rate for the variants that are not in S_h. Clearly, if $S_h = N$, no computation is necessary. Figure 8.1 presents an example of observed demand rates and the computed true demand rates for a subcategory with ten products.

5.3 *Estimation of Non-parametric Choice Models*

Farias et al. (2013) study the problem of modeling consumer choice, when the amount of data available is limited. They show that optimizing the assortment based on a mis-specified choice model can lead to highly suboptimal revenues. They consider a generic consumer choice model, where choices are modelled as

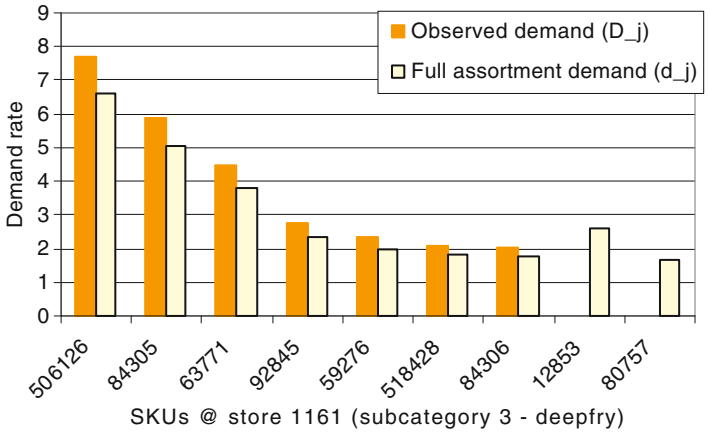


Fig. 8.1 Estimates of observed and original demand rates for a subcategory

distributions over preference lists. They develop a non-parametric approach to learn the right choice model, using limited data on customer purchase decisions. They apply their method on a real data set consisting of automobile sales transactions from a major US automaker, and show that it leads to a 20 % improvement in prediction accuracy over other state-of-the-art models, which results in a 10 % increase in revenues. They addresses the crucial issue of choice model identification, which is key to optimizing the assortment.

van Ryzin and Vulcano (2013) extend their previous work to estimate demand for a set of substitutable products using readily-available sales transactions and product availability data. They model demand as consisting of bernoulli arrivals followed by a general, non-parametric discrete choice model, that is compatible with an arbitrary random utility model. They apply the EM algorithm to jointly estimate the arrival rates and the probability distribution of customer choices. They use numerical experiments to demonstrate that their approach allows them to rapidly identify customer types and produce good estimates of demand.

6 Assortment Planning in Practice

The goal of this section is to describe assortment planning practice as illustrated by the processes used by a few retailers with whom we have interacted: Best Buy, Borders Books, Tanishq and Albert Heijn (Levy and Weitz 2004, Chapter 12), also provides a description of retail assortment planning.

6.1 Best Buy

Most retailers divide their products into various segments, usually called categories and sub categories. The assortment planning process begins by forecasting the sales of each segment for a future planning period ranging from a several month season to a fiscal year. Then scarce store shelf space and inventory purchase dollars are allocated to each segment based in part on the sales projections. Finally, given these resource allocations, the number of SKUs to be carried in each segment is chosen. As such, assortment planning in practice is essentially a strategic planning and capital budgeting process.

Best Buy offers a good example of this process. In their planning process, conventional still cameras and digital still cameras are two of the product segments. The starting point for a forecast of next year's sales is last year's sales adjusted for trend. Figure 8.2 shows sales of digital and traditional cameras through 2002. A logical forecast for 2003 would be less than 2002 sales for traditional cameras and more than 2002 sales for digital cameras.

The forecasts based on sales history are then adjusted based on information from trade shows, vendors, observations of competitor moves and reviews of new technology. The goal of assimilating these inputs is to identify changes in sales for a product category that might not be apparent from a straight forward extrapolation of sales history.

The next step is to set goals for each segment for sales, margin and market share based on the sales forecast, to allocate shelf space and inventory purchase dollars and then to determine how many SKUs to carry in each product segment. A critical input in deciding how many SKUs to carry is the importance to the customer of a

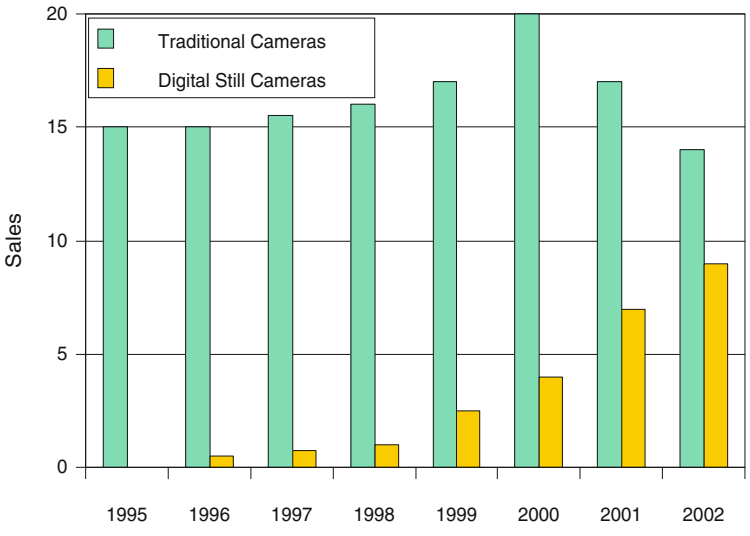


Fig. 8.2 Historical sales of traditional and digital cameras

Category	Promo	Labor	Impulse	Price	Selection
Computer	High	High	Low	High	Medium
Refrigerator	Medium	High	Low	Medium	High
Accessories	Low	Low	High	Low	Low
Movies	High	Med	High	High	High

Fig. 8.3 The impact of sales drivers for various types of products

broad selection in a particular category. Figure 8.3 was created by Best Buy to show the factors that influence sales and the importance of these factors for different types of products. For example, an accessory item such as a surge protector is often an impulse buy whose sales would be significantly increased by placing it on display near the check out register or in some other high traffic area. However, the customer is not particularly sensitive to price and doesn't require a broad selection. By contrast, placing a refrigerator next to the cash register to drive sales would be silly, because this isn't an impulse purchase for customers. However, they do value a broad selection and low prices. Another way of interpreting the data in this table is that Best Buy believes customers shopping for accessories are very willing to substitute if they don't find exactly what they are looking for, but refrigerator and movie customers are relatively unwilling to substitute.

This matrix is used to guide the number of SKUs to be carried in each product category. Other things being equal, a greater number of SKUs would be carried for those products where selection has a high impact on sales.

Once the number of SKUs to be carried in a product segment has been determined, it is left to the buyer for that segment to determine exactly which SKUs to carry. As an example, in flat panel TV's, Best Buy might carry 82 different SKUs. By contrast, the number of potential SKUs is much larger, comprising of eight diagonal widths (e.g. $19'', 25'', 32'', 35'', 40''$, etc.), five screen types (plasma, LCD, projection, etc.), seven resolutions (analog, 480i, 720p, 1080i, etc.) and nine major vendors (Sony, Panasonic, Pioneer, etc.) for a total of $8 \times 5 \times 7 \times 9 = 2,520$ potential SKUs. It is left to the buyer through a largely manual process to determine which 82 out of these 2,520 SKUs will be carried by Best Buy. The buyer incorporates a number of factors into the choice of SKUs. For example, it is highly desirable to carry products from several vendors so that Best Buy can benefit from competition when negotiating with vendors on price.

The Best Buy example suggests that practice and academic research are complementary, in that practice ends with delegating to the buyer the decision of which products to carry from the universe, and this is precisely the problem that has been emphasized in the academic literature.

6.2 *Borders*

Two interrelated issues in assortment planning are the division of decision rights between corporate and stores and the degree to which the assortment varies by store. Figure 8.4 below depicts alternatives of these two factors.

By far the most common approach is for corporate headquarters to decide on a single common assortment that is carried by all stores of the chain, except that in smaller stores, the breadth of the assortment may be reduced by removing some of the least important SKUs. A relatively small number of retailers (Bed Bath & Beyond would be an example) allow their store managers considerable authority in deciding which SKUs to carry in their stores. Usually, a portion of the assortment is dictated by corporate, and the remainder is chosen by store management from a corporate approved list of options. Obviously a result of this approach is that the assortment is different in all stores, and is hopefully tuned to the tastes of that store's customers.

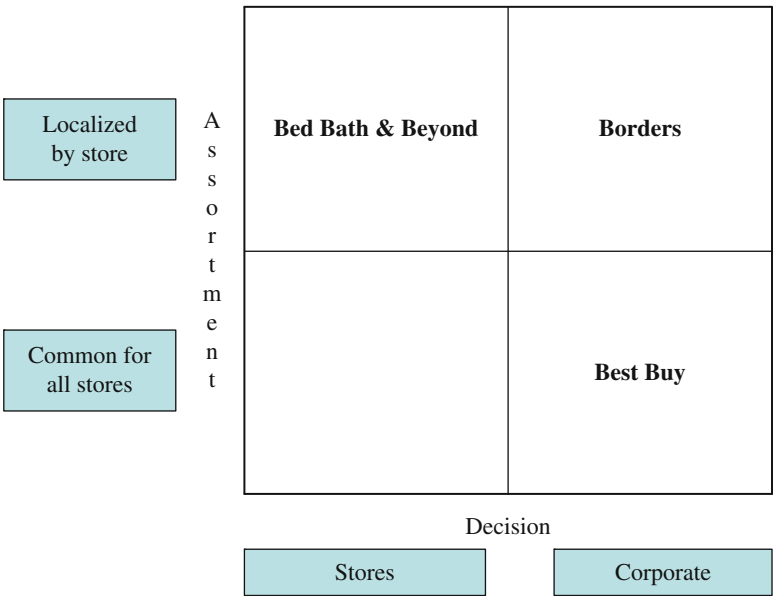


Fig. 8.4 Approaches to assortment planning

Borders Books is one of the few retailers that have developed a central approach to creating a unique assortment for each store. They segment their products into about 1,000 book categories and define the assortment at a store by the number of titles carried in each category. To choose these parameters they rely on a measure called Relative Sales per Title (RST) that equals the sales in a category over some history period divided by the number of titles carried in the category over the same period. If RST is high for a store-category in a recent period, then they increase the number of titles in that category, and conversely, reduce the titles in low RST categories. For example, a rule could be to divide the 1,000 categories in a store into the upper, middle and lower third of RST values and then increase number of titles carried in upper third by Δ and reduce lower third by Δ, where Δ and the frequency of adjusting the assortment are parameters of the process that determine how quickly and aggressively the assortment is adjusted based on history. Their overall process also takes seasonality into account, but that is outside the scope of this survey article.

6.3 Tanishq

Tanishq, a division of Titan Industries Ltd. (India's largest watch maker) is India's leading branded jewelry manufacturer and retailer in the country's $10 billion jewelry market. Tanishq jewelry is sold exclusively through a company controlled retail chain with over 60 boutiques spread over 39 cities. This network of boutiques is supplied and supported by a strong distribution network.

Assortment planning is a key activity at Tanishq involving significant challenges. First, jewelry is a complex product category with a very broad offering to choose from (more than 30,000 active SKUs) making assortment selection non-trivial. Second, given the small to medium size of most of the retail outlets, there were inventory limitations; as a consequence, getting the assortment decision right was critical. Significant differences in customer profile across its 60 boutiques and the frequent introduction of new products added further layers of complexity to the assortment planning process.

Traditionally, each store placed its own order, subject to guidelines on total inventory drawn up by the supply chain team at the corporate headquarters. This was done since the store associates were the ones closest to the customers and hence believed to have the best understanding of their preferences. This was true to a large extent, as the jewelry buying process in the Indian market was highly interactive, with store associates playing a significant role in guiding the customer through the product offerings based on their preferences (e.g. price range, design). Consequently, the store associates had a fairly accurate knowledge of customer choices, their willingness to substitute across product attributes, and reasons that led them to reject certain product variants.

However, there were issues with this model. First, store associates were already burdened with monthly sales targets and hence had little time to do full justice to the

ordering process. Second, their knowledge was limited only to product variants that the store had stocked in the past. Hence, they were missing out on potential product opportunities. This necessitated the need to modify the existing assortment planning process and address those shortcomings.

Tanishq accomplished this by moving from a store-centric model to a hybrid model involving both the store associates and a central supply chain team. The supply chain team at the corporate headquarters had the best access to sales and inventory data from all stores. They had detailed information about market trends and were in the best position to analyze historical data to detect selling patterns, and best selling variants at the state, regional, and national levels. This, combined with the local, store specific knowledge of the store associates, resulted in a more refined process for Tanishq.

The first step was the identification of product attributes relevant to the customers' choice process. This was done by the central supply chain team, based on inputs from the store associates. For example, the product category of rings was defined by the following attributes: theme, collection, design, gem type and size.

The next step was the determination of an appropriate assortment strategy for each product category. Again, this was carried out by the central supply chain team. They analyzed historical sales and inventory data in order to understand differences in sales mix across stores by attribute, to identify best sellers, and to develop an understanding of basic selling patterns.

The assortment strategy for each product category was developed based on a simple 2×2 matrix of percent contribution to sales vs. sales velocity (see Fig. 8.5).

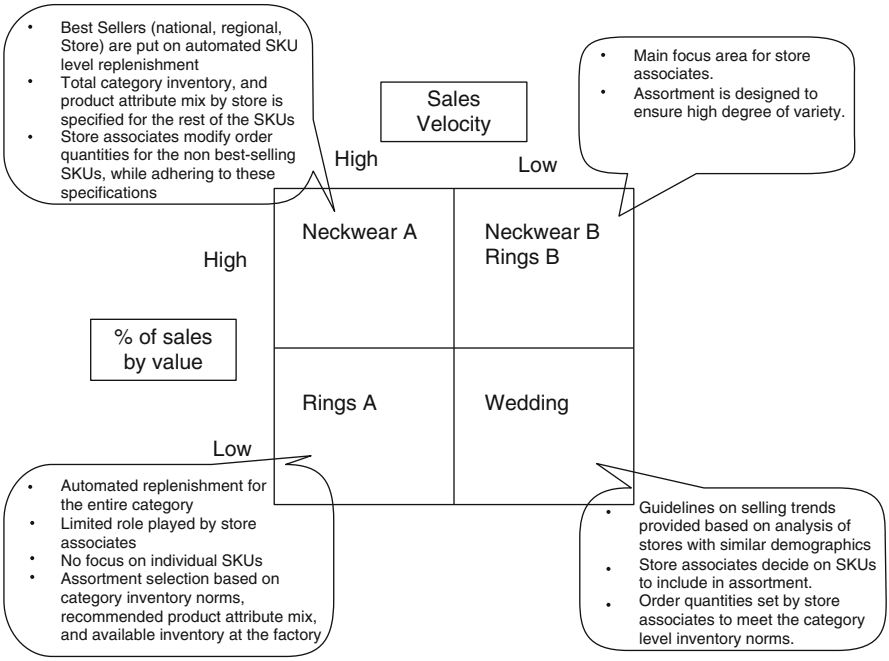


Fig. 8.5 Assortment strategy based on percent sales vs. sales velocity matrix

For example, in the case of a product category like Daily Neckwear, which has high percent sales contribution as well as high velocity, the high volume SKUs were put on replenishment, with inventory levels decided based on simple EOQ models. For the rest of the category, norms were drawn for overall inventory level and product attribute mix at each store (e.g. at Store A, overall inventory of Neckwear should be $ 2 million and the mix should be: Themes—50 % traditional, 30 % contemporary, and 20 % fashion; Gem—40 % large, 30 % medium, and 30 % small).

Based on the assortment strategy, the supply chain team developed a preliminary assortment plan for each store, with suggested products and inventory levels. With a bulk of the products put on SKU level replenishment, the work of store associates has been considerably reduced.

For the products not on SKU-level replenishment, the store associates were at liberty to modify the products selected and order quantities based on their knowledge of localized customer preferences. This was subject to the overarching inventory and product attribute mix guidelines drawn by the central team. This is done through a visual interface that provides the store associates a dynamic picture of how the modified order is stacking up against corporate guidelines.

Through the adoption of a hybrid model, Tanishq was thus able to customize its product offering to suit each store's clientele, while at the same time automating a bulk of the assortment planning process.

6.4 Albert Heijn

Albert Heijn, BV is a leading supermarket chain in the Netherlands with 1,187 stores and about $10 billion in sales.[2] In the grocery industry, supermarkets often carry more than 30,000 stock keeping units (SKUs). At the top level of the hierarchy, SKUs are divided into three groups: chilled products, dry goods, and groceries. Each group then is divided into merchandising categories, such as wines, bread spreads, butter and margarine. A subcategory is defined as a group of variants such that the difference between products within a subcategory is minimal, but the difference between subcategories is significant. For example, the subcategories in the butter and margarine category include deep-fry fat, regular butter, healthy butter, and margarines. We assume that substitution takes place within a subcategory but not across subcategories. The assortment planning models reviewed in this chapter focused on the selection and inventory/space allocation within a subcategory given a fixed shelf space and other constraints. Albert Heijn follows a hierarchical approach to assortment planning. First, store space is allocated to categories. Then product selection and facing allocation to products are

[2] Albert Heijn, BV is a subsidiary of Ahold Corporation, which owns many supermarket chains around the world with about 8,500 stores and $50 billion in sales.

carried out, subject to the shelf space constraint. In this subsection, we describe the details of this hierarchical approach.

Albert Heijn solves the following optimization problem to allocate shelf space between categories for each store.

$$\max\left\{\sum\nolimits_i P_i(x_i) : \sum\nolimits_i x_i \leq StoreShelfSpace;\ x_i \geq 0, \forall i.\right\}$$

$P_i(x_i)$ is the category gross profit when x_i meters of shelf space is allocated to category i. The function P_i is assumed to have a logarithmic form whose parameters are estimated using data from multiple stores $(x_i, P_i(x_i))$. The optimization is done by a Greedy Heuristic—allocating 1 m of shelf space at each step to the category with the highest incremental gross profit. Note that this shelf space allocation approach is similar to Corstjens and Doyle (1981), except that cross-space elasticities are not included in the formulation (i.e., category gross profit depends only on the category shelf space).

(Contrast this with the shelf space allocation approach at Borders Bookstores. Borders grouped 300,000 titles into 300 categories and allocated shelf space to categories on the premise that, "Except for best sellers, a customer is interested not in title but category". Category popularity is assessed by computing RST (Relative sales per title = Category sales/Number of titles). Shelf space is periodically reassigned from low RST to high RST. Following the principle of "Survival of the Fittest", categories "fight" for shelf space. Store managers are allowed to pick titles to be stocked within each category, thereby decentralizing a part of the decision process. Assuming that the number of titles is a proxy for category shelf space, RST is equivalent to $P_i(x_i)/x_i$. The Borders approach is similar to that of Albert Heijn except that rather than allocating the last meter of shelf space based on the marginal return, Borders allocates space based on average return from a category.

At Albert Heijn, it is the category manager's responsibility to choose the number of products and their shelf space allocation in each category, given a fixed shelf space. Category managers use several heuristics and their expertise about the category in order to make these decisions. Firstly, Albert Heijn wants to be known as the high variety, high quality supermarket in the Netherlands. One of the guidelines to achieve this strategical mandate is to carry 10 % more variety than the nearest competitor. The minimum number of SKUs in a subcategory, the minimum number of facings in a subcategory, the minimum and maximum number of facings for particular SKUs are also specified by category managers. If there is a need to reduce variety in a subcategory, the likely candidate is the subcategory with the highest substitution rate. To introduce new products periodically, m worse products are discarded and m new products are included in the assortment. Given the product selection, facings are allocated to products proportional to their demand rates.

Inventory management operates within the given facing allocations for a selection of products. For non-perishable items, the assigned facings are filled as much as possible at all times, even in the non-peak-load periods. That is achieved by

ordering an integral number of case packs such that the inventory position is as close as possible to and less than the maximum inventory level that would fit in the allocated facings. For perishable items that have a shelf life of a few days or less (e.g., produce), the inventory control is done in a more dynamic way. Albert Heijn uses a real-time system that estimates the demand for each product in the assortment based on the sales in the last few hours, and places an order to maximize each product's expected revenues minus cost of disposed inventory.

6.5 *Comparison of Academic and Industry Approaches to Assortment Planning*

This section compares and contrasts the approaches taken by academia and industry to assortment planning. Industry has taken a more strategic and holistic approach, while academics use a more operational and detail oriented approach. In some respects these approaches are nicely complementary in that the aspects of assortment planning that have received least attention in practice have received the most attention in academia, and academic research has the potential to fill a void in retail practice.

For most retailers, the process of assortment planning starts at the strategic level. The breadth of product categories carried and the depth of products offered in each of them is a function of the retailer's position in the competitive landscape. For example, a retailer like Best Buy would carry a rarely demanded product such as a 10 mega-pixel camera, just to maintain consumer perception of Best Buy as offering the latest technologies. In other words, the assortment would carry products which are otherwise unprofitable, but are a strategic necessity. While academic research does acknowledge such phenomenon (Cachon et al. 2005), there is little research that focuses on incorporating these strategic considerations while optimizing the assortment.

The other strategic aspect that retailers are concerned with is the role of a product category in their mix. Going back to the Best Buy example, it might be the case that Best Buy offers a very extensive assortment of HDTV's, more than what might be the optimal number when looked upon in isolation, for they are the main traffic drivers for the store. In other words, customers prefer to shop at Best Buy as they see extensive variety on offer in key categories, and as a result end up buying at Best Buy. There is little academic research (except Cachon and Kök 2007) that models this aspect of an assortment. On the other hand, the pricing version of this phenomenon (loss leaders and advertising features to drive traffic into the store and the razor-blade model) is extensively studied in the marketing literature.

One common theme across all the industry examples is that retailers recognize the fact that not all categories should be treated the same. The major drivers of sales in each category are different. While product variety may be the most important factor in a consumers store choice and purchasing decisions for one category,

promotions, in-store service experience, and impulse buying (aisle displays) may be more critical for another category. For example, Dhar et al. (2001) find that increasing the breadth and depth of the assortment does not have a positive effect on the performance of high penetration, high frequency categories like coffee and cereals.

Most retailers consider product selection as one among several levers (like promotions, pricing, etc.) that influence sales. Hence, they find it critical to integrate assortment planning decisions with the other influencing parameters. For example, if an apparel retailer is advertising a certain line of clothing heavily, then the variety that needs to be offered is higher than what might have been required without the attention due to advertising. Hence, retailers make assortment decisions in conjunction with other key factors that influence sales.

Retailers are well aware of the dynamic nature of the problem. At many retailers, the initial assortment developed by the buyers is tested across a sample of stores to get an early read, prior to the actual selling season. The test results are used to understand trends on winners and losers and gaps in the portfolio so as to redesign the assortment. As there are several other factors such as promotions, pricing, display, etc. which affect sales on an ongoing basis, the assortment is reviewed from time to time and appropriate changes are made. Academic papers, with the exception of Caro and Gallien (2007), consider static assortments. Even in mature categories, the frequent introduction of new products make it a necessity to revise the assortments. In practice, categories in different stages of their life cycles or categories with seasonal products require different assortment planning approaches. Growth potential is another strategic consideration that influences a retailer's assortment. For example, a dying product category like VCRs might not have the variety that a growing category like DVDs would.

The Tanishq example illustrates how assortment planning and replenishment can be attribute-focused rather than product-focused. For non-best sellers, Tanishq chooses a certain theme and gem size distribution as the defining properties of the target assortment. This approach is sensible, especially for categories in which attributes of the products are critical in driving traffic and influencing consumers' choice behavior. The attribute-focused approach is common in apparel retailing as well. Levy and Weitz (2004) describe the assortment plan for a jeans category where the size distribution, colors and styles are the main attributes that define the assortment. The total inventory budget is then allocated to products given the required distribution of the assortment over these attributes. Academic assortment planning models are mostly product-focused.

Customization of the assortment at the store level has gotten scant attention from retailers and no attention from academics. The Tanishq example illustrates a hybrid approach, where either the assortment or the guidelines for the assortment of the categories are planned at the corporate level, and for some categories store associates tinker with the assortment given the guidelines. Albert Heijn also follows the hybrid approach in that the store assortments are chosen from a chain-wide assortment. Borders Books is the best example we know of a retailer that aggressively customizes assortments at the store level.

Retailers take supply chain considerations into account in assortment planning. For example, Best Buy considers vendor relations, vendor performance and the number of products in other categories from a vendor while developing the assortment plans. However, there is very limited discussion of assortment planning from a supply chain view in the academic literature.

We performed a search on Google for "retail assortment planning" and found more than 700 references. Most of these references are to the product description of software providers and consulting firms, indicating a strong industry interest in the topic. Some academic papers come up in the search as well. One interesting observation that complements the discussion above is that there is a huge disconnect between the two groups: the language or the terminology of each group is substantially different and neither group acknowledges the existence of the other.

7 Directions for Future Research

There has been strong interest in assortment planning research since the first edition of this book chapter in 2008. Four research avenues emerge as important future research directions based on our discussion in this chapter.

First, more empirical work is needed in understanding the impact of assortment variables on consumers' store choice and purchasing behavior. Second, most of the existing theoretical models have not been implemented as part of industry applications (or their theoretical predictions have not been empirically tested). The field would benefit from such applications and empirical tests, as a validation of the assumptions in the increasingly complicated assortment planning models being formulated in the academic literature. Third, it seems that there are significant opportunities in generalizing the existing theoretical work to handle more complex problems faced by the retailers. One example would be to allow customization of the assortment by store. Fourth, incorporating the empirical findings on consumer behavior and perception of variety in assortment optimization models seems a worthy area of research. Below we describe some possible research topics from these four avenues in no particular order.

Demand arrival is assumed to be exogenous in most academic models. Understanding the drivers of store traffic through market share or store choice models, and incorporating those in assortment planning is a possible research direction. Lower prices, for example, would increase store traffic, but on the other hand, lower margins would lead to narrower assortments. Retailers recognize these interactions but make these decisions sequentially and in rudimentary ways. The joint pricing and assortment planning problem has not been studied in depth. Aydin and Ryan (2000) study optimal pricing under MNL model but do not consider operational costs. Cachon et al. (2008) are interested in the impact of competitive intensity on the variety level and prices.

Academic models take a static view of the assortment planning problem, whereas in practice, assortment decisions in a category can be made several times

throughout the season. The problems that industry faces include not only multi-period problems, but also managing the assortment for multiple generations of products, as in the digital versus traditional camera example. The dynamic assortment problem provides a rich set of research questions.

A significant number of papers have started studying dynamic assortment planning. Demand learning through tests in sample stores or online environments remain a topic worthy of investigation. Online retail environments and omni-channel retailing bring up many novel applications of dynamic assortment planning and open research questions.

Assortment planning models assume that there is a well defined set of candidate products, for which the consumer choice behavior is known perfectly. It may be interesting to take an attribute view of this problem, where consumers are interested in particular attributes rather than products. Mostly, a category is assumed to be composed of homogenous products that are potential substitutes from a consumer's perspective. Assortment planning for vertically differentiated products (i.e., varying quality) or more general choice models (e.g., subgroups of products that are more likely than others to be substitutes) can be studied to generalize the existing results on properties of optimal assortments. There is a significant body of literature in marketing on consumers' perception of variety as mentioned in Sect. 2.4. Incorporating some of those concepts in assortment planning may increase the applicability of the theoretical models.

Consumers are usually assumed to be a homogenous group. However, marketing literature places particular emphasis on understanding consumer segments. Estimation papers attempt to identify the latent consumer segments, and products are carefully positioned to achieve price discrimination between consumer segments in the product line design literature. Similarly in retail assortment planning, the consideration of multiple consumer segments may lead to optimal assortments that are composed of clusters of products that target these different segments. Recent work on mixed logit models and assortment customization provide a starting point in this direction.

Consumer purchase decisions across product categories may not always be independent. For example, a consumer's decision to buy a red colored sheet might depend on his being able to find a matching pillow. Explicitly incorporating this basket effect of consumer behavior while optimizing the assortment is an interesting research avenue. Agrawal and Smith (2003) and Cachon and Kök (2007) are first examples of this.

Estimating model parameters such as substitution probabilities, is another area that needs further research. There is an extensive body of literature in marketing (conjoint analysis) and econometrics that deal with parameter estimation for a wide variety of consumer choice models. However, there is little application of these in the assortment planning literature. For academic research to impact the industry, it is critical to invest research time in this area and to come up with innovative techniques to estimate the parameters which form the backbone of the several optimization models.

It is usually assumed that each individual buys a single unit of a single product in a category. This may not be true, even among substitutable products. For example, one shopper may buy multiple units of multiple flavors of yogurt in the same purchase occasion. This behavior violates the assumptions of standard choice models like the MNL, and it might be interesting to develop alternate models and study the properties of the resulting assortment. It would also be worthwhile to study the structure of the optimal assortment for product categories in situations when consumers are variety-seeking, causing the inventory-variety trade-off to take a different form.

Clearly, it is necessary to develop methods to understand the role of categories and to measure the intangible factors (such as the strategic importance of a category, the impact of assortment breadth or inventory levels on attractiveness of a store). The relation of assortment and inventory decisions with other levers such as pricing, promotions, and advertising has not been studied empirically. Joint optimization of some of these variables may lead to interesting results. It may be possible to draw from the literature on economics of product differentiation and the marketing/operations literature on product line design, both of which have extensively studied these variables and their impact on industry structures or product variety.

Assortment planning in multi-store, multi-tier supply chains is a completely open research area. Singh et al. (2005) and Aydin and Hausman (2003) are the only cases in the literature that incorporate supply chain considerations into assortment planning. The pros and cons of the hierarchical approach, the benefits of localization, and the execution problems associated with them have not been studied empirically or analytically. Balancing the benefits of customizing assortments by store with the increased cost of complexity is increasingly seen by retailers as a significant source of competitive advantage. An extremely interesting research question here is how to strike the balance, find the sweet spot between a "one size fits all" and "each store is its own" philosophies.

Incentive conflicts between the levels of the hierarchy may be a hurdle in deployment of the corporate assortment plans to the store level. Corporate level plans that are built based on strategic considerations may be imperfectly executed because the store managers' incentives are based on more short term objectives. The conflict of incentives between store managers, buyers, and vendors in a decentralized supply chain is yet another potential research area. For example, it is not clear how a category level assortment plan and the vendor-managed inventory agreements should be reconciled.

In conclusion, it seems to us that academics could make a tremendous contribution to retailing in the area of assortment planning. Retailers have developed practices that enable them to incorporate the complexities of the world in which they live, but they realize their approaches are too much based on art and judgment and that they could benefit from more rigorous use of the huge quantities of data available to them. If academics would be willing to work with individual retailers to understand their true complexity, they could make an enormous contribution in adding rigor and science to the retailer's planning process, much as academics have done in other areas like finance, marketing and strategy.

References

AC Nielsen. (1998). *Eighth annual survey of trade promotion practices*. Chicago, IL: ACNielsen.

Agrawal, N., & Smith, S. A. (1996). Estimating negative binomial demand for retail inventory management with unobservable lost sales. *Naval Research Logistics, 43*, 839–861.

Agrawal, N., & Smith, S. A. (2003). Optimal retail assortments for substitutable items purchased in sets. *Naval Research Logistics, 50*(7), 793–822.

Alptekinoğlu, A. (2004). Mass customization vs. mass production: Variety and price competition. *Manufacturing & Service Operations Management, 6*(1), 98–103.

Alptekinoğlu, A., & Grasas, A. (2014). When to carry eccentric products? Optimal retail assortment under consumer returns. *Production and Operations Management, 23.5*, 877–892.

Alptekinoğlu, A., Honhon, D., & Ulu, C. (2012). Positioning and pricing of horizontally differentiated products. Available at SSRN 2166570.

Alptekinoğlu, A., & Semple, J. (2013). *The exponomial choice model*. Working Paper, Pennsylvania State University.

Anderson, S.P., de Palma, A., & Thisse, J. F. (1992). *Discrete choice theory of product differentiation*. Cambridge, MA: The MIT Press.

Anupindi, R., Dada, M., & Gupta, S. (1998). Estimation of consumer demand with stockout based substitution: An application to vending machine products. *Marketing Science, 17*, 406–423.

Avsar, Z. M., & Baykal-Gursoy, M. (2002). Inventory control under substitutable demand: A stochastic game application. *Naval Research Logistics, 49*, 359–375

Aydin, G., & Hausman, W. H. (2003). *Supply chain coordination and assortment planning*. Working Paper, University of Michigan.

Aydin, G., & Ryan, J. K. (2000). Product line selection and pricing under the multinomial logit choice model. In *Proceedings of the 2000 MSOM Conference*.

Bassok, Y., Anupindi, R., & Akella, R. (1999). Single-period multiproduct inventory models with substitution. *Operations Research, 47*, 632–642.

Basuroy, S., & Nguyen, D. (1998). Multinomial logit market share models: Equilibrium characteristics and strategic implications. *Management Science, 44*(10), 1396–1408.

Baumol, W. J., & Ide, E. A. (1956). Variety in retailing. *Management Science, 3*, 93–101.

Bell, D. R., Ho, T.-H., & Tang, C. S. (1998). Determining where to shop: Fixed and variable costs of shopping. *Journal of Marketing Research, 35*, 352–369.

Bell, D. R., & Lattin, J. M. (1998). Shopping behavior and consumer preference for store price format: Why large basket shoppers prefer EDLP. *Marketing Science, 17*, 66–88.

Ben-Akiva, M., & Lerman, S. R. (1985). *Discrete choice analysis: Theory and application to travel demand*. Cambridge, MA: The MIT Press.

Bernstein, F., Gürhan Kök, A., & Xie, L. (2011). *Dynamic assortment customization with limited inventories*. Working Paper, Duke University.

Besbes, O., & Saure, D. (2011). *Product assortment and price competition with informed consumers*. Working Paper, Columbia University.

Boatwright, P., & Nunes, J. C. (2001). Reducing assortment: An attribute-based approach. *Journal of Marketing, 65*(3), 50–63.

Borin, N., & Farris, P. (1995). A sensitivity analysis of retailer shelf management models. *Journal of Retailing, 71*, 153–171.

Broniarczyk, S. M., Hoyer, W. D., & McAlister, L. (1998). Consumers' perception of the assortment offered in a grocery category: The impact of item reduction. *Journal of Marketing Research, 35*, 166–176.

Bucklin, R.E., & Gupta, S. (1992). Brand choice, purchase incidence, and segmentation: An integrated modeling approach. *Journal of Marketing Research, 29*, 201–215.

Bultez, A., & Naert, P. (1988). SHARP: Shelf allocation for retailers profit. *Marketing Science, 7*, 211–231.

Cachon, G. P., & Kök, A. G. (2007). Category management and coordination of categories in retail assortment planning in the presence of basket shoppers. *Management Science, 53*(6), 934–951.

Cachon, G. P., Terwiesch, C., & Xu, Y. (2005). Retail assortment planning in the presence of consumer search. *Manufacturing & Service Operations Management, 7*(4), 330–346.

Cachon, G. P., Terwiesch, C., & Xu, Y. (2008). On the effects of consumer search and firm entry in a multiproduct competitive market. *Marketing Science, 27.3*, 461–473

Campo, K., Gijsbrechts, E., & Nisol, P. (2004). Dynamics in consumer response to product unavailability: Do stock-out reactions signal response to permanent assortment reductions? *Journal of Business Research, 57*, 834–843.

Caro, F., & Gallien, J. (2007). Dynamic assortment with demand learning for seasonal consumer goods. *Management Science, 53.2*, 276–292.

Chen, F., Eliashberg, J., & Zipkin, P. (1998). Customer preferences, supply-chain costs, and product line design. In T.-H. Ho & C. S. Tang (Eds.), *Product variety management: Research advances*. Norwell: Kluwer Academic Publishers.

Chiang, J. (1991). A simultaneous approach to the whether, what and how much to buy questions. *Marketing Science, 10*, 297–315.

Chintagunta, P. K. (1993). Investigating purchase incidence, brand choice and purchase quantity decisions of households. *Marketing Science, 12*, 184–208.

Chong, J. K., Ho, T. H., & Tang, C. S. (2001). A modeling framework for category assortment planning. *Manufacturing & Service Operations Management*, **3**(3), 191–210.

Cooper, L. G., & Nakanishi, M. (1988). *Market-share analysis: Evaluating competitive marketing effectiveness*. Boston: Kluwer Academic Publishers.

Corstjens, M., & Doyle, P. (1981). A model for optimizing retail space allocations. *Management Science, 27*, 822–833.

Davis, J. M., Guillermo, G., & Topaloglu, H. (2014). Assortment optimization under variants of the nested logit model. *Operations Research, 62*(2), 250–273.

de Groote, X. (1994). Flexibility and marketing/manufacturing coordination. *International Journal of Production Economics, 36*, 153–167.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B, 39*, 1–38.

Desai, P., Radhakrishnan, S., & Srinivasan, K. (2001). Product differentiation and commonality in design: Balancing revenue and cost drivers. *Management Science, 47*, 37–51.

DeHoratius, N., & Raman, A. (2008). Inventory record inaccuracy: An empirical analysis. *Management Science, 54.4*, 627–641.

Dhar, S. K., Hoch, S. J., & Kumar, N. (2001). Effective category management depends on the role of the category. *Journal of Retailing, 77*(2), 165–184.

Dobson, G., & Kalish, S. (1993). Heuristics for pricing and positioning a product line. *Management Science, 39*, 160–175.

Downs, B., Metters, R., & Semple, J. (2002). Managing inventory with multiple products, lags in delivery, resource constraints, and lost sales: A mathematical programming approach. *Management Science, 47*, 464–479.

Dreze, X., Hoch, S. J., & Purk, M. E. (1994). Shelf management and space elasticity. *Journal of Retailing, 70*, 301–326.

Eliashberg, J., & Steinberg, R. (1993). Marketing-production joint decision-making. In J. Eliashberg & G. L. Lilien (Eds.), *Handbooks in OR & MS* (Vol. 5). Amsterdam: Elsevier

Emmelhainz, L., Emmelhainz, M., & Stock, J. (1991). Logistics implications of retail stockouts. *Journal of Business Logistics, 12*(2), 129–141.

Fader, P. S., & Hardie, B. G. S. (1996). Modeling consumer choice among SKUs. *Journal of Marketing Research, 33*, 442–452.

Farias, V. F., Jagabathula, S., & Shah, D. (2013). A nonparametric approach to modeling choice with limited data. *Management Science, 59.2*, 305–322.

Fisher, M. L., & Raman, A. (1996). Reducing the cost of demand uncertainty through accurate response to early sales. *Operations Research, 44*, 87–99.

Fisher, M., & Vaidyanathan, R. (2014). A demand estimation procedure for retail assortment optimization with results from implementations. *Management Science 60*(10), 2401–2415.

Gaur, V., & Honhon, D. (2006). Assortment planning and inventory decisions under a locational choice model. *Management Science, 52*(10), 1528–1543.
Greene, W. H. (1997). *Econometric analysis*. Englewood Cliffs, NJ: Prentice Hall.
Gruca, T. S., & Sudharshan, D. (1991). Equilibrium characteristics of multinomial logit market share models. *Journal of Marketing Research, 28*(11), 480–482.
Gruen, T. W., Corsten, D. S., & Bharadwaj, S. (2002). Retail out-of-stocks: A worldwide examination of extent, causes and consumer responses. Grocery Manufacturers of America.
Guadagni, P. M., & Little, J. D. C. (1983). A logit model of brand choice calibrated on scanner data. *Marketing Science, 2*, 203–238.
Hadley, G., & Whitin, T. M. (1963). *Analysis of inventory systems*. Englewood Cliffs, NJ: Prentice Hall.
Hoch, S. J., Bradlow, E. T., & Wansink, B. (1999). The variety of an assortment. *Marketing Science, 18*(4), 527–546.
Honhon, D., Gaur, V., & Seshadri, S. (2010). Assortment planning and inventory decisions under stockout-based substitution. *Operations Research, 58.5*, 1364–1379.
Honhon, D., Jonnalagedda, S., & Pan, X. A. (2012) Optimal algorithms for assortment selection under ranking-based consumer choice models. *Manufacturing & Service Operations Management, 14.2*, 279–289.
Hopp, W. J., & Xu, X. (2008). A static approximation for dynamic demand substitution with applications in a competitive market. *Operations Research, 56.3*, 630–645.
Hotelling, H. (1929). Stability in competition. *Economic Journal, 39*, 41–57
Huffman, C., & Kahn, B. E. (1998). Variety for sale: Mass customization or mass confusion? *Journal of Retailing, 74*, 491–513.
Irion, J., Al-Khayyal, F., & Lu, J. (2012). A piecewise linearization framework for retail shelf space management models. *European Journal of Operational Research, 222*(1), 122–136.
Jain, A., Rudi, N., & Wang, T. (2014). Demand estimation and ordering under censoring: Stock–out timing is (almost) all you need. *Operations Research, 63*(1), 134–150.
Kahn, B. E. (1995). Consumer variety-seeking in goods and services: An integrative review. *Journal of Retailing and Consumer Services, 2*, 139–48.
Kohli, R., & Sukumar, R. (1990). Heuristics for product line design. *Management Science, 36*(3), 1464–1478.
Kök, A. G. (2003). *Management of product variety in retail operations*. Ph.D. Dissertation, The Wharton School, University of Pennsylvania.
Kök, A. G., & Fisher, M. L. (2007). Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research, 55*(6), 1001–1021.
Kök, A. G., & Xu, Y. (2011). Optimal and competitive assortments with endogenous pricing under hierarchical consumer choice models. *Management Science, 57.9*, 1546–1563.
Kök, A., & Martínez-de-Albéniz, V. (2013). *A Competitive Model for Quick–Response Product Decisions*. Working Paper, Duke University.
Kurt Salmon Associates. (1993). *Efficient consumer response: Enhancing consumer value in the grocery industry*. Food Marketing Institute Report # 9–526, Food Marketing Institute.
Kurtulus, M. (2005). *Supply chain collaboration practices in consumer goods industry*. Ph.D. Dissertation, INSEAD.
Kurtulus, M., & Toktay, B. (2007). *Category captainship: Outsourcing retail category management*. Working Paper, Vanderbilt University.
Lancaster, K. (1966). A new approach to consumer theory. *Journal of Political Economy, 74*, 132–57.
Lancaster, K. (1975). Socially optimal product differentiation. *American Economic Review, 65*, 567–585.
Lancaster, K. (1990). The economics of product variety: A survey. *Marketing Science, 9*, 189–210.
Levy, M., & Weitz, B. A. (2004). *Retailing management* (pp. 398–400). New York: McGraw-Hill/ Irwin.

Li, Z. (2007). A single-period assortment optimization model. *Production and Operations Management, 16.3*, 369–380.

Lippman S. A., & McCardle, K. F. (1997). The competitive newsboy. *Operations Research, 45*, 54–65.

Maddah, B., & Bish, E. K. (2004). Joint pricing, assortment, and inventory decisions for a retailer's product line. 2007. *Naval Research Logistics, 54*(3), 315–330.

Mahajan, S., & van Ryzin, G. J. (1999). Retail inventories and consumer choice. Chapter 17. In S. Tayur, et al. (Eds.), *Quantitative methods in supply chain management*. Amsterdam: Kluwer.

Mahajan, S., & van Ryzin, G. (2001a). Stocking retail assortments under dynamic consumer substitution. *Operations Research, 49*(3), 334–351.

Mahajan, S., & van Ryzin, G. (2001b). Inventory competition under dynamic consumer choice. *Operations Research, 49*(5), 646–657.

Manchanda, P., Ansari, A., & Gupta, S. (1999). The "shopping basket": A model for multicategory purchase incidence decisions. *Marketing Science, 18*(2), 95–114.

Martínez-de-Albéniz, V., & Roels, G. (2011). Competing for shelf space. *Production and Operations Management 20*(1), 32–46.

McBride, R. D., & Zufryden, F. S. (1988). An integer programming approach to the optimal product line selection problem. *Marketing Science, 7*(2), 126–140.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics*. New York: Academic.

McGillivray, A. R., & Silver, E. A. (1978). Some concepts for inventory control under substitutable demand. *INFOR, 16*, 47–63.

Miller, C. M., Smith, S. A., McIntyre, S. H., & Achabal, D. D. (2010). Optimizing retail assortments for infrequently purchased products. *Journal of Retailing, 86*(2), 159–171

Miranda Bront, J., Mendez-Diaz, I., & Vulcano, G. (2009). A column generation algorithm for choice-based network revenue management. *Operations Research, 57*(3), 769–784.

Moorthy, S. (1984). Market segmentation, self-selection, and product line design. *Marketing Science, 3*, 288–307.

Musalem, A., et al. (2010). Structural estimation of the effect of out-of-stocks. *Management Science, 56.7*, 1180–1197.

Mussa, M., & Rosen, S. (1978). Monopoly and product quality. *Journal of Economic Theory, 18*, 301–317.

Netessine, S., & Rudi, N. (2003). Centralized and competitive inventory models with demand substitution. *Operations Research, 51*, 329–335.

Netessine, S., & Taylor, T. A. (2007). Product line design and production technology. *Marketing Science, 26*(1), 101–117.

Noonan, P. S. (1995). *When consumers choose: A multi-product, multi-location newsboy model with substitution*. Working Paper, Emory University.

Pan, X. A., & Honhon, D. (2012). Assortment planning for vertically differentiated products. *Production and Operations Management, 21.2*, 253–275.

Parlar, M. (1985). Optimal ordering policies for a perishable and substitutable product: A Markov decision model. *Infor, 23*, 182–195.

Parlar, M., & Goyal, S. K. (1984). Optimal ordering policies for two substitutable products with stochastic demand. *Opsearch, 21*(1), 1–15.

Progressive Grocer. (1968a, October). The out of stock study: Part I. S1–S16.

Progressive Grocer. (1968b, November). The out of stock study: Part II. S17–S32.

Quelch, J. A., & Kenny, D. (1994). Extend profits, not product lines. *Harvard Business Review, 72*, 153–160.

Rajaram, K. (2001). Assortment planning in fashion retailing: Methodology, application and analysis. *European Journal of Operational Research, 129*, 186–208.

Rajaram, K., & Tang, C. S. (2001). The impact of product substitution on retail merchandising. *European Journal of Operational Research, 135*, 582–601.

Raman, A., McClellan, A. d., & Fisher, M. L. (2001). Supply chain management at world Co. Ltd. Harvard Business School Case # 601072.

Rusmevichientong, P., & Topaloglu, H. (2012). Robust assortment optimization in revenue management under the multinomial logit choice model. *Operations Research, 60.4*, 865–882.

Rusmevichientong, P., Shen, Z.-J. M., & Shmoys, D. B. (2010). Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations Research, 58.6*, 1666–1680.

Russell, G. J., Bell, D. R., et al. (1997). Perspectives on multiple category choice. *Marketing Letters, 8*(3), 297–305.

Saure, D., & Zeevi, A. (2013). Optimal dynamic assortment planning with demand learning. *Manufacturing & Service Operations Management, 15*(3), 387–404.

Schary, P., & Christopher, M. (1979). The anatomy of a stockout. *Journal of Retailing, 55*(2), 59–70.

Simonson, I. (1999). The effect of product assortment on buyer preferences. *Journal of Retailing, 75*, 347–370.

Singh, P., Groenevelt, H., & Rudi, N. (2005). *Product variety and supply chain structures*. Working Paper, University of Rochester.

Smith, S. A., & Agrawal, N. (2000). Management of multi-item retail inventory systems with demand substitution. *Operations Research, 48*, 50–64.

Song, J.-S. (1998). On the order fill rate in multi-item, base-stock systems. *Operations Research, 46*, 831–845.

Song, J.-S., & Zipkin, P. (2003). Supply chain operations: Assemble-to-order systems. In S. Graves & T. De Kok (Eds.), *Handbooks in operations research and management science. Supply chain management* (Vol. XXX). North-Holland: Amsterdam.

Talluri, K., & van Ryzin, G. (2004). Revenue management under a general discrete choice model of consumer behavior. *Management Science, 50*, 15–33.

Ulu, C., Honhon, D., & Alptekinoğlu, A. (2012). Learning consumer tastes through dynamic assortments. *Operations Research, 60.4*, 833–849.

Urban, T. L. (1998). An inventory-theoretic approach to product assortment and shelf space allocation. *Journal of Retailing, 74*, 15–35.

Vaidyanathan, R., & Fisher, M. (2012). *Assortment planning*. Working Paper, The Wharton School, University of Pennsylvania.

van Herpen, E., & Pieters, R. (2002). The variety of an assortment: An extension to the attribute-based approach. *Marketing Science, 21*(3), 331–341.

van Ryzin, G., & Mahajan, S. (1999). On the relationship between inventory costs and variety benefits in retail assortments. *Management Science, 45*, 1496–1509.

van Ryzin, G., & Vulcano, G. (2013). *An expectation-maximization algorithm to estimate a general class of non-parametric choice-models*. Working Paper.

Vulcano, G., Van Ryzin, G., & Ratliff, R. (2012). Estimating primary demand for substitutable products from sales transaction data. *Operations Research, 60.2*, 313–334.

Walter, C., & Grabner, J. (1975). Stockout models: Empirical tests in a retail situation. *Journal of Marketing, 39*, 56–68.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics, 11*, 95–103.

Zinn, W., & Liu, P. (2001). Consumer response to retail stockouts. *Journal of Business Logistics, 22*(1), 49–71.

Chapter 9
Fast Fashion: Business Model Overview and Research Opportunities

Felipe Caro and Victor Martínez-de-Albéniz

1 Introduction

The global apparel industry has experienced a compound annual growth rate of 4.3 % since 2000, reaching a market size of USD 1.7 trillion in 2012 (Euromonitor International 2013). The growth has not only been in terms of revenue. The number of pieces of clothing purchased per capita increased from 9.0 in 2000 to 13.9 in 2012 worldwide, and in countries like the United Kingdom it has increased from 18.7 to 29.5 over the same period (Euromonitor International 2013). Part of the growth embedded in these figures has been attributed to the emergence of new industry players—collectively known as "fast-fashion retailers"—which have seen an explosive expansion since the turn of the century. In fact, stores like Hennes and Mauritz (H&M) from Sweden and Zara—the flagship brand of Inditex from Spain—have established themselves as recognized brands (Interbrand 2013) and have grown to become the largest apparel retailers in the world, see Fig. 9.1.

Fast fashion brought fresh air into the textile and apparel industries and it quickly struck a chord with the consumer. From a management and economics perspective,

F. Caro (✉)
UCLA Anderson School of Management, Los Angeles, CA 90095, USA
e-mail: fcaro@anderson.ucla.edu

V. Martínez-de-Albéniz
IESE Business School, University of Navarra, Barcelona, Spain
e-mail: valbeniz@iese.edu

N. Agrawal, S.A. Smith (eds.), *Retail Supply Chain Management*,
International Series in Operations Research & Management Science 223,
DOI 10.1007/978-1-4899-7562-1_9

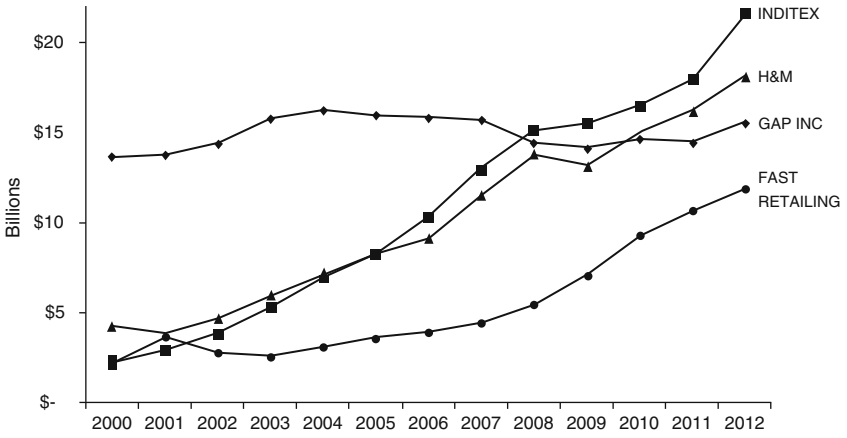


Fig. 9.1 Select specialty apparel retailer revenues in 2000–2012. *Source*: annual reports

fast fashion has been the long-awaited realization of "lean retailing" with items produced in small batches and within short lead times. Moreover, fast fashion's reliance on near-shore production has given a lifeline to an otherwise dying industry in developed countries (Abernathy et al. 2006; Doeringer and Crean 2006). On the other hand, fast fashion has been associated with a disposable culture and its social responsibility is constantly under scrutiny (Siegle 2011; Cline 2012).

Fueled by the success and growth of fast-fashion retailers, the term *fast fashion* has become ubiquitous and it has been used indiscriminately to describe almost any specialty apparel retailer below a certain price threshold, spanning stores like Old Navy and Chico's that have almost nothing in common besides the fact that they sell clothes. Hence, given the prominent role of fast fashion in the last decade, it is worth asking: which retailers are fast fashion and how do they operate? To find an answer to this question, in Sect. 1.1 we first follow a qualitative approach based on online sources and then in Sect. 1.2 we provide a more precise academic definition and we postulate metrics to measure "degrees" of fast fashion.

1.1 Which Firms Are Fast Fashion and How Do They Operate?

The Wikipedia entry for *fast fashion* lists 21 firms.[1] The list is quite diverse, but most of the firms have the following in common. First, they are specialty apparel retailers with brick and mortar stores and some online presence. Second, they are not "haute couture" or trend-setters but rather fashion followers that target the

[1] http://en.wikipedia.org/wiki/Fast_fashion, accessed January 17, 2014.

mid-to-low price range. To elaborate a more definite list of firms, we performed a frequency count using the Factiva database. We first searched for all the media publications in the last 2 years that contained the exact phrase "fast fashion" and we looked for brand names to form a preliminary list. Then, for each brand, we counted in how many of these media publications the brand was mentioned. A ranking of the brands that appeared in at least 20 publications is shown in Table 9.1 and a word-cloud representation is shown in Fig. 9.2. As a form of validation, we performed the same frequency count using all the PDF documents available through Google that contained the exact phrase "fast fashion". The corresponding ranking using the latter is also reported in Table 9.1.

The first remark from Table 9.1 is that the firms in the top ten are the same in both lists except for Wet Seal, which is a newcomer in the fast-fashion market so it appears more often in the Factiva search because the articles are more recent. Second, from Table 9.1 and Fig. 9.2 it is clear that H&M and Zara stand out with a number of appearances that outshines the rest. Therefore, it is safe to say that these two specialty retailers embody what fast fashion is or at least they are widely recognized as the exemplary representation of fast fashion. H&M is a rather secretive company that does not disclose its operations but the annual report

Table 9.1 Frequency count of specialty apparel retailers in media publications that mention fast fashion (data retrieved August 26, 2013)

	Number of appearances in Factiva search		Number of appearances in PDF online search	
Specialty apparel retailer	Rank	% appearances	% appearances	Rank
H&M	1	31.7 %	41.0 %	2
Zara/Inditex	2	29.2 %	45.9 %	1
Gap	3	11.9 %	18.2 %	3
Uniqlo/Fast Retailing	4	9.9 %	9.4 %	8
Topshop	5	9.3 %	13.7 %	4
Forever 21	6	7.5 %	11.2 %	6
Mango	7	4.3 %	12.4 %	5
Wet Seal	8	3.2 %	0.6 %	16
Benetton	9	3.1 %	10.1 %	7
New Look	10	2.8 %	6.2 %	9
Esprit	11	2.8 %	4.7 %	10
C&A	12	1.9 %	4.7 %	11
American apparel	13	1.2 %	2.6 %	13
Urban outfitters	14	0.9 %	2.8 %	12
Peacocks	15	0.5 %	1.1 %	15
Charlotte Russe	16	0.5 %	0.2 %	17
Armani Exchange	17	0.3 %	1.5 %	14

The search in the Factiva database was among 7,587 articles published in the last 2 years that mentioned fast fashion. The PDF search was among 466 PDF files available to download in Google.com that mentioned fast fashion

Fig. 9.2 Word-cloud representation of fast-fashion specialty retailers based on number of appearances in Factiva search (cf. Table 9.1). The figure was generated by wordle.net

describes H&M's business concept as "fashion and quality at the best price" (H&M 2012). On the other hand, Zara has been repeatedly studied and its mode of operation has been widely documented, see Ferdows et al. (2002), Ghemawat and Nueno (2003), McAfee et al. (2004) and Lewis et al. (2004) or Caro (2012).

Zara—and H&M to a similar extent—have undertaken a radical change to the design cycle in order to provide fashion almost on demand. Specifically, these retailers have chosen to work at the item level—which includes all the sizes and colors of a given garment—rather than using collections. They can do this because they do not have a wholesale channel that is demanding a full collection, and they control the retail point of sales. Such control structure allows them to avoid batching thousands of products together. In particular, it is no longer necessary to design together products with quick and slow supplier lead times. In the words of H&M: "The time from an order being placed until the items are in the store may be anything from a few weeks up to 6 months. The best lead time will vary. For high-volume fashion basics and children's wear it is advantageous to place orders further in advance. In contrast, trendier garments in smaller volumes have to be in the stores much quicker" (H&M 2007).

Overall, the lead time of each product in the assortment depends on where it fits in the *fashion triangle* (see Fig. 9.3). At the bottom of the triangle are basic products. These items are the perennial products that are present at the store year after year with slight variations in design, such as a grey pullover or a white t-shirt. Basics are typically sourced in large quantities from low-wage countries and have long lead times. The center of the triangle is composed of fashion-basics or updated classics, which represent "basics with a feel for fashion" (H&M 2010). Fashion-basics have some fashion component—e.g., a non-traditional cut or a special trim—but they are produced as basics in varying volume. The line between basics and fashion-basics can be blurry. Moreover, since they share the same lead times, they tend be lumped

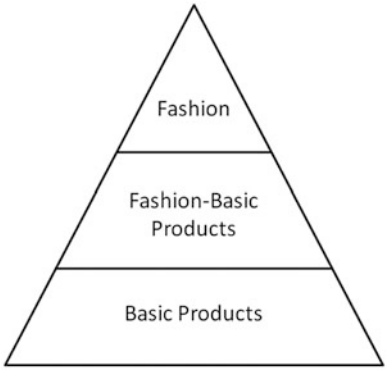


Fig. 9.3 The fashion triangle. Based on Abernathy et al. (1999)

in one category (which for ease of exposition we refer to as basics). At H&M, basic items roughly represent 70 % or more of the product assortment. At Zara, basics have increased from less than 20 % in the late 1990s to 40 % or more nowadays.

The top section of the fashion triangle corresponds to the (true) fashion products. For these items, H&M and Zara have typically used *quick-response* production to reach stores as soon as possible, thereby allowing them to respond to nascent demand trends first, so as to provide and capture more value from the consumers. This requires them to accelerate the production phase—using near-shore suppliers close to market in countries such as Portugal, Morocco, Bulgaria, Romania or even Turkey—and also the design phase, by directing the creative aspects towards a commercial need to reduce design iterations, and by using standard methods and materials to reduce efforts on samples. As a result, the total design-to-market time for an item to be launched in January can be reduced to a mere 6 weeks if the appropriate fabric is used and the go decisions (authorizations to move from sample to industrialization) are not delayed. In a way, they are like a surfer that is able to catch a wave before any other notices it. Figure 9.4 compares the planning process of fashion versus basic products (this figure also serves as a comparison with respect to a more traditional collection-based retailer that only carries basic items). The coexistence of fashions and basics calls for a dual supply chain. Moreover, the two types of products play different marketing roles. The fashion products generate customer traffic, sometime even playing the role of a loss leader, whereas the basics bring in the revenue.

An important advantage of working at the item level is that it gives the freedom to introduce products in the store continuously, not only twice a year. This implies that the utilization of all resources—designers, factories, distribution—can be balanced better over time, avoiding unnecessary peaks twice a year (see Fig. 9.5). Costs and response times can thus be reduced. The frequent assortment changes are also necessary for fashion items to keep up with the trends. Indeed, a retailer like H&M "buys items on an ongoing basis throughout the season to optimise fashion precision" (H&M 2011). Therefore, fast-fashion retailers combine supply chain agility to respond quickly, and constant product introductions to attract variety-seeking/

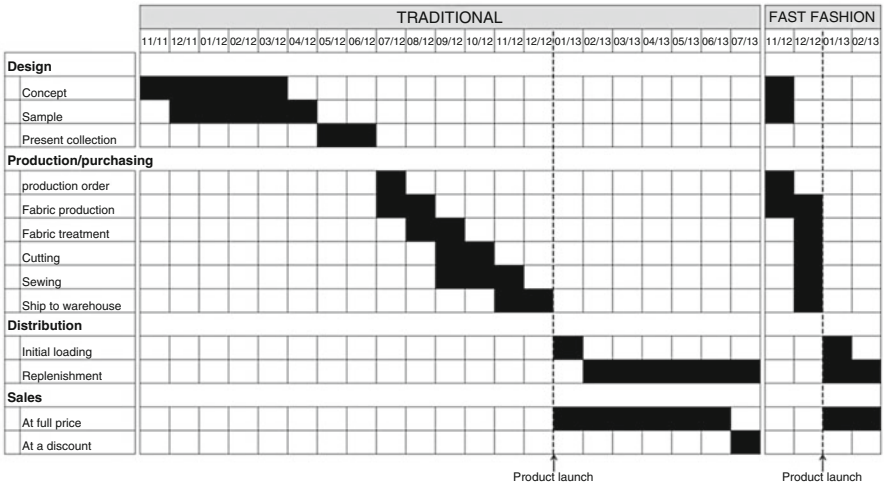


Fig. 9.4 Traditional vs. fast-fashion design-to-sales processes for a product introduced in January 2013. *Source*: Caro and Martínez-de-Albéniz (2013)

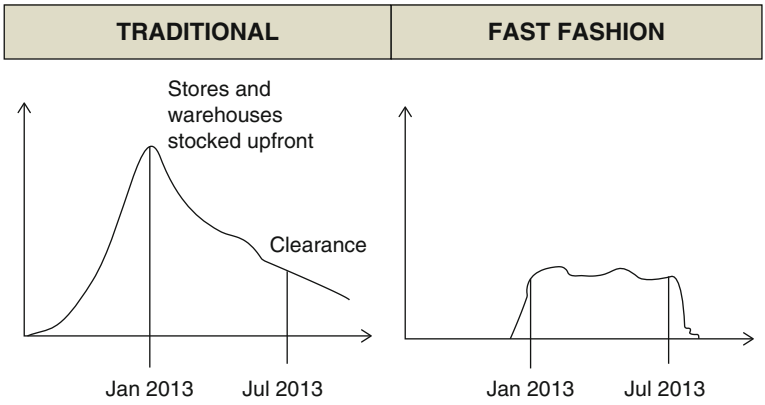


Fig. 9.5 Resource (designers, factories, distribution) utilization in a typical season

fashion-conscious customers. It is these two key features—operational agility and time-based variety—that we use next to measure the execution of the fast-fashion business model.

1.2 Defining and Measuring Fast Fashion

Based on the discussion above, fast fashion can be defined as a business model that combines three elements: (a) quick response; (b) frequent assortment changes; and

(c) fashionable designs at affordable prices. Note that the first two elements are fundamentally operational and allow the execution of fast fashion, whereas the last element represents the value proposition that the operational backend strives to deliver. Though this definition is quite broad, it does put a boundary and it leaves out several (fashion) retailers that sometimes are mistaken as being fast fashion. For instance, the fashion powerhouse Prada sells at a much higher price point—and the responsiveness of its supply chain is unclear—so it would not be fast fashion according to our definition. On the other end, there are many retailers that sell at affordable prices but they do not qualify as fast fashion either. For instance, Old Navy has very competitive prices but lacks quick response capabilities; or in the case of Chico's, the assortment is refreshed regularly but the products are mostly basics and fashion-basics (Chico's 2012).

The first two elements in our definition—namely, quick response and frequent assortment changes—characterize a fast-fashion supply chain, and for that reason we devote more attention to them in this book chapter and we postulate metrics to measure their effectiveness. Since the purpose of quick response is to reduce markdowns and stockouts, its effective implementation should lead to a better gross margin and less inventory. Therefore, an appropriate metric to measure the effectiveness of quick response is the gross margin return on inventory (GMROI), which is defined as the ratio between the gross margin and the average, where both quantities are measured at the aggregate firm level. The GMROI metric is largely used among retailers but several other ratios could serve the same purpose. For instance, Hausman and Thorbeck (2010) use Operating Income/Inventory as a markdown/stockout performance metric.

Measuring the dynamic assortment capability is less straightforward. Ideally, one would want to monitor and keep track of the product assortment on display at the stores, but collecting this data is impractical. Instead, we resort to the online stores in the USA. Specifically, for each specialty apparel retailer we considered the "new arrivals" of the Women's section and counted how many items were less than a week old. In other words, we counted the number of products that had become available less than a week ago. We disregarded variations in color and prints to only count those products that were really new introductions. Then, we took the average over a 20-week period.[2]

In Fig. 9.6 we plot the GMROI versus the weekly number of new arrivals for the top four specialty retailers in Table 9.1, which are publicly traded companies (the three retailers that follow on the list are privately held). It is noteworthy that Fig. 9.6 confirms that H&M and Zara are "in a different ball game" compared to Gap and Uniqlo. Not only do H&M and Zara have better dynamic assortment capabilities—in the order of 120 new product introductions per week on average—but they also get more margin out of their inventory, roughly 50 % better

[2] Zara has a separate section for Women in their teens (TRF), which we included in the count. The other retailers in the study have a single section for Women that includes teenagers.

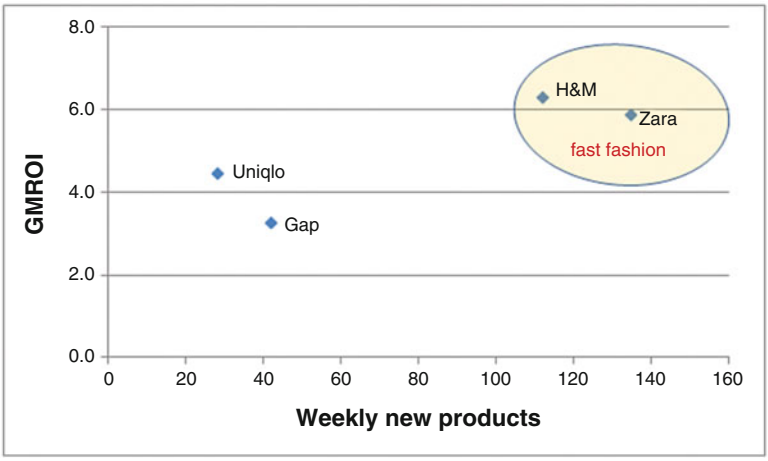


Fig. 9.6 GMROI versus the average number of weekly new products introduced by mid-to-low price specialty apparel brands. GMROI is a 5-year average. For Zara and Uniqlo we report the GMROI of the holding company (Inditex and Fast Retailing, respectively)

GMROI, which speaks to their ability to respond quickly with the right product/quantity so markdowns are less of an issue.[3] It is also interesting to observe from Fig. 9.6 that, though there is not a straight correlation between new arrivals and GMROI, there does seem to be a few local "sweet spots". In fact, H&M and Uniqlo introduce less products than their nearest competitor (Zara and Gap, respectively) and manage to achieve a higher GMROI. Finally, Fig. 9.7 shows the new arrivals over the 20-week period considered. Both Zara and H&M have big spikes when a new season is launched, but during the season Zara's assortment rotation tends to be more stable with a standard deviation of 37 new products versus 53 for H&M.

The reminder of this book chapter is structured as follows. Sections 2 and 3 explore in depth the literature on quick response and dynamic assortment, respectively. In Sect. 4 we survey papers related to the design and pricing strategies of fast-fashion retailers. We conclude the chapter in Sect. 5 by discussing ongoing challenges for fast-fashion retailers and we identify future research opportunities.[4]

[3] Topshop and Forever 21 introduce three times more products than H&M and Zara but it is unclear whether that pays off because their GMROI is unavailable.

[4] We focus on analytical and empirical research. For more qualitative work on fast fashion, we refer the reader to Choi (2013a).

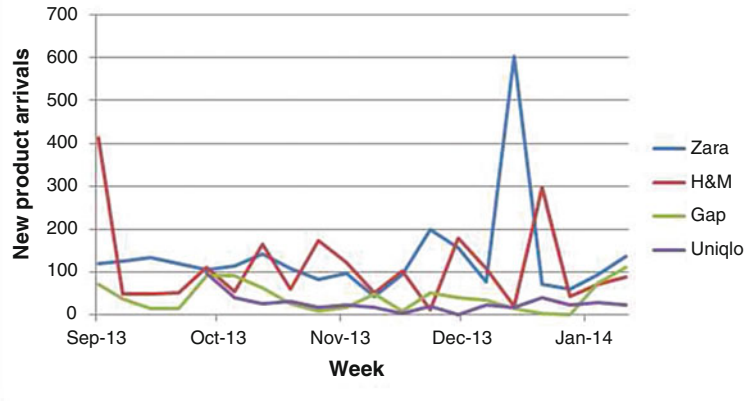


Fig. 9.7 Weekly new arrivals in the women section in Fall 2013

2 Sourcing and Quick Response

Quick Response (QR) was developed in the textile and apparel industry and since then it has been a prominent topic in Operations Management. QR was originally a set of standards for information exchange and supply chain management that allowed lead times to be shortened and increased supply chain efficiency (Palmer and Markus 2000). Over time, the use of the term QR has evolved into a broader interpretation, which is conceptually very simple: postpone all risky production decisions, e.g., commit to purchases that may not be needed in case of low sales, until there is enough evidence that the market demand is there. QR thus allows to reduce finished goods excess inventory, although per-unit costs (manufacturing and shipment) may increase. The concept is related to postponement and delayed differentiation (Feitzinger and Lee 1997; Lee and Tang 1997), as QR often requires holding raw materials ready to be died, cut and sewed after item-level demand forecasts have improved.

The early literature on QR, such as Iyer and Bergen (1997) or the classic Sport Obermeyer paper by Fisher and Raman (1996), centered on a single firm and brought to light the value of early information. Further academic contributions around QR for a single firm have focused on two main issues: advanced models for demand uncertainty and in particular how forecasts are improved over time; and integrating production constraints into the decision models. In addition, competition and externalities on the supply chain have been studied as well. Finally, empirical research is a promising new field of work for QR.

2.1 Demand Forecasting

Information is a key driver of QR decisions. It is widely accepted that it is impossible to forecast fashion at the item level a priori (Christopher et al. 2004). The only feasible approach is to start selling the product and use early sales data to generate more reliable forecasts. Dynamic demand models are thus required. Iyer and Bergen (1997) consider a model where demand is normally distributed with mean θ and standard deviation σ, where θ itself is unknown and follows a normal distribution with mean μ and standard deviation τ. Early sales will provide more accurate information on θ, which will help improve the demand forecast. Hence, if no information about θ is available, then demand is normally distributed with mean μ and standard deviation $\sqrt{\sigma^2+\tau^2}$. But if early sales d_1 are available, the demand forecast becomes normally distributed with mean $\mu(d_1) = \frac{\sigma^2}{\sigma^2+\tau^2}\mu + \frac{\tau^2}{\sigma^2+\tau^2}d_1$ and standard deviation $\sqrt{\sigma^2+\frac{1}{\rho}}$ where $\rho = 1/\tau^2 + 1/\sigma^2$, i.e., smaller than $\sqrt{\sigma^2+\tau^2}$. Hence, the higher $\tau^2\rho$ (i.e., the higher τ/σ), the better the forecast improvement due to observation of early sales. Fisher and Raman (1996) suggest a similar model where demand arrives in two time-windows: early and late sales follow a bivariate normal distribution and, after observing early sales, the distribution of late sales is updated. This updating scheme generally falls under the Martingale Model of Forecast Evolution (MMFE), see Heath and Jackson (1994). Other models have been used too. In particular, Lago et al. (2013) use a demand model where demand is exponentially decreasing over time, with an uncertain rate which is only revealed after the product is introduced. Demand is decreasing because inventory levels are reduced over time, thus decreasing the display, availability and consequently sale of the items. Higher rates imply that products sell out faster.

2.2 Production

The other main ingredient of QR is the consideration of production factors. Fisher (1997) provides a high level picture of the different types of supply chains, from efficient (long lead-times and rigid production schedules) to responsive (short lead-times and flexibility). If production costs are linear and there are no volume constraints, the problem is a relatively simple extension of the newsvendor model, see e.g., Martínez-de-Albéniz (2011) or Song and Zipkin (2012). The main trade-off there is to balance the higher costs of QR orders with the higher exposure to excess inventory costs of early orders. Specifically, letting q_1 be the early order quantity and q_2 the QR order quantity, and assuming that QR orders can be placed after demand D is realized, we can formulate the problem as follows:

$$\max \mathbb{E}_D \left[p \min\{D, q_1 + q_2(D)\} - c_1 q_1 - c_2 q_2(D) \right].$$

where p is the revenue per unit, and $c_1 \leq c_2$ the per-unit production cost of early and QR orders respectively, both less than p. It is optimal to set $q_2(D) = (D - q_1)^+$ and q_1 satisfying the critical fractile equation $Pr[D \geq q_1] = c_1/c_2$. Thus, if costs are relatively similar, QR orders will dominate, while if costs are very different, QR orders will be seldom used. Beyond this simplistic model, Fisher and Raman (1996) incorporate relevant apparel production constraints: minimum order quantities and capacity constraints. These are strong drivers of QR orders: QR capacity constraints imply that inflating early orders is desirable; minimum order quantities introduce binary decisions into the problem, which may reduce or increase early and QR orders, when the unconstrained order quantity is below the minimum. They describe an application to the Sport Obermeyer case study. Fisher et al. (2001) consider the possible cost of back-ordering between issuing and receiving the QR order, which makes the optimization problem intractable (expected profit is neither convex nor concave), so they suggest a heuristic and describe an application to a catalog retailer. A practical implementation of advanced optimization is suggested in Agrawal et al. (2002), who develop a methodology for managing a portfolio of retail products with different lead time requirements by using vendors that differ in costs and production flexibility.

2.3 Competitive Implications

Given the prevalence of QR, an essential step in the analysis is to consider how the practice changes firm behavior under competition. Indeed, QR was conceived as a competitive strategy expected to change "the rules of the game", in the words of Hammond and Kelly (1990), similar to what just-in-time manufacturing had meant to the auto industry.

A key paper in this line of work is Caro and Martínez-de-Albéniz (2010). They present a two-period model where firms make inventory decisions taking into account that demand will spill-over to the competitor whenever there is a stock-out. The two-period setting allows for demand updates, which is a fundamental feature of QR. Moreover, motivated by the emergence of fast-fashion retailers and their co-existence with more traditional apparel retailers, Caro and Martínez-de-Albéniz study in particular the asymmetric game where only one firm has the QR capability while the other firm uses "slow response" (SR) and cannot leverage early demand information. The main contribution of the paper resides in the insights for the asymmetric duopoly. It is shown that in equilibrium the QR firm will stock less while the SR firm will stock more compared to the case when both firms are SR (see Fig. 9.4 in the paper). The dynamics of this result are quite interesting. If the QR competitor committed to a high inventory level, the SR firm would actually want to stock less (see Proposition 3), but since such kind of commitment is not credible,

there is an opportunity for demand spill-overs that the SR firm seizes by stocking more. These spill-overs turn out to work well for the QR competitor since it depletes inventory that would otherwise be carried over to the next period. So, by stocking less the QR competitor lets the SR firm take most of the inventory risk upfront, and even in those scenarios where demand in the initial period is high, the QR firm benefits because then it faces less competition in the last period. This effect becomes even more pronounced with demand correlation because the QR firm can also learn at the competitor's expense. Though both firms move their inventory in opposite directions, it is shown that in equilibrium the aggregate industry inventory level decreases.

Another important implication from the paper is that with equal costs, QR is a dominant strategy. In other words, QR is a no-brainer regardless of the competitor's actions. This adds another layer to the significance of QR and gives a stronger message to firms that are yet to adopt it. Of course, a QR firm would be better off competing against a SR firm rather than another QR firm, which confirms that QR provides a competitive advantage. What is not so obvious is that a SR firm would prefer a QR over a SR competitor. This is due to the spill-overs in the first period that can favor the SR firm, so the asymmetric scenario can be beneficial to both competitors.

It is also possible that QR might involve higher costs (e.g., due to expediting or local production). In that case, Caro and Martínez-de-Albéniz show that QR pays more for "fashion" goods while SR is better for "basic" items with low demand variability or low correlation across periods. This is an analytical confirmation of the fundamental rule that the supply chain should match the type of product (Fisher 1997). Interestingly, the paper shows that with unequal cost structures the asymmetric competitive scenario can still be preferred by both competitors, and this continues to hold true even when the firms endogenously choose their supply chains. This provides support for the co-existence of QR and SR retailers observed in practice. Nasser and Turcic (2013) analyze a similar context and also observe an asymmetric equilibrium when the competing firms offer products with an intermediate level of differentiation.

Another related paper that studies QR under competition is Lin and Parlaktürk (2012). They propose a two-period production model where two retailers compete in a Cournot setting. Namely, the market clearing price is $A - \sum_{i=1}^{2} X_i$ where A is an uncertain parameter, and X_i is the quantity brought to market by retailer i. They analyze different scenarios where none, one or both retailers have access to QR from the manufacturer, and study the manufacturer's optimal pricing strategy. They find that for the manufacturer it may be best to offer QR to just one or to both retailers. In addition, in contrast with Caro and Martínez-de-Albéniz (2010), they show that QR can hurt a retailer when demand uncertainty on the market potential (parameter A) is low. This effect is due to the fact that a retailer without QR can credibly inflate its initial order, thereby forcing the fast retailer to reduce its order, and hence its profits.

2.4 Impact on Consumers and Suppliers

It is worth pointing out that there are several papers studying the externalities of QR on other stakeholders within the supply chain. Cachon and Swinney (2009) study the effect of QR on strategic consumers, those that may delay their purchases until the discount season, where price is lower. They show that, by reducing the amount of early orders, QR decreases the probability of having excess inventory at the end of the season, thereby reducing the incentive of strategic consumers to wait for discounts. As a result, QR becomes even more valuable when consumers are strategic, as opposed to myopic. The opposite effect is shown in Iyer and Bergen (1997) when there is an intermediary (e.g., a retail partner such as department store) between the manufacturer using QR and customers. Indeed, the manufacturer adopting QR may lose sales from the retailer, its "sell-in" (as opposed to the "sell-out" from retailer to final consumers). This is because, without QR, the retailer may be ordering a very high sell-in and taking most of the inventory risk, while with QR, it may reduce the expected sell-in to shift all the demand risk to the manufacturer. The way to make the transition to QR profitable for both retailer and manufacturer is then to put in place quantity discount or volume commitment schemes. Krishnan et al. (2010) incorporate retailer effort considerations: the retailer usually puts an effort that can influence the pace of sales. With such model in mind, the inventory reduction associated with QR will reduce the risk of excess inventory costs, thereby requiring less effort from the retailer's part, which may switch it to competing products. As in Iyer and Bergen (1997), the final outcome is that QR may be detrimental to the manufacturer, unless new contracts (beyond flat wholesale pricing) are put place. Finally, the impact on supplier pricing has also been studied in Calvo and Martínez-de-Albéniz (2012). They present a model where a retailer makes use of dual sourcing (advance orders with a slow, efficient supplier; and QR orders with a fast, more expensive supplier). The price quotes from the suppliers are endogenous to the retailer decisions regarding procurement. Specifically, if the retailer commits to single sourcing, then prices may in equilibrium be lower than if the retailer accepts to place both early and QR orders, which results in the retailer sometimes being worse off. This implies that using QR also removes pressures for both slow and fast suppliers to keep prices low, which may deteriorate overall retailer and supply chain performance.

2.5 Empirical Work

Finally, there is scarce empirical literature on QR. So far, the only exception is Lago et al. (2013) who evaluate the value of QR sourcing. They study the sales of products of a fast fashion firm over the Fall–Winter 2008 season. Each item, defined by a model and a color, may be introduced at a different time, and may be sourced from a different origin (from East Asia, South Asia, East Europe,

West Europe or North Africa). Such input variability allows Lago et al. to study how product performance, measured by the speed of sales, depends on different factors. They focus on the interaction between time of design and sourcing origin. Their results confirm most of the intuitions about QR: an item with a shorter time-to-market (Europe or Africa for the company under study) sells faster; and the speed-of-sales difference between QR and slow production is higher early in the season, thereby confirming that firms can learn as the season advances. Furthermore, the paper provides quantitative estimates of the advantage of QR. Namely, a product sourced under QR sells about twice as fast compared to one sourced with long lead-times. This provides a strong business case for QR if the sourcing cost difference is small compared to the value of inventory and space at the store.

3 Dynamic Assortment

Besides QR, the other main difference between fast fashion and traditional retailing is the way assortments are managed. Indeed, for many years the industry has worked around the concept of collections. Assortments are updated twice a year: at the beginning of the calendar year, the Spring–Summer collection is introduced; at the end of the summer the Fall–Winter collection is released. This industry-wide pace of change has been supported by design (cool hunting), communication (catwalks and store mock-ups where media and wholesale customers are invited), sales and marketing (catalogs, advertising) that follow similar bi-annual patterns. As a result, assortment planning with this approach can be considered as static. The chapter by Kök et al. (2015) in this handbook discusses extensively the academic literature around that problem.

In contrast, fast-fashion players rely much less on collection advertising and wholesale channels. As a result, they are able to design, produce and distribute new products dynamically, both at the beginning and the middle of the season. This raises interesting research problems that have only been explored recently.

One line of work extends the static assortment problem to multiple periods and incorporates demand learning. The set of products that can be included into each period's assortment is typically fixed, and the focus is on balancing exploration, i.e., including a product in order to learn about its demand rate, and exploitation, i.e., including a product with high demand rate and thus high profit. Caro and Gallien (2007) is the first paper to develop such a model, using a multi-armed bandit formulation. They decouple the dynamic assortment problem into a set of single-product dynamic programs and propose an index policy such that, in each period, only the products with the highest index should be included. The index includes both information about the expected demand rate and the potential value of better information on demand. Rusmevichientong et al. (2010) include a capacity constraint and design an algorithm for the dynamic problem, where parameters are estimated in parallel with revenue generation. Sauré and Zeevi (2013) focus on the asymptotic performance of such algorithms. Farias and Madan (2011) introduce

an irrevocability constraint, i.e., a product cannot be introduced again after it is removed; they design a heuristic that performs well. Alptekinoglu et al. (2011) use a locational model with unknown demand distributions that can be discovered by varying the assortment over time. All these papers assume that the demand parameters are stationary and need to be learnt.

Three important features are missing in the papers above: new products may be introduced also in the middle of the season, not all at the beginning; they cannot be introduced, removed and introduced again (Farias and Madan 2011); and demand is not stationary but typically decreases over time because, at the store, new products typically get better displays and generate more interest than older ones, everything else being equal.

Some recent papers have recognized that demand may change over time. Caldentey and Caro (2010) assume they follow a stochastic process over time, which they call the "vogue". Caro and Martínez-de-Albéniz (2012) use a satiation model where consumers progressively move away from stores that do not refresh their assortments often enough. But Caro et al. (2012) is the first paper to consider the three elements from above together in assortment planning. They take the entire set of products I as given and decide when each should be introduced over the season. The products compete for customer attention, and to capture such effect a demand attraction model is proposed: in period t, if product $i \in I$ is included in the assortment, its demand will be equal to $v_{it} / \left(v_0 + \sum_{j \in S_t} v_{jt} \right)$, where S_t is the set of product present in the assortment in period t, and v_{jt} is the attractiveness of the product in the period. Moreover, to incorporate decreasing demands over time, once introduced a product's attractiveness varies dynamically: $v_{jt} = \kappa_{j,t-intro_j} v_j$, where v_j is the attractiveness of the product when it is first introduced and $\kappa_{j,l}$ is the decay parameter that depends on the age l of the product. The focus of the paper is put on exponential attractiveness decays, i.e., $\kappa_{j,l} = \kappa_j^l$ with κ_j the decay parameter. Note that this demand model is supported by real sales data, as shown in their paper. It has also been used in describing the box office sales of movies (Ainslie et al. 2005). The parameters $v_j, \kappa_{j,l}$ are product characteristics, inputs into the model, as well as r_j the per-unit margin of product. Letting α_t denote the market size of period t, the optimization problem of Caro et al. (2012) can thus be written as an integer program:

$$\max \quad \sum_{t=1}^{T} \alpha_t \sum_{i=1}^{n} r_i \times \left(\frac{v_i \sum_{u=1}^{t} \kappa_{i,t-u} x_{iu}}{v_0 + \sum_{j=1}^{n} v_j \sum_{u=1}^{t} \kappa_{j,t-u} x_{ju}} \right)$$

$$\text{s.t.} \quad \sum_{t=1}^{T} x_{it} \leq 1 \qquad \forall\, i \in I,$$

$$x_{it} \in \{0,1\} \qquad \forall\, i \in I, t = 1, \ldots, T.$$

Caro et al. show that the optimization problem is in general NP-complete. They propose a fluid approximation that can be solved easily and can also be used for developing heuristics. In particular, the fluid approximation is a concave nonlinear maximization problem when product margins are identical; otherwise, the problem may not be concave, but their numerical study suggests that the optimal solution can be found quickly. Some appealing insights are derived: when decays are exponential and margins identical across products, the approximation's optimal solution is to introduce the products with less decay (i.e., higher κ_j) first. This implies that basic products, with stable demand, should be introduced in the beginning of the season. In contrast, fashionable products for which customer interest quickly drops should be spaced over the entire season and used to refresh the assortment. Moreover, Caro et al. show that the heuristics based on the fluid approximation generally perform very well, even when margins are not identical.

The framework presented in Caro et al. (2012) can be extended to capture most of the realities of fast fashion. In particular, rather than taken the set I of possible products as a given, it is important to let the retailer decide whether a new product should be designed and introduced in the middle of the season, depending on the most recent information. In other words, the model should incorporate closed-loop controls into the assortment decision. Çınar and Martínez-de-Albéniz (2013) propose a dynamic programming formulation to allow for such closed-loop decisions. Instead of binary introduction decisions, they allow for continuous amounts of products u_{it} to be introduced in category $i \in I$ in period t. These depend on the current attractiveness present in category i in period t, denoted x_{it}. As a result, the profit-to-go of the retailer in period t, J_t, can be written as $J_{T+1} \equiv 0$ (terminal condition) and

$$
\begin{aligned}
J_t\left((x_{it})_{i\in I}\right) = \max_{u_{1t},\ldots,u_{nt}\geq 0} \; & \frac{\sum_{i\in I} r_i y_{it}}{v_0 + \sum_{i\in I} y_{it}} - \sum_{i\in I} c_{it} u_{it} + \beta \mathbb{E}\left[J_{t+1}\left((x_{it+1})_{i\in I}\right)\right] \\
\text{s.t.} \; & y_{it} = x_{it} + u_{it} && \forall i \in I \\
& x_{it+1} = \widetilde{\varepsilon}_{it} y_{it} && \forall i \in I
\end{aligned}
$$

The decay of attractiveness is similar to Caro et al. (2012), since attractiveness randomly decays with parameter $\widetilde{\varepsilon}_{it}$; this extends the deterministic decay κ_j of Caro et al. However, the way of assortment attractiveness can be increased is quite different. Caro et al. improve the value of the assortment by introducing new products $i \in I$, at a date specified up-front. In contrast, Çınar and Martínez-de-Albéniz can increase the attractiveness of an existing category $i \in I$, continuously and as a function of the latest information about how much decay there has been in category i's attractiveness. The model provides some insights that are complementary to Caro et al. (2012). When category margins are identical, the problem is well behaved. Again, products that decay less will be used early in the season, even if their introduction cost is higher, while products that are cheaper but decay faster should be used more at the end of the season.

The two models above open a number of interesting research opportunities. Mainly, the nature of dynamic demand evolution needs to be better understood. Real data shows that indeed individual product sales decrease over time, as new products are introduced into the assortment. However, the detailed process of how this happens is unclear: is the age of the product the determinant decay factor? Or is it because of the decrease of inventory availability over time, as Lago et al. (2013) suggest? Furthermore, there are other drivers of demand that need to be incorporated to the demand model, such as pricing or display. The increasing amount of available point-of-sales data should definitely spark more empirical work on these questions.

4 Pricing Strategy and Fashionable Design

Fast-fashion retailers mostly sell products at affordable prices—i.e., they sell "inexpensive fashion"—so the posted prices at different retailers are usually within the same price range.[5] Therefore, the main difference in pricing strategies across fast-fashion retailers is whether they use in-season promotions and markdowns or not. H&M is an example of the former whereas Zara follows the latter and avoids price changes during the selling season. Regardless of the in-season policy, fast-fashion retailers usually have well-announced clearance sales at the end of the regular season in which markdowns are introduced to liquidate stock and free up space for the new season.

The theoretical research on pricing for fast fashion has centered on price positioning and pricing strategies. On the former, Caro and Martínez-de-Albéniz (2012) present a model in which firms compete on price and product "freshness". Specifically, an inter-temporal utility model is introduced to account for product satiation. The satiation effect is incorporated through a retention factor that captures the carryover effect of consumption from one period to the next. In plain words, the retention factor measures how fast the consumer is willing to consume again. Offering a less satiating product—i.e., one with a lower carryover effect—is costly but it attracts more customers. When firms are symmetric, it is shown that there is a product strategy that is mostly dominant and firms can essentially ignore competition. However, this no longer holds if a firm breaks the symmetry by improving its processes to offer a fresher product. An important finding is that firms price incorrectly and are worse off when they ignore product satiation. Moreover, firms should aim at developing capabilities to offer less satiating products more efficiently, but since all firms have the same incentive, major improvements might be needed to guarantee an increase in profits.

[5] Note that H&M, and especially Zara, have deviated from the "affordable" pricing strategy to enter Asian countries—most notably Japan and China—where they are perceived as high-end European brands that signify status and therefore consumers are willing to pay a price premium.

Interestingly, depending on the current cost structure and the magnitude of the improvements, all firms can be better off after a "product war". This result is in contrast to price wars, which always hurt profits. Caro and Martínez-de-Albéniz (2009) present a variation of this model that relates satiation to assortment rotation, which is how fast-fashion retailers counteract product satiation in practice.

A separate stream of literature has focused on how to price fashion or seasonal products when consumers are forward-looking, in the sense that they anticipate the usual markdown policy used by retailers and might wait until prices goes down. The consumers' logic is quite simple: if nobody buys early, then the retailer will be forced to decrease prices. Su and Zhang (2008) show that a price commitment strategy in which the retailer makes a credible commitment not to lower prices can be effective in deterring consumers' strategic behavior. An alternative and equally effective strategy is allowing markdowns but rationing capacity (Liu and van Ryzin 2008). The latter resembles Zara's practice of having limited production to create shortages and induce consumers to buy at the regular season price. In the same vein, Liu and van Ryzin (2011) study rationing strategies when consumers can learn over repeated seasons and Yin et al. (2009) analyze strategies that restrict inventory display in order to create a perceived sense of scarcity.

Fashionable design is the last element of fast fashion that has not been discussed so far. This subject has been almost absent in the operations literature, and for a good reason since design is the part of retailing that has remained closer to an art rather than a science, at least until now. One paper that does deal with design at a high level is Cachon and Swinney (2011). This paper looks at whether the quick-response and (enhanced) design capabilities of a fast-fashion retailer are strategic complement or substitutes under the presence of forward-looking consumers. Though there are some exceptions, for the most part the paper shows that the two elements are strategic complements, which confirms that fast fashion is really an "all or nothing" proposition.

The economics and marketing literature has delved further into the drivers and dynamics of fashion. Sproles (1981) provides a comprehensive survey of the different—and sometimes competing—perspectives that try to explain the "fashion process". These perspective differ on the level at which the fashion process takes place (individual or societal) and whether the factors driving the process are endogenous or exogenous. Miller et al. (1993) categorize the different perspectives in a conceptual framework, which they formalize mathematically in a system of difference equations that are able to explain several of the fashion trends described in the literature. Pesendorfer (1995) provides an alternative model of fashion cycles in which fashion designs are used as a signaling device in a matching game. Consumers adopt fashions to show that they are "in" and the widespread adoption leads to lower prices, giving the firm selling fashion an optimal time for innovation. Kuksov and Wang (2013) build on the signaling idea and show that in equilibrium consumers randomize over designs, which explains fashion's "unpredictability". From an empirical standpoint, not too many attempts have been made to validate the theoretical findings. A few exceptions are Yoganarasimhan (2012) and

Martínez-de-Albéniz and Sáez-de-Tejada (2014) who use decades of data to analyze the presence of fashion cycles in the choice of names for newborns and Nunes et al. (2012) who study how fashion designs evolve based on the feedback from critics and reviewers. The lack of data is frequently cited as a reason that has prevented further empirical studies, but this is likely to change with the recent surge of social media where fashion dynamics can be tracked more easily (e.g., see Wang et al. 2013).

5 The Evolution of Fast Fashion

We began this chapter by noting that fast fashion has changed the industry dynamics significantly in recent years. We have outlined the set of practices that characterize fast fashion: sourcing with quick response and assortment planning with dynamic in-season introductions. Beyond these intrinsically operational levers, fast-fashion retailers have adopted alternative pricing and product strategies. We have discussed in detail all these elements in this chapter. But this overview would not be complete without a discussion on the current trends around the fast-fashion phenomenon, as well as the related research questions that arise from its evolution. Indeed, the fast-fashion model keeps evolving. There are numerous trends that retailers must take into account and that are affecting the operational implementation of fast fashion.

5.1 Leveraging Business Analytics

Business analytics is one trend that seems poised to grow in importance. It has gained notoriety with the copious amount of data that has become available lately, but the underlying concepts and techniques are not new to retailing. Good examples include Smith et al. (2001) and Fisher and Raman (2010). Though data-driven decision making is arguably relevant to any retailer, it is becoming a necessity for fast-fashion retailers that want to excel operationally, and in particular want to scale their internal processes to sustain continued growth. Zara, for instance, has taken up the challenge and since 2005 it has embedded model-based decision making into its daily operations. Caro et al. (2010) and Caro and Gallien (2010) describe a model developed and implemented at Zara to optimize the allocation of scarce inventory across its global network of stores. An interesting feature of the model—and quite unique to Zara—is how the model accounts for the interaction between the inventory levels of the different sizes of a given garment. The model aims at keeping the key sizes in stock to avoid negative customer perception and to ensure that the overall product remains on display. The use of the model led to a 3–4 % increase in sales.

Zara has also ventured into business analytics to optimize clearance sales. Caro and Gallien (2012) describe in detail the implementation of a model-based decision

support system for markdowns at Zara. Though this is a classic revenue management problem, there are at least two distinguishing characteristics: (a) the model considers multiple items which contrast with most of the literature that focuses on a single item; and (b) the lack of in-season price response data poses a challenge that is overcome by leveraging past season data combined with an adaptive procedure. The model was tested in a controlled field experiment with a symmetric design in which half of the assortment in Ireland was priced using the model and half was priced manually. The same happened in Belgium but with the opposite pricing methods. The rest of Western Europe was priced manually and was used a baseline. Using double differences to control for confounding effects, it is shown that the model increased clearance revenue by 6 %, which amounted to $90M in 2008.

Despite some isolated efforts, there is room for more research focused on business analytics in fast fashion. In particular, it would be interesting to see how business analytics can enhance the fundamental operational capabilities that define fast fashion, even more so as retailing evolves rapidly and steadily to cater to omni-channel consumers.

5.2 Creating or Following Fashion Trends

The most intriguing changes are happening in the design space. What initially gave birth to the fast-fashion model was the rapid and unpredictable changes of what customers want. These fickle trends are getting more numerous and shorter. Thus, quickly identifying a nascent trend becomes vital to retailers. Currently, fast-fashion players rely mostly on own sales data and competitor intelligence—i.e., paying attention to their new releases, in particular to determine whether these are successful—as an input for design. But this means that the original design decision, whether it was internal or at a competitor, was a wild guess that was not customer-driven. This may change: we have seen some design crowdsourcing platforms appear, a form of open innovation (Salfino 2013). For example, Threadless was started in 2000 and now boasts a community of over two million creators that can post their print designs on the Threadless website. Each week, the company selects the most voted designs for production, i.e., printing over T-shirts, hoods, tops, etc. The designer is rewarded with USD 2,000, plus additional payments for every reprint (Pozin 2012). Over 500,000 designs have been submitted to date and 1 % of them have been chosen for production. ModCloth uses a similar model, except that designs are not only prints, but full product specifications including fabric, cut, etc. This online retailer was started in 2002, and currently gathers 700 independent designers and suppliers, who create and keep ownership of original product designs. Once a design is ready, it is posted on modcloth.com and online customers can rate it. Successful products are then manufactured; this task's responsibility falls on the designers/suppliers (Indvik 2013). Similar initiatives have been tried out of apparel retailing too. The Danish toy company Lego experienced in 2006–2012 with

DESIGN byME, an online platform where users could submit their brick construction designs and Lego would custom produce them (Lego 2012). Popular designs could then inspire mass production designs. Furthermore, it is worth noting that the examples above introduce a pure pull logic into the design process, where design is only approved after sufficient people have endorsed it.[6]

Models with a clear push logic also exist. For example, JustFab is a subscription service for shoes and accessories where users initially take a test to learn their fashion preferences, and later on are offered customized assortments that fit their tastes (Chang 2011). The company's role is thus to curate new designs that each user will like. Since the assortment is constantly renewed and prices are rather low, some investors have called this subscription model "the new fast fashion" (Reuters 2013). Another business model known as *flash sales* also has a push logic and borrows elements of fast fashion. Flash sale websites offer "one deal a day" in which a selection of fashion items are sold at a discount for a very short period of time (usually less than a day). Imposing a narrow time window serves the same purpose than limiting inventory: it creates a perceived sense of scarcity and stimulates impulsive buying. Numerous websites—e.g., Zulily, Gilt Groupe, Ideeli, Net-a-Porter, Vente Privée or Privalia—adopted this business model; so many, that the market could be drying up (Roof 2014).

From a research perspective, these changes open numerous research opportunities. Models can be developed to understand what is the best way to capture demand trends. Clearly, different approaches have different impacts in terms of demand forecast accuracy (e.g., using votes or "likes" from Facebook provides a less accurate picture than pre-orders with full payment), reach (e.g., online will reduce access costs to the consumers but will also be less targeted than physical displays at a store) and costs (e.g., virtual displays are cheaper than real samples that require production). There are also interesting problems regarding the allocation of costs and profits, especially when retailers are the ones collecting revenues while designers are incurring the fixed costs of design, and design quality is hard to codify, so engineering effective incentive systems is a challenge.[7] Finally, understanding better how consumers dynamically choose between current styles and future ones is another interesting direction of work (Lobel et al. 2013; Bernstein and Martínez-de-Albéniz 2014).

5.3 Sourcing and Corporate Social Responsibility

There are also various developments on the production side of fast fashion. Deciding where to produce a garment usually depends on three aspects: (a) there are

[6] In manufacturing, a pull system is make-to-order, whereas a push system is make-to-stock.

[7] Chan et al. (2013) present a method to codify and identify styles in product designs. It works well for design patents, but it might be less applicable to fashion due to the lack of IP protection.

technical capabilities that are product-specific, e.g., treatment of leather requires significant expertise and access to water; (b) lead time requirements may eliminate some possible sourcing origins, although nowadays air transportation has mostly removed such constraints; and finally (c) cost competitiveness, including materials costs, energy costs, wages and freight charges, provides the last and perhaps most important element for decision-making. Thus, determining the optimal sourcing strategy becomes a complex task, especially when most of these factors change over time. For example, wage developments in China are triggering the offshoring of production to countries such as Vietnam, Cambodia or Bangladesh (Roland Berger 2011).

Offshoring for purely economic motives raises ethical questions: it is not always clear that working conditions are appropriate. For instance, the Rana Plaza factory collapse in April 2013 showed that workplace safety standards were not being followed (The Economist 2013). Moreover, the search for low costs is usually credited as one of the reason that has pushed factories into non-compliance, with consumers' appetite for fast fashion getting much of the blame (Lamson-Hall 2013). In fact, the Rana Plaza incident immediately put H&M on the spot for being the largest exporter of clothing from Bangladesh, even though it was not directly involved with that factory (Kerppola et al. 2014). Fast-fashion retailers have been taking note, and in response are developing corporate social responsibility (CSR) policies, e.g., Inditex has a code of conduct and responsible practices, and a committee of ethics, see Inditex (2012). It is not clear how to implement such CSR measures and what control mechanisms and incentives work best. Indeed, even when CSR policies exist, they are difficult to enforce, especially when there is limited visibility as work is offshored and subcontracted. Laudal (2010) identifies sector-specific variables that drive the risk of violating CSR standards, which suggests that regulation may be more effective than individual-firm actions. Besides literature in business ethics, there is some nascent research in operations on these subjects—including Babich and Tang (2012), Guo et al. (2013) and Kim (2013)—but much more is needed.

These ethics concerns are starting to be shared by some consumers. Siegle (2011) and Cline (2012) point out that fast fashion is unsustainable by nature as it encourages disposability, low durability, low quality, the loss of craftsmanship and ultimately uniformity. Some hard indicators can support this observation, e.g., Allwood et al. (2006) point out that consumers in the United Kingdom throw away 30 kg of clothing and textiles per capita each year, on average. Beyond economics, in a review of Siegle's book, Anderson (2011) states that "our bulimic passion for fashion is symptomatic of a broader malaise. Disposability, instant gratification, the idea that impulses are there be indulged, regardless of impact—these sentiments permeate our lives." Some retailers are taking a similar position. For instance, Zady states that it "began with a grand vision: to combat the fast-fashion craze by providing a platform for only those companies that care about timeless style and solid construction" (Zady.com 2013); it sells products with a

traceable origin. Adidas is supporting a community project in Brazil to design bags and caps with favela-inspired graphics (Clarke 2013). These critiques of fast fashion raise the question of how to make the entire business model more sustainable. Recycling is one option (Salfino 2014). From the research standpoint, there is already some work on this topic, e.g., Choi (2013b) examines how to use carbon footprint taxation to encourage local sourcing. But this is a broad research line that should be further explored, in connection with the work on closed-loop supply chains (Daniel et al. 2002).

Furthermore, if retailers continue to search for the current-day lowest-cost options, garment manufacturers choosing to close down high-wage operations and ramp up low-wage ones will experience inefficient investments (capacity installation, employee training and skill development). And it is not only a matter of costs: moving away from a region may have irreversible consequences. For instance, we have worked with an Italian jeans manufacturer that can no longer source and treat denim fabrics in Italy because most of the suppliers disappeared during the offshoring waves in the 1990s and 2000s. Similarly, there are few suppliers with QR capabilities left in Spain, after most retailers moved their QR operations to Portugal, North Africa and East Europe. It thus seems necessary to shape dynamic sourcing strategies that pay attention to cost dynamics and longer term implications, i.e., that a region's capabilities are being shaped by the retailer's sourcing decisions.

5.4 Beyond Apparel

We would like to conclude this chapter by discussing how fast-fashion practices can be extended beyond apparel retailing. The general ideas behind this phenomenon apply to any industry where numerous new products appear every day and consumers are searching for novelty. One such industry is food (grocery stores and restaurants). There, the fast-fashion formula would amount to changing offers and menus to satisfy customers' desire for new tastes and to providing the items from on-the-spot sources, as opposed to long-planned supplies, e.g., fresh preparations where ingredients are combined at the last minute. Some companies already have such capabilities, e.g., Seven Eleven Japan (Matsuo and Ogawa 2007). Another such example could be consumer electronics. A fast-fashion electronics manufacturer or retailer would have to significantly reduce the time between new product introductions, and be able to install flexible production capacity so as to respond quickly to demand, with low supply chain inventories. Interestingly, releases of smartphones have been more and more frequent, and product upgrades have less to do with technology breakthroughs and more with simple added functionalities and aesthetics (Knowledge @ Wharton 2013). Many other industries may also be ripe for a fast-fashion revolution.

Acknowledgements The authors thank Pol Boada Collado for helping collect the data used in Figs. 9.6 and 9.7. V. Martínez-de-Albéniz's research was supported in part by the European Research Council—ref. ERC-2011-StG 283300-REACTOPS and by the Spanish Ministry of Economics and Competitiveness (Ministerio de Economía y Competitividad, formerly Ministerio de Ciencia e Innovación)—ref. ECO2011-29536.

References

Abernathy, F. H., Dunlop, J. T., Hammond, J. H., & Weil, D. (1999). *A stitch in time: Lean retailing and the transformation of manufacturing–lessons from the apparel and textile industries*. Oxford: Oxford University Press.

Abernathy, F. H., Volpe, A., & Weil, D. (2006). The future of the apparel and textile industries: Prospects and choices for public and private actors. *Environment and Planning A, 38*, 2207–2232.

Agrawal, N., Smith, S. A., & Tsay, A. A. (2002). Multi-vendor sourcing in a retail supply chain. *Production and Operations Management, 11*(2), 157–182.

Ainslie, A., Drèze, X., & Zufryden, F. (2005). Modeling movie life cycles and market share. *Marketing Science, 24*(3), 508–517.

Allwood, J. M., Laursen, S. E., de Rodríguez, C. M., & Bocken, N. M. P. (2006). *Well dressed? The present and future sustainability of clothing and textiles in the United Kingdom*. Cambridge: University of Cambridge Institute for Manufacturing.

Alptekinoglu, A., Honhon, D., & Ulu, C. (2011). Learning consumer tastes through dynamic assortments. *Operations Research, 60*(4), 833–849.

Anderson, H. (2011). To die for: Is fashion wearing out the world? By Lucy Siegle review. *The Guardian*, June 12 online.

Babich, V., & Tang, C. (2012). Managing opportunistic supplier product adulteration, deferred payments, inspection, and combined mechanisms. *Manufacturing & Service Operations Management, 14*(2), 301–314.

Bernstein, F., & Martínez-de-Albéniz, V. (2014). *Using product rotation to induce purchases from strategic consumers*. Working Paper, IESE Business School.

Cachon, G., & Swinney, R. (2009). Purchasing, pricing, and quick response in the presence of strategic consumers. *Management Science, 55*(3), 497–511.

Cachon, G., & Swinney, R. (2011). The value of fast fashion: Quick response, enhanced design, and strategic consumer behavior. *Management Science, 57*(4), 778–795.

Caldentey, R., & Caro, F. (2010). *Dynamic assortment planning*. Working Paper, UCLA.

Calvo, E., & Martínez-de-Albéniz, V. (2012). *Sourcing strategies and supplier incentives for short life-cycle goods*. Working Paper, IESE Business School.

Caro, F. (2012). *Zara: Staying fast and fresh*. Technical Report, The Case Center. Reference number 612-006-1.

Caro, F., & Gallien, J. (2007). Dynamic assortment with demand learning for seasonal consumer goods. *Management Science, 53*(2), 276–292.

Caro, F., & Gallien, J. (2010). Inventory management of a fast-fashion retail network. *Operations Research, 58*(2), 257–273.

Caro, F., & Gallien, J. (2012). Clearance pricing optimization for a fast-fashion retailer. *Operations Research, 60*(6), 1404–1422.

Caro, F., Gallien, J., Díaz, M., García, J., Corredoira, J., Montes, M., et al. (2010). Zara uses operations research to reengineer its global distribution process. *Interfaces, 40*(1), 71–84.

Caro, F., & Martínez-de-Albéniz, V. (2009). The effect of assortment rotation on consumer choice and its impact on competition. In S. Netessine & C. Tang (Eds.), *Operations management models with consumer-driven demand*. Dordrecht: Springer.

Caro, F., & Martínez-de-Albéniz, V. (2010). The impact of quick response in inventory-based competition. *Manufacturing & Service Operations Management, 12*(3), 409–429.
Caro, F., & Martínez-de-Albéniz, V. (2012). Product and price competition with satiation effects. *Management Science, 58*(7), 1357–1373.
Caro, F., & Martínez-de-Albéniz, V. (2013). Operations management in apparel retailing: Processes, frameworks and optimization. *BEIO, Boletín de Estadística e Investigación Operativa, 29*(2), 103–116.
Caro, F., Martínez-de-Albéniz, V., & Rusmevichientong, P. (2012). *The assortment packing problem: Multiperiod assortment planning for short-lived products*. Working Paper, IESE Business School.
Chan, T., Mihm, J., & Sosa, M. (2013). *Identifying styles in product design*. INSEAD - Technology and Operations Management, Working Paper.
Chang, A. (2011). Online shoe clubs are in step with fashion-forward women. *Los Angeles Times*, December 29 online.
Chico's. (2012). Annual Report.
Choi, T.-M. (Ed.). (2013a). *Fast fashion systems: Theories and applications. Communications in cybernetics, systems science and engineering*. Leiden: CRC Press.
Choi, T.-M. (2013b). Local sourcing and fashion quick response system: The impacts of carbon footprint tax. *Transportation Research Part E, 55*, 43–54.
Christopher, M., Lowson, R., & Peck, H. (2004). Creating agile supply chains in the fashion industry. *International Journal of Retail and Distribution Management, 32*(8), 367–376.
Çınar, E., & Martínez-de-Albéniz, V. (2013). *A closed-loop approach to dynamic assortment planning*. Working Paper, IESE Business School.
Clarke, C. (2013). Two MBA graduates help promote artisans from Africa and Brazil. *Financial Times*, August 13 online.
Cline, E. L. (2012). *Overdressed: The shockingly high cost of cheap fashion*. New York: Portfolio/Penguin Group.
Daniel, V., Guide, R., Jr, & Van Wassenhove, L. N. (2002). *Closed-loop supply chains*. Berlin: Springer.
Doeringer, P., & Crean, S. (2006). Can fast fashion save the US apparel industry. *Socio-Economic Review, 4*, 353–377.
Euromonitor International. (2013). *GMID passport*. Retrieved September 17, 2013.
Farias, V. F., & Madan, R. (2011). Irrevocable multi-armed bandit policies. *Operations Research, 59*(2), 383–399.
Feitzinger, E., & Lee, H. (1997). Mass customization at Hewlett-Packard: The power of postponement. *Harvard Business Review, 117*(1), 116–121.
Ferdows, K., Machuca, J. A. D., & Lewis, M. (2002). *Zara*. Technical Report, ECCH case 603-002-1.
Fisher, M., & Raman, A. (2010). *The new science of retailing*. Boston: Harvard Business Press.
Fisher, M. L. (1997). What is the right supply chain for your product? *Harvard Business Review, 75*, 105–117.
Fisher, M. L., Rajaram, K., & Raman, A. (2001). Optimizing inventory replenishment of retail fashion products. *Manufacturing & Service Operations Management, 3*(3), 230–241.
Fisher, M. L., & Raman, A. (1996). Reducing the cost of demand uncertainty through accurate response to early sales. *Operations Research, 44*(1), 87–99.
Ghemawat, P., & Nueno, J. L. (2003). *Zara: Fast fashion*. Technical Report, Harvard Business School Multimedia Case 9-703-416.
Guo, R., Lee, H., & Swinney, R. (2013). *The impact of supply chain structure on responsible sourcing*. Stanford GSB Working Paper.
Hammond, J. H., & Kelly, M. G. (1990). *Quick response in the apparel industry*. Technical Report, Harvard Business School Note 9-690-038.

Hausman, W. H., & Thorbeck, J. S. (2010). Fast fashion: Quantifying the benefits. In T. C. E. Cheng & T.-M. Choi (Eds.), *Innovative quick response programs in logistics and supply chain management, international handbooks on information systems* (pp. 315–329). New York: Springer.

Heath, D. C., & Jackson, P. L. (1994). Modeling the evolution of demand forecasts with application to safety stock analysis in production/distribution systems. *IIE Transactions, 26* (3), 17–30.

H&M. (2007). Annual Report.

H&M. (2010). Annual Report.

H&M. (2011). Annual Report.

H&M. (2012). Annual Report.

Inditex. (2012). Annual Report.

Indvik, L. (2013). How ModCloth went from a college dorm to $100 million a year. *Mashable.com*, August 13 online.

Interbrand. (2013). Best global brands 2013.

Iyer, A. V., & Bergen, M. E. (1997). Quick response in manufacturer-retailer channels. *Management Science, 43*(4), 559–570.

Kerppola, M., Moody, R., Zheng, L., & Liu, A. (2014). *H&M's global supply chain management sustainability: Factories and fast fashion*. Technical Report, GlobaLens Case, Univesity of Michigan.

Kim, S. (2013). *Time to come clean? Disclosure and inspection policies for green production*. Yale School of Management, Working Paper.

Knowledge @ Wharton. (2013, September 11). *Still hot or not? Technology firms face faster product cycles*. Accessed January 23, 2014, from http://knowledge.wharton.upenn.edu/article/still-hot-or-not-technology-firms-face-faster-product-cycles/

Kök, A. G., Fisher, M., & Vaidyanathan, R. (2015). Assortment planning: Review of literature and industry practice. In N. Agrawal & S. Smith (Eds.), *Retail supply chain management* (2nd ed.). New York, NY: Springer Science+Business Media.

Krishnan, H., Kapuscinski, R., & Butz, D. (2010). Quick response and retailer effort. *Management Science, 56*(6), 962–977.

Kuksov, D., & Wang, K. (2013). A model of the "it" products in fashion. *Marketing Science, 32*(1), 51–69.

Lago, A., Martínez-de Albéniz, V., Moscoso, P., & Vall, A. (2013). *The role of quick response in accelerating sales of fashion goods*. Working Paper, IESE Business School.

Lamson-Hall, P. (2013). The rise of fast fashion pressures factories into non-compliance. *Sourcing Journal Online*, September 30 online.

Laudal, T. (2010). An attempt to determine the CSR potential of the international clothing business. *Journal of Business Ethics, 96*(1), 63–77.

Lee, H., & Tang, C. (1997). Modelling the costs and benefits of delayed product differentiation. *Management Science, 43*(1), 40–53.

Lego. (2012). *What happened to DESIGN byME?* Accessed January 23, 2014, from http://ldd.lego.com/en-us/subpages/designbyme/

Lewis, M. A., Machuca, J. A., & Ferdows, K. (2004). Rapid-fire fulfillment. *Harvard Business Review, 82*(11), 104–110.

Lin, Y.-T., & Parlaktürk, A. (2012). Quick response under competition. *Production and Operations Management, 21*(3), 518–533.

Liu, Q., & van Ryzin, G. (2008). Strategic capacity rationing to induce early purchases. *Management Science, 54*(6), 1115–1131.

Liu, Q., & van Ryzin, G. J. (2011). Strategic capacity rationing when customers learn. *Manufacturing & Service Operations Management, 13*(1), 89–107.

Lobel, I., Patel, J., Vulcano, G., & Zhang, J. (2013). *Optimizing product launches in the presence of strategic consumers*. Working Paper, NYU Stern.

Martínez-de-Albéniz, V. (2011). Using supplier portfolios to manage demand risk. In P. Kouvelis, O. Boyabatli, L. Dong, & R. Li (Eds.), *Handbook of integrated risk management in global supply chains* (pp. 425–445). Hoboken: Wiley.

Martínez-de-Albéniz, V., & Sáez-de-Tejada, A. (2014). *Dynamic choice models for newborn name preferences*. IESE Working Paper.

Matsuo, H., & Ogawa, S. (2007). Innovating innovation: The case of Seven-Eleven Japan. *International Commerce Review, 7*(2), 104–114.

McAfee, A., Dessain, V., & Sjöman, A. (2004). *ZARA: IT for fast fashion*. Technical Report, Harvard Business School Case 9-604-081.

Miller, C. M., McIntyre, S. H., & Mantrala, M. K. (1993). Toward formalizing fashion theory. *Journal of Marketing Research, 30*(2), 142–157.

Nasser, S., & Turcic, D. (2013). *To commit or not to commit: Revisiting quantity vs. price competition in a differentiated industry*. Olin Business School, Washington University in St. Louis, Working Paper.

Nunes, J., Drèze, X., Cillo, P., Prandelli, E., & Scopelliti, I. (2012). *The end of designer as dictator: How fashion critics affect aesthetic innovation*. University of Southern California Working Paper.

Palmer, J. W., & Markus, M. L. (2000). The performance impacts of quick response and strategic alignment in specialty retailing. *Information Systems Research, 11*(3), 241–259.

Pesendorfer, W. (1995). Design innovation and fashion cycles. *The American Economic Review, 85*(4), 771–792.

Pozin, I. (2012). How to start a business without really trying. *Inc.com*, June 28 online.

Reuters. (2013, September 26). *JustFab sews up $40M to become a global fast-fashion empire*. Accessed January 23, 2014, from http://www.reuters.com/article/2013/09/26/idUS228927694220130926

Roland Berger. (2011, December). *The end of the China cycle?* Accessed January 23, 2014, from http://www.rolandberger.com/media/pdf/Roland_Berger_End_of_China_cycle_short_version_20120104.pdf

Roof, K. (2014). Groupon buys Ideeli at a discount. *Forbes*, January 4 online.

Rusmevichientong, P., Shen, Z. J. M., & Shmoys, D. B. (2010), November. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations Research, 58*(6), 1666–1680.

Salfino, C. (2013). Apparel industry hears crowdsourcing's roar. *Sourcing Journal Online*, August 29 online.

Salfino, C. (2014). Recycling denim at retail adds life to old jeans. *Sourcing Journal Online*, February 27 online.

Sauré, D., & Zeevi, A. (2013). Optimal dynamic assortment planning with demand learning. *Manufacturing & Service Operations Management, 15*(3), 387–404.

Siegle, L. (2011). *To die for: Is fashion wearing out the world?* London, UK: Fourth Estate.

Smith, S., Agrawal, N., & Tsay, A. (2001). SAM: A decision support system for retail supply chain planning for private label merchandise with multiple vendors. In P. M. Pardalos, H. E. Romeijn, & J. Geunes (Eds.), *Supply chain management: Models, applications and research directions*. New York: Springer.

Song, J.-S., & Zipkin, P. H. (2012). Newsvendor problems with sequentially revealed demand information. *Naval Research Logistics (NRL), 59*(8), 601–612.

Sproles, G. B. (1981). Analyzing fashion life cycles: Principles and perspectives. *Journal of Marketing, 45*(4), 116–124.

Su, X., & Zhang, F. (2008). Strategic customer behavior, commitment, and supply chain performance. *Management Science, 54*(10), 1759–1773.

The Economist. (2013, May 4). *Workplace safety: Avoiding the fire next time*. Accessed January 23, 2014, from http://www.economist.com/news/business/21577078-after-dhaka-factory-collapse-foreign-clothing-firms-are-under-pressure-improve-working

Wang, J., Aribarg, A., & Atchadé, Y. F. (2013). *Modeling choice interdependence in a social network*. Stephen M. Ross School of Business, University of Michigan, Working Paper.

Yin, R., Aviv, Y., Pazgal, A., & Tang, C. S. (2009). Optimal markdown pricing: Implications of inventory display formats in the presence of strategic customers. *Management Science, 55*(8), 1391–1408.

Yoganarasimhan, H. (2012). *Identifying the presence and cause of fashion cycles in the choice of given names*. UC Davis Working Paper.

Zady.com. (2013). *Mission statement*. Accessed January 23, 2014, from https://zady.com/about/mission

Chapter 10
Managing Variety on the Retail Shelf: Using Household Scanner Panel Data to Rationalize Assortments

Ravi Anupindi, Sachin Gupta, and M.A. Venkataramanan

1 Introduction

Two fundamental retailer decisions are which items to stock in a category (the assortment decision) and how much to stock of each item (the inventory decision). While these decisions have always been key to retailer profitability, they have received renewed attention because of industry initiatives labeled Efficient Consumer Response (ECR). Category Management, a component of ECR, emphasizes the need to recognize the inter-relatedness (e.g., substitutability) of items within a category when making decisions. Thus, categories need to be managed as strategic business units, with an emphasis on total category performance. Point-of-sale information can potentially play a critical role in providing insights into consumer behavior to help develop sound category strategies.

Retailers recognize that wider assortments help their business by catering to the needs of multiple consumer segments (Coughlan et al. 2006), as well as by offering variety to variety-seeking consumers. However, there are limits to the value of variety. Adding items with small differences offers little in the way of "real" variety to the consumer (Boatwright and Nunes 2001), yet adds to costs of operations such as

R. Anupindi (✉)
David B Hermelin Professor of Business Administration, Professor of Technology and Operations, Stephen M. Ross School of Business, University of Michigan, Ann Arbor, MI, USA
e-mail: anupindi@umich.edu

S. Gupta
Director, Graduate Studies and Henrietta Johnson Louis Professor of Management, Johnson Graduate School of Management, Cornell University, Ithaca, NY, USA

M.A. Venkataramanan
Vice Provost for Strategic Initiatives, Jack R. Wentworth Professor, School of Business, Indiana University, Bloomington, IN, USA

N. Agrawal, S.A. Smith (eds.), *Retail Supply Chain Management*,
International Series in Operations Research & Management Science 223,
DOI 10.1007/978-1-4899-7562-1_10

administrative costs and cost of warehouse space. The sharp growth of warehouse clubs and deep discount drug stores in recent years is attributed, in part, to their cost advantages arising from their limited variety offering. The resultant loss of market share has re-focussed attention of supermarkets on the need to manage variety. It is believed that there is substantial potential for lowering supermarket operating costs without hurting business by making store assortments more efficient; see, for example, a report by the Food Marketing Institute (1993).

Managing retail space entails solving two types of problems. The first is allocating space to categories, called the *inter-category* space allocation problem. The second is allocating space to items within a category or the *intra-category* space allocation problem. This second problem is often referred to as the assortment problem. Ideally, assortment decisions need to incorporate a variety of factors. On the demand side, one needs to consider the (heterogenous) customer purchase behavior including substitution patterns when their preferred items are not available (either temporarily due to stock-out or permanently due to limited assortment), the stochastic nature of demand arising due to the uncertainty inherent in consumer choice, the effect of product display on sales, etc. On the supply side, retailers face a finite shelf-space constraint for a category and incur fixed costs to include items in the assortment. Further, since limited assortments may have longer term consequences on profitability, a retailer needs to balance current profits with implications of the assortment on future profits. Finally, such a model for decision making should be driven by actual data and the solution strategy should be scalable to address the large problem sizes that any realistic assortment decision would entail.

In this chapter, we outline a modeling framework that incorporates some of the above features to assist the retailer in determining the optimal subset of items to carry in a category, from the set currently carried, and the quantity to stock of each item. We propose the use of household purchase data collected via scanners to estimate intrinsic preferences of consumers and to infer their substitution patterns. Such information is key to ensuring that the assortment carried caters to heterogeneous consumers' tastes, while avoiding unnecessary and expensive duplication.

Previous research on the retailer's assortment problem has typically not modeled consumer substitution behavior explicitly. Empirical evidence from several studies suggests that in packaged goods markets, consumers are often willing to substitute a less preferred item for their (non-available) preferred item. A Food Marketing Institute survey reports that only 12–18 % of shoppers said they would not buy an item on a shopping trip if their favorite brand-size was not available; the rest indicated they would be willing to buy another size of the same brand, or switch brands. A number of other studies (Emmelhainz et al. 1991; Carpenter and Lehmann 1985; Urban et al. 1984; Gruen et al. 2002) support a similar conclusion. A 1993 study by Willard Bishop Consulting Ltd. and Information Resources, Inc. found that when duplicative items were removed, 80 % of consumers saw no difference (Business Week 1996). Other evidence suggests that consumers make about two-thirds of their purchase decisions about grocery and health-and-beauty

products while they are in the store (Nielsen Marketing Research 1992). Thus, it is important to take account of substitution behavior of consumers when rationalizing assortments.

It is likely that consumers who do not find their preferred item in the store assortment are not fully satisfied, whether or not they buy another item in the category. The decision to rationalize assortments needs to take account of the potential adverse impact on customer retention. Traditional formulations of the assortment problem typically assume that the retailer is a myopic profit maximizer. Such formulations disregard the longer-term adverse impact on profits of not satisfying consumers' demand for their preferred items. In our proposed formulation, the retailer's objective function is a weighted sum of profits and a penalty for disutility caused to consumers who do not find their preferred items in the assortment. The rationale for including a penalty is that dissatisfied customers may take their future business elsewhere, thereby hurting longer term profits, even if they purchase less preferred items in the current period. Our proposed model can be used by a retailer to balance short term profits and customer disutility when choosing assortments.

Another contrast of our proposed approach with previous research lies in our accommodation of differences in item preferences between consumers. Most previous work assumes an aggregate demand model. Aggregate demand specifications do not allow us to distinguish between the extent of disutility or dissatisfaction caused by not stocking a particular item to, for example, more versus less loyal groups of consumers. Clearly this distinction is relevant for a retailer who cares about retaining customers in the longer run. The existence of consumer heterogeneity has been established by a number of previous empirical studies. Our proposed model allows for completely idiosyncratic patterns of substitution, as well as disutility due to non-stocking, between consumers.

To demonstrate an empirical application of the proposed model, we estimate consumer preferences for eight items in the canned tuna category using household scanner panel data, a commonly available source of market research information. A hierarchical Bayesian approach is used to estimate an interval scaled measure of each household's utility for the eight items, and the household's price and promotion sensitivity. The retailer's decision problem is then solved as an integer programming problem. Although the problem is large in terms of the number of decision variables and constraints, we show that it can be solved efficiently. Our solution reveals that a significant reduction in customer disutility can be accomplished at the cost of a small reduction in the current period profits.

Our model should be considered as an illustrative first step. While we have captured the richness of customer heterogeneity, substitution behavior, and the current vs. future profit tradeoff, we also have made simplifying assumptions on other aspects of this complex problem. In Sect. 6 we outline several ways to enhance our proposed model to incorporate these remaining aspects, which we hope will inform further research in this important field.

The rest of the chapter is organized as follows. In Sect. 2 we review related research. We discuss the consumer model in Sect. 3. In Sect. 4 we develop an optimization framework for the assortment decision, discuss special cases and

some properties of the model. In Sect. 5, we demonstrate an empirical application of our proposed model using household panel data. We conclude in Sect. 6 with a brief summary and a discussion of extensions and further research.

2 Literature Review

Two broad streams of literature are relevant to this study—one in marketing, the other in operations management. Early research in marketing deals with issues of retail shelf space allocation and is empirical in nature. Corstjens and Doyle (1981) proposed a model to optimize space allocation across categories, given an overall store space constraint. Direct and cross space elasticities were measured via a multiplicative sales response model using cross-sectional data. Their model does not explicitly include the assortment decision, although allocation of zero space to an item may be interpreted as exclusion of the item. However, as pointed out by Borin et al. (1994), the multiplicative sales response model predicts zero sales for a given category if the space of any of the store's other categories is set to zero. Bultez and Naert (1988) and Bultez et al. (1989) model the intra-category space allocation problem. Space elasticities are measured experimentally with item sales as the criterion variable. However, the assortment decision is not explicitly modeled. Borin et al. (1994) incorporate both the space allocation and assortment decisions in a retailer model. However, this study does not empirically estimate the demand model. Instead, parameter values are assumed. More recently, van Dijk et al. (2004) use observed variation in shelf-space allocation across stores to infer shelf-space elasticities.

The focus of these studies is on allocation of a scarce resource—space—given that different items show varying responsiveness to space. Thus, emphasis is placed on methods and data for measurement of space elasticities (own and cross) and on algorithms to solve the retailer profit maximization problem efficiently. By contrast, our focus is on estimating consumers' brand preferences to infer their willingness to substitute, thereby determining the optimal assortment of items to stock. In the present study we do not tackle issues of responsiveness of demand to space allocations, but leave that for future research. The primary emphasis in our work is motivated by the empirical observation that in most consumer packaged goods categories, consumers can often be (imperfectly) satisfied by one of several items. This characteristic of consumer behavior is used in determining optimal assortments.

Recent empirical findings in the marketing literature provide strong support for the idea that assortment reductions may be profitable for retailers. Broniarczyk et al. (1998) conduct controlled lab experiments as well as field experiments in which assortments were reduced in five categories in convenience stores. They measure consumer perceptions of variety, which are shown to mediate store choice.

A key finding is that elimination of low-selling items had little or no impact on shoppers' perceptions of variety, as long as favorite items were available and category shelf space was held constant.

Boatwright and Nunes (2001) analyze data from a natural experiment conducted by an online grocer, in which 94 % of the categories experienced dramatic reductions in the number of SKUs offered. Sales increased an average of 11 % across the 42 categories examined.[1] An important finding especially relevant to our work is that customers who lose their favorite item when the assortment is reduced are significantly less likely to purchase in the category on a future purchase occasion.

Borle et al. (2005) use household panel data of the same online grocer that Boatwright and Nunes study to analyze the effects of assortment reductions in several categories on overall store sales. They find that although the effect is positive in several categories, overall store sales are reduced due to decreases in the number of store visits and the size of the shopping basket. To our knowledge, this is the first study that demonstrates that customer retention, i.e., customers' repeat store visit behavior, is adversely affected by reductions in category assortments.

Sloot et al. (2006) distinguish between short and long term sales effects of a 25 % item reduction in the assortment in one category. They find that while short-term category sales suffer a sharp reduction, long-term category sales display only a weak negative effect.

The findings of both Broniarczyk et al. (1998) and Boatwright and Nunes (2001) highlight that the impact of assortment reductions is heterogeneous across consumers, depending on the extent of loyalty exhibited towards the lost item. Borle et al. (2005) show conclusively that assortment reductions may reduce a shopper's probability of returning to this store on the next shopping visit. Although our data do not permit us to directly model the effects of assortment availability on consumers' store choice decisions, in our assortment optimization model we formalize the idea by including in the retailer's objective function the disutility incurred by consumers as a result of not finding their preferred items in the available assortment. This disutility is idiosyncratic to each consumer, and serves as a proxy for the reduced profits resulting from the lower probability of consumers choosing this retailer in future.

In the operations literature, work on assortment problems was motivated by the textile industry where decisions regarding which sizes (e.g., in-seam lengths for slacks) to carry had to be made. Pentico (1974) considers the single dimension assortment problem with probabilistic demands, with assumptions about substitution behavior of consumers. Pentico (1988) extends the earlier work to two-dimensional assortment problems with deterministic demands. Other related work deals with determining optimal stock levels for multiple items given stochastic demands and a pattern of substitution based on non-availability; see, for example, Bassok et al. (1997)

[1] Part of the increase is attributed to enhanced utility due to reduced clutter in the category. Our model does not allow for such an effect.

and the references therein. In this work, however, substitution is determined by the supplier firm and not by the buyer or consumer.

van Ryzin and Mahajan (1999) study a stochastic single period assortment planning problem under a Multinomial Logit (MNL) Choice model. A consumer's choice depends on the variants that the store carries and they assume that consumers do not substitute in the event of a stock-out. Using a newsvendor framework with identical exogenous retail prices across all variants, they show that the optimal assortment always consists of a certain number of the most "popular" products. They also illustrate that retail prices and profits increase when consumer preferences are more "fashion" oriented. In a follow-up paper, Mahajan and van Ryzin (2001) incorporate both assortment-based as well as stock-out based substitution behavior and present a stochastic sample path optimization method to solve for the optimal assortment. In contrast to these papers that assume a MNL model of choice, Gaur and Honhon (2006) use a locational choice model to study the assortment problem.

Smith and Agrawal (2000) study the assortment planning problem using a general probabilistic model of demand allowing for substitution behavior. Using a substitution matrix, they estimate the derived demand for a given assortment. They then present a methodology to determine the assortment and stocking levels jointly when retailers incur a fixed cost for carrying an item in stock as well as the classical inventory and shortage costs for excess inventory and shortage at the end of the period.

Some recent papers have focused on jointly addressing demand estimation as well as assortment planning. Chong et al. (2001) present a category assortment planning problem. Consumer choice is represented as a combination of a category-purchase-incidence model and a brand-share model. While the former predicts the probability of an individual consumer's purchase from a category on a given shopping trip, the latter predicts which brand will be purchased. The optimization problem then determines the optimal number of facings for the various products to maximize profits, subject to a shelf space constraint. They illustrate their methodology using data from five stores in eight food categories.

Kok and Fisher (2004) present a demand estimation as well as an assortment optimization model. Using cross-sectional data across stores that carry different assortments, they estimate the substitution behavior of a homogenous set of customers. Using a probabilistic model of choice, they posit an assortment optimization model and develop heuristics to determine the number of facings of a particular product that a retailer should carry. They apply their method to a supermarket chain in the Netherlands and illustrate that their methodology for assortment planning potentially leads to a 50 % increase in profits.

Miller et al. (2006) propose an approach to optimize retail assortments with demand specified as a multinomial logit model. Consumers' utilities for products are estimated via a conjoint approach wherein consumer heterogeneity is allowed. In an empirical application they find that there is a significant negative impact on profits when heterogeneous consumers are assumed to be homogeneous.

Like the papers just discussed, our chapter focuses on a joint demand estimation and assortment planning problem. Demand is modeled at the household level using

a discrete choice framework, specifically a probit model. Households are modeled as heterogeneous in unobserved utility function parameters, and the heterogeneity distribution is estimated using household scanner panel data. Thereby, posterior estimates of households' preference are derived.

The formulation of our optimization model is similar to the one studied by Dobson and Kalish (1988, 1993) in the context of positioning and pricing a product line. They present welfare and profit maximization formulations for positioning and pricing respectively. Our formulation is also similar to McBride and Zufryden (1988) who apply integer programming techniques to the optimal product line selection problem. Their model formulation recognizes heterogeneity in consumer preferences. Our approach of incorporating consumer disutility into the retailer's objective function is, however, more general than that of Dobson and Kalish (1993) or McBride and Zufryden (1988). The idea of penalizing the objective function for lost goodwill due to non-availability of stock is not new. In stochastic inventory theory (Arrow et al. 1958; Lee and Nahmias 1994) a penalty cost for shortages is routinely included in the objective function. However, to our knowledge, this chapter is the first to operationalize the penalty based on disutilities estimated from market-place data.

A key point of distinction between our paper and most of the literature discussed previously is with respect to the model of consumer heterogeneity. The classical multinomial logit (MNL) model as used in van Ryzin and Mahajan (1999) and Mahajan and van Ryzin (2001) allows for heterogeneity between consumers only via the stochastic term in the random utility. However, these differences between consumers are unobservable to the firm a priori, since the *expected* utility of a product is identical across consumers. This is why the model is sometimes referred to as the "homogeneous" MNL model. By contrast, we explicitly incorporate differences between consumers in the expected utility via a distributional assumption on the utility function parameters. The distribution of these parameters is then empirically estimated and can be used when determining the optimal assortment. Our approach is similar in theory to conjoint models (e.g., Miller et al. 2006) in which idiosyncratic utility functions are estimated.

3 Consumer Model

Our model of the retailer's decision problem of which items to carry and how much to carry, discussed at length in the next section, assumes that each consumer chooses that item from the available assortment which maximizes the consumer's utility. Solving this problem requires empirical estimates of consumers' preferences. We discuss in this section our approach to estimate consumer preferences.

Traditionally, data on consumer preferences have been collected via surveys as stated preferences (ordinal- or interval-scaled), or trade-offs that individuals would be willing to make on particular attributes (e.g., conjoint studies). An alternative approach is revealed preference data as obtained from reported or observed brand choices of consumers in actual purchase situations. For most product categories

in the grocery industry these data are readily available from syndicated sources (e.g., household panels of Nielsen and Information Resources Inc.). The primary advantage of stated preference data is the ability to measure preferences for items currently not stocked (in particular, for new products). The major disadvantages of stated preference data relative to brand choice data are potentially lower validity of the data, and often substantially higher cost of data gathering.

Since the focus of our empirical work is on assortment decisions for supermarket product categories, we consider a model to estimate preferences that can be applied to observed brand choices of consumers—a multinomial probit model of brand choice. The probit model can be derived by assuming that the utility a consumer obtains from purchasing an item in the category is composed of a deterministic component and a stochastic component. The stochastic component represents unobserved (to the researcher) components of utility. In the typical formulation of the brand choice model, the utility of item $j, j = 1, 2, \ldots, J$ to consumer i on occasion t is given by U_{ijt}, thus: $U_{ijt} = \tilde{V}_{ijt} + \varepsilon_{ijt}, \varepsilon_{ijt} \sim N(0, \Sigma)$ where

$$\tilde{V}_{ijt} = \tilde{\alpha}_{ij} - \beta_i p_{ijt} + \tilde{\gamma}_i X_{ijt} \tag{10.1}$$

where for consumer i and item j, $\tilde{\alpha}_{ij}$ is the *intrinsic* utility or valuation, p_{ijt} is the price of the item on occasion t, X_{ijt} represents other attributes of the item (such as in-store promotions) on that occasion, and β_i and $\tilde{\gamma}_i$ are parameters. The assumption that the stochastic term has a multivariate normal distribution leads to the multinomial probit model of brand choice. We use a diagonal covariance structure $\varepsilon_{ijt} \sim N(0, \Sigma)$ where Σ is a $J \times J$ diagonal matrix, coupled with the identifying restriction that the first diagonal element is one. The choice of diagonal covariance structure simplifies the calculation of choice probabilities, while obviating the restrictive IIA property associated with a scalar covariance matrix, as well as with a multinomial logit model.

Note that the parameters of the utility function are individual specific, thus allowing for heterogeneity in both the intrinsic preferences and the effects of price and other attributes. As we demonstrate subsequently, this characteristic of the model has important implications for the optimal assortment decision of the retailer. The objective of model estimation is to recover the unknown parameters of the deterministic component of the utility function. Data required to estimate the model are observations of consumer choices as well as prices and in-store promotional conditions on each purchase occasion. Such information is typically available in household scanner panel data.

We model heterogeneity by specifying a series of conditional distributions in a Hierarchical Bayesian fashion. The reader is referred to Imai and van Dyck (2005) and McCullogh and Rossi (1994) for details of the estimation approach. A key benefit of using this approach is that it yields posterior estimates of utility function parameters at the individual level. These estimated utility functions are inputs into the retailer's optimization problem.

To obtain item-specific intrinsic utilities, we assume that prices are determined exogenously.[2] Furthermore, for simplification they are assumed to remain constant at their observed mean level p_j. We also assume the in-store promotion variables are fixed at their average levels X_j, again for simplification. Since utility is linear in prices, we divide utilities by the estimated price coefficient β_i (Kalish and Nelson 1991) to obtain a $-metric utility, thus:

$$V_{ij} = \alpha_{ij} - p_j + \gamma_i X_j \tag{10.2}$$

where $\alpha_{ij} = \tilde{\alpha}_{ij}/\beta_i$, $\gamma_i = \tilde{\gamma}_i/\beta_i$ and p_j is the (constant) price of item j.[3]

The difference in $-utility between two items may be considered the cost of substituting one item for the other for the consumer; see Krishna (1992) and Bawa and Shoemaker (1987) for a similar notion of substitution costs. An alternative interpretation of this difference is the reduction in price of the less preferred item necessary to make the consumer indifferent between the two items.

We assume that a consumer is willing to substitute lower utility items when higher utility items are not carried in the retail assortment. This assumption is strongly supported by empirical studies (Urban et al. 1984; Emmelhainz et al. 1991). The order of substitution is described by the rank-ordering of estimated preferences for items. When such substitution occurs, however, the consumer is assumed to incur a disutility equal to the difference in $-metric of intrinsic utility between the most preferred item in the category and the item bought (i.e., the substitute item).

Empirical evidence also suggests that consumers may be willing to incur disutility due to downward substitution only upto a point. Below this point they may be unwilling to substitute and may choose to either postpone purchasing in the category or purchase at a different store (Borle et al. 2005). In an ideal setting, one would estimate the utility of a no-purchase decision and expect that consumers will be willing to substitute items as long as the utility of these items is above the utility for no-purchase. However, in the form they are currently available, household scanner panel data do not allow empirical estimation of the no-purchase threshold of households. Thus, in the subsequent empirical illustration we posit alternate mechanisms for operationalizing the no-purchase decision; we outline some options in Sect. 4.2.

Since the vector of intrinsic brand utilities is unique to each consumer, our consumer model allows completely idiosyncratic patterns of substitution. Not only is the highest preference brand allowed to be different across consumers, consumers who have a given brand as the most preferred may substitute a different brand in the event the most preferred item is not carried in the assortment. Such heterogeneity in substitution behavior between consumers has been documented in empirical

[2] In Sect. 6 of the chapter, as future research, we discuss the possibility of extending the model to determine optimal prices as well.

[3] The transformation of utilities by dividing by the price coefficient also serves to remove the influence of the unidentified scale factor that confounds the vector of parameter estimates (Swait and Louviere 1993).

studies (Emmelhainz et al. 1991). Furthermore, since we obtain an interval-scaled measure of preference, consumers who have exactly the same rank-ordering of brand preferences may incur differing amounts of disutilities due to non-availability of the most preferred item. This allows us to capture differences in intensities of brand preferences between consumers (e.g., loyals vs. switchers) that are relevant for the assortment and inventory decision.

To summarize, our model of the process consumers follow to choose an item to purchase in a category after entering the store is as follows. Consumers have preferences for various items in a category; these preferences vary from consumer to consumer. A consumer observes the available assortments (and the prices of items) and picks the highest utility item from those available or choses not to purchase. The exact operationalization of the no-purchase decision is discussed in the next section.

To use the consumer demand model in the retailer optimization problem, we revert to the utility measures V_{ij} in (10.2) (at constant prices) and use the estimated utilities $\widehat{V}_{ij}$. Disutilities form an important component of the retailer's objective function in our model, as detailed in the subsequent section. Ideally, we should use the random utility function U_{ijt} shown earlier. However, since U_{ijt} contains both a deterministic and a stochastic component, its use will lead to a potentially complex stochastic programming formulation. While accurate, this formulation does confound the impact of heterogeneity and probabilistic choice on the assortment decision. Instead, to focus exclusively on the heterogenous model of consumer behavior, we use only the deterministic component of the utility given by V_{ijt}. Our modeling choice is not without precedence; see Dobson and Kalish (1988, 1993) and McBride and Zufryden (1988). We comment on alternative approaches that could incorporate stochastic choice in the concluding section.

4 The Retailer Assortment and Stocking Problem

In this section, we describe a model to solve the retailer's assortment and stocking problem. We first develop a basic model that incorporates profits and disutility. We then discuss some special cases and properties of the formulation.

4.1 Basic Formulation

The retailer's problem can be defined as follows: We are given a set of N items indexed by j. There is a fixed cost of stocking each item. Consumers belong to one of the s index segments,[4] $s \in \{1, \cdots, S\}$. There exists a (monetary) utility

[4] The consumer model in Sect. 3 was developed assuming each consumer is a separate segment, i.e., the number of consumers in each segment is one. Other models of brand choice that provide estimates for "segments" of consumers could be employed, such as formulations of Kamakura and Russell (1989) and Chintagunta et al. (1991).

measurement, V_{sj}, for every segment s for every item j (see Sect. 3). As noted previously, for solving the retailer's optimization problem we assume that prices and promotional activities are held constant at their average levels. As a consequence, item utilities are time invariant. A consumer (segment) chooses from all available items the one that maximizes its utility.[5] The retailer's problem is to select an assortment and determine the stock for items in the assortment to maximize profits. The profit function can be written as:

$$PR(\mathbf{x}, \mathbf{y}) = \sum_j \left[\sum_s (p_j - c_j) x_{sj} n_s - K_j y_j \right] \tag{10.3}$$

where p_j is the per unit (regular) price of item j, c_j is the per unit variable cost of stocking item j, x_{sj} is a 0–1 variable which takes on a value of one if segment s customers are assigned to item j and zero otherwise (a decision variable),[6] n_s is the number of consumers in segment s, K_j is the fixed cost of stocking item j, and y_j is a 0–1 decision variable which takes the value one if item j is stocked and zero otherwise. Finally, $\mathbf{x}$ is a $S \times N + 1$ matrix of x_{sj} and $\mathbf{y}$ is an $N + 1$-vector of y_j. We let no-purchase decision be a "product" that is always available, thus expanding the product space to $N + 1$; further, $p_0 = c_0 = K_0 = 0$ and $y_0 = 1$.

Typically a retailer may do assortment planning for its stores twice a year; thus the planning horizon for assortments is about 6 months. In our formulation, we have not specified any planning horizon explicitly. The data can be scaled to accommodate any planning horizon. We need to, however, consider the fixed costs—which include costs relating to sourcing, supplier selection, negotiations, etc.,—appropriate for the planning horizon. Due to fixed costs of carrying an item in the assortment, not all items may be stocked. As a consequence, the following situations are possible:

1. A customer segment buys a less preferred item because its most preferred item is not available.
2. A customer segment does not purchase at all because no satisfactory item is available.

In either case the customer incurs a disutility. We postulate that such disutility adversely affects the customer's likelihood of repurchasing at this store, thereby affecting long-run profits.[7] We propose the following measure of customer disutility:

[5] We assume, for simplification, that each consumer buys exactly one unit in each restocking period. This assumption can be relaxed by weighting each consumer by the number of units bought. In general, the number of units bought by a consumer within any stocking period may be uncertain. Incorporating this uncertainty will result in a stochastic programming formulation. We elaborate upon this idea in the discussion of future work in Sect. 6.

[6] In the optimization model, the item "assigned" to a consumer will be the one that maximizes the consumer's utility. Thus, consumers will in effect self-select their best alternative from the available assortment.

[7] Notice that this disutility is due to non-stocking of items and not due to stock-out of an item.

$$DU(\mathbf{x}) = \sum_s n_s \left[\sum_k \{(V_{sj_1} - V_{sk})x_{sk}\} + (V_{sj_1} - V_{sj_0})(1 - \sum_k x_{sk}) \right] \tag{10.4}$$

where, $V_{sj_1} = \max_j\{V_{sj}\}$, and V_{sj_0} is the no-purchase utility, as discussed later in Sect. 4.2.

For those customers who are assigned an item k, the disutility is the difference between the utility of item k and their most preferred item.[8] Similarly, customers who do not purchase are also dissatisfied. The disutility incurred by these customers is the difference in utility between their highest utility and their utility for no-purchase. Clearly, customers who find their most preferred item in the assortment do not incur any disutility.

We propose that the overall objective function for a retailer should be a weighted combination of profits as measured by (10.3) and disutility as measured by (10.4). The extent to which a retailer should weight consumer disutility will depend on the product category. Customer dissatisfaction with some categories is likely to have a larger adverse impact on store choice. In the context of pricing, for example, Harris and McPartland (1993) classify categories into "traffic generators" (i.e., affect store choice) and others. We model this by taking a convex combination of the profit and disutility functions. Thus the objective function of the retailer is:

$$\Pi(\mathbf{x}, \mathbf{y}, w_c) = (1 - w_c)PR(\mathbf{x}, \mathbf{y}) - w_c DU(\mathbf{x}) \tag{10.5}$$

where $0 \le w_c \le 1$. w_c may be interpreted as a control or policy parameter whose value is to be subjectively determined by the decision maker.[9]

The optimization problem of the retailer is then written as follows:

$$(\text{P1}) \max_{\mathbf{x},\mathbf{y}} \Pi(\mathbf{x}, \mathbf{y}, w_c)$$

such that,

$$\sum_k V_{sk} x_{sk} \quad \ge \quad V_{sj} y_j \qquad \forall s, j \tag{10.6a}$$

[8] Dissatisfaction measured as sum across segments of the differences in utilities implies that a large number of small disutilities is equivalent to a small number of large disutilities; e.g., two segments with one unit of disutility each is equivalent to one segment (of same size) with two units of disutility. This may not be desirable since larger differences in utilities signify consumers *loyal* to certain brands, and smaller differences in utilities signify *switchers*. A non-linear (say, e.g., exponential) function of difference in utilities will allow us to distinguish between *loyals* and *switchers*.

[9] A similar objective function (weighted combination of profits and consumer utility) was also considered by Little and Shapiro (1980) in the context of pricing nonfeatured products in supermarkets. Similarly, there is extensive literature on bi-criterion optimization problems; see, for example, French and Ruiz-Diaz (1983).

$$\sum_j x_{sj} \leq 1 \quad \forall s \tag{10.6b}$$

$$x_{sj} \leq y_j \quad \forall s, j \tag{10.6c}$$

$$x_{sj} = 0, 1 \quad \forall s, j \tag{10.6d}$$

$$y_j = 0, 1 \quad \forall j \neq 0 \tag{10.6e}$$

$$y_0 = 1 \tag{10.6f}$$

Constraints (10.6a) ensure that of the items stocked, a customer is assigned his/her most preferred item. Constraints (10.6b) ensure that segment s is assigned to at most one item; finally, constraints (10.6c) ensure that only items that are offered are chosen by the customers.

At first glance it may appear that incorporating consumer disutility through $DU(\cdot)$ in the objective function makes constraints (10.6a) redundant. The constraints are redundant (or trivially satisfied) only when a retailer sets $w_c = 1.0$. Otherwise, in the absence of constraints (10.6a) it is possible that a retailer may assign a less preferred item (with a higher contribution margin) to a consumer even though a more preferred item (with a lower contribution margin) is stocked, albeit for a different consumer. Such an assignment is problematic from an implementation viewpoint in the context of supermarkets since a consumer walks into a store and necessarily picks his most preferred item if it is available. Constraints (10.6a) ensure that the retailer incorporates this fact into its decision making.

4.2 Modeling No Purchase

As discussed previously, a customer may decide to not purchase in the category if its preferred item is not stocked. Since scanner data do not report non-purchasing on account of unavailability in the assortment, we model this outcome and assume its value.[10] There are at least two ways one could model no purchase in the optimization problem. For a customer segment s, first rank order the utilities V_{sj} in decreasing order to write:

$$V_{sj_1} \geq V_{sj_2} \geq \cdots \geq V_{sj_N}.$$

[10] Category purchase incidence is frequently modeled using scanner data (e.g. Bucklin and Gupta 1992). However, the consumers' decision is considered to be one of choosing to buy one of the items in the assortment at today's prices and promotions, versus postponing the purchase decision to a future occasion when prices may be better, and relying meanwhile on available household inventory for consumption. Thus, the impact of assortment unavailability is not modeled.

Then,

1. For all customer segments s, assume that customers do not purchase if their most preferred d (exogenously specified) items are not stocked (see Smith and Agrawal 2000 for a similar operationalization). We call d the *depth of no purchase*. Clearly $d \in [1, N]$. An alternate interpretation of d is that it captures the (store) switching cost of a consumer; a large d implies high switching cost. Intuitively, a large d implies that a customer is willing to substitute less preferred items when more preferred items are not stocked rather than not purchase, regardless of the magnitude of disutility incurred. Under this operationalization, we set the no-purchase utility $V_{sj_0} = V_{sj_{d+1}}$ if $d < N$ and $V_{sj_0} = V_{sj_N} - \varepsilon$ (for some $\varepsilon > 0$) if $d = N$
2. Alternately, let T be an exogenously specified threshold level of disutility that signifies no purchase. Suppose there exists an item j_{k+1} for segment s, such that $V_{sj_1} - V_{sj_{k+1}} \geq T$. Then we infer that a customer in segment s will not purchase if items j_1 through j_k are not available in the assortment. Under this operationalization, we set the no-purchase utility $V_{sj_0} = V_{sj_{k+1}}$.

While either formulation is easily incorporated in our model, in this chapter, we use the former approach to model no-purchase. Later, we will analyze the sensitivity of the assortment solution to the depth of no purchase, d. To incorporate the depth of no purchase into problem P1, we modify constraint (10.6a) as follows. For each customer segment, s, define an order set consisting of d elements $\mathcal{N}_s^d = \{j_1, j_2, \ldots, j_d, j_0 | V_{sj_1} \geq V_{sj_2} \geq V_{sj_d} \geq V_{sj_0}\}$. We then rewrite (10.6a) as:

$$\sum_{k=0}^{d} V_{sj_k} x_{sj_k} \geq V_{sj_i} y_{j_i} \qquad \text{for } j_i \in \mathcal{N}_s^d \text{ and } \forall s \tag{10.6a'}$$

Furthermore, to ensure that a customer is assigned a product within their first d choices or no-purchase, we need to modify constraint (10.6c) to:

$$\sum_{j=0}^{d} x_{sj} \leq 1 \qquad \forall s \tag{10.6b1'}$$

$$\sum_{j=d+1}^{N} x_{sj} \leq 0 \qquad \forall s \tag{10.6b2'}$$

4.3 Reformulation

In this section we reformulate problem (P1), specifically constraint (10.6a′) which facilitates solution of (P1) as a linear program when integrality constraints on x_{sj} are relaxed. We observe that constraint set (10.6b)–(10.6e) is of the same form as that for

an uncapacitated plant/warehouse location problem (Cornuejols et al. 1977). We now reformulate constraint set (10.6a′) that results in a tighter formulation for (P1). Observe that (10.6a′) ensures that a customer segment is assigned its most preferred product amongst the ones stocked. Thus it merely depends on the rank order of products for any given consumer segment and not on the interval scaled utilities as measured by V_{sj}. We exploit this structure to replace (10.6a′) with.

$$1 - \sum_{k=i+1}^{d} x_{sj_k} \geq y_{j_i} \qquad \text{for } j_i \in \mathcal{N}_s^d \text{ and } \forall s \tag{10.6a''}$$

We also relax the constraints on x_{sj} in (10.6d) as follows:.

$$x_{sj} \leq 1 \tag{10.6d'}$$

Proposition 4.1. *Problem (P1) with (10.6a″) set of constraints is at least as tight a formulation as (P1) with (10.6a′) set of constraints. Furthermore the relaxation of integrality constraints to (10.6a′) still guarantees an integer solution for* x_{sj}.

A proof is provided in the appendix.

Thus the new constraint set (10.6a″) achieves the same results as (10.6a′), i.e., ensuring that of the items stocked a customer segment is assigned its most preferred item. Furthermore, this reformulation does not increase the number of constraints. Finally, the relaxation guarantees an integer solution. In the sequel we will use (P1) with (10.6a″) and (10.6d′).

4.4 Discussion of the Optimization Model and Some Special Cases

Readers familiar with the literature on plant location will see that problem (P1) has an embedded uncapacitated plant location model (when $w_c = 0$, and constraints (10.6a) are relaxed). This problem is extensively researched by Cornuejols et al. (1977) and they show that the problem is NP-hard. Hence problem (P1) is also NP-hard. Our computational study shows that similar to the uncapacitated plant location model (Erlenkotter 1978), the solution to problem (P1) is easily obtained for problem sizes (relatively small) of interest in this study. Large scale models comprising several products in a product line and a larger number of customer segments will call for development of heuristics.

We now consider a few special cases of Problem P1. First, we consider the situation when a retailer places zero weight on the disutility incurred by the consumers due to his assortment decision; we shall identify a retailer with $w_c = 0.0$ as a *myopic retailer* who maximizes just short-term profits.

To highlight the need to model "no purchase", consider the myopic retailer who solves P1 with $w_c = 0.0$ and with a depth of no purchase $d < N$. Recall that as d increases, consumers are more willing to substitute to the available items in the assortment and less willing to not purchase. We then observe that in a model without a no-purchase decision, a myopic retailer will stock only one product. Effectively, we solve problem P1 with $w_c = 0.0$ and $d = N$; that is, the retailer does not care about disutilities incurred by the consumers and all consumers purchase some product. This implies that the total demand is unaffected by the choice of items available. Then a retailer carries just one product $j^* = argmax_j\{(p_j - c_j)n_s - K_j\}$ which maximizes his profit.

We would like to be able to study the behavior of the assortment decision with respect to parameters like weight on disutility (w_c), depth of no-purchase (d), contribution margins ($p_j - c_j$), etc. In general, (P1) is a complex optimization problem and usually does not permit many comparative statics results. Analytically, we were unable to get any general sensitivity results with respect to p_j, w_c and d. The main difficulty appears to be the very general formulation of the heterogeneity of consumers. Any change in these parameters affects the substitution pattern through change in the interval scaled utilities and hence the demand patterns. The obvious case is when profit margins increase due to decrease in marginal costs. This increases the contribution margin and with fixed p_j, d and w_c, the retailer will find it optimal to increase his assortment sizes, since for $d < N$ it may help him satisfy more consumers and/or decrease disutility if $w_c > 0$.

5 Computational Study

5.1 Description of Household Scanner Panel Data

The data were collected by the AC Nielsen Company and are available for a 2 year period. A panel of households provided information on their purchasing in several categories. These data were supplemented with data on prices, in-store displays, and feature advertising collected from the supermarkets in the city. We include purchases of the eight largest brand-sizes of canned tuna made by 1,097 panelist households in our estimation sample. These eight items account for approximately 90 % of category volume. Brand names are disguised to meet confidentiality requirements of the data provider.

In Table 10.1 we provide descriptive statistics of the data. Besides shelf price, we include in-store displays and retailer feature advertising in the choice model. Table 10.1 indicates that there is considerable variation in shelf prices and promotional activity between brands, highlighting the need to control for the effects of these variables when measuring intrinsic brand preference or valuation.

Bayesian posterior estimates of the model parameters are obtained for each household using the approaches of Imai and van Dyck (2005) and McCullogh and Rossi (1994). Table 10.2 contains the mean value of the estimated posterior estimates. The coefficients of price, display, and feature, have the expected signs.

Table 10.1 Descriptive statistics of data

Item	Average price (cents/oz.)	Display (% occasions)	Feature (% occasions)
1	12.3	3.9	25.9
2	21.8	0	1.7
3	12.0	4.0	29.9
4	11.5	8.7	24.4
5	15.1	0	0
6	24.2	0	0
7	11.3	4.3	24.4
8	9.8	4.2	13.7

Table 10.2 Mean value of household parameter estimates of probit model demand

Mean brand specific constants	
Item 1	0.815
Item 2	2.350
Item 3	0.494
Item 4	1.030
Item 5	0.267
Item 6	2.885
Item 7	−0.273
Price ($/oz.)	−26.882
Display	0.597
Feature	0.163

We use the estimated $-metric intrinsic preferences for items V_{ij} to infer patterns of primary demand and likely substitution between items. We computed optimal assortments under two separate assumptions about consumers' willingness to substitute. First we assume that consumers are willing to make one substitution. That is, they will not purchase in the category if their first preference and second preference brands are not available (i.e., $d = 2$). Therefore, we focus on the top two brands for each consumer. Note that customers who do not find their most preferred brand but do find their second-most-preferred brand still incur a disutility, which our decision model incorporates. Next, we also solved for the optimal assortment under the assumption that consumers are willing to substitute twice (i.e., $d = 3$). In the subsequent discussion we describe the solution under the $d = 2$ assumption in detail and thereafter briefly talk about the $d = 3$ case.

Table 10.3 shows the cross classification of the first and second preference brands for the sample of 1,097 consumers.[11] Row total $N_{i.}$ indicates the number

[11] Note that only the rank ordering of preferences is used to construct Table 10.3 to illustrate the nature of substitution between items. The retailer optimization problem uses interval-scaled values of preferences.

Table 10.3 Cross classification of first and second preference brands (cell entries are in %)

First preference product	Second preference product								No. of consumers
	1	2	3	4	5	6	7	8	
1	0.00	0	71.7	3.3	1.7	0	10.0	13.3	60
2	0	0	0	0	0	0	0	0	0
3	69.7	1.4	0	1.9	0	0	17.1	9.9	211
4	3.4	0	6.8	0	1.1	0	78.4	10.2	88
5	12.5	6.3	3.1	0	0	0	28.1	50.0	32
6	0	0	0	0	0	0	0	0	0
7	2.3	0	14.7	55.4	2.5	0	0	25.1	354
8	14.5	0	21.3	2.8	13.6	0	47.7	0	352
Number of consumers	213	5	177	212	59	0	288	143	1,097

of consumers whose first preference brand is brand i. Similarly, column total $N_{.j}$ is the number of consumers whose second preference brand is brand j. Each cell entry in the table denotes the percentage of $N_{i.}$ consumers who have brand j as their second preference brand.

The row totals are indicative of primary demands for items. For example, it is clear that items 3, 7 and 8 are the first-preference products of a large number of consumers, while none of the consumers in our sample prefer items 2 and 6. Similarly, items 1, 4, and 5 have relatively weak primary demand. Column totals indicate whether items are acceptable as substitutes. Item 1, for example, is the brand of second choice for a large number of consumers (213) as compared with its primary demand (60). A similar preference pattern is evident for items 3 and 4. Item 8 has the opposite kind of preference pattern, with large number of consumers (352) preferring it in first place while only 143 prefer it in second place. Large cell entries indicate items that are more substitutable. For example, we see that 71.7 % of consumers who have item 1 as their first preference have item 3 as their second preference. Conversely, 69.7 % of those who prefer item 3 are willing to accept item 1. There is some evidence of asymmetries in patterns of substitution between brands. For instance, the entry in row 5 and column 8 is 50.0 % while that in row 8 and column 5 is only 13.6 %. These data further confirm the existence of substantial heterogeneity in patterns of substitution between consumers.

5.2 Solution Technique for Assortment Problem

We used LINDO, a commercial linear programming package, to solve the reformulated optimization model. The problems are generated from the preference, price, and cost data using a program written in C. This program allows the decision maker to vary the weight w_c (weight on consumer welfare and profit objectives) and d (depth of no purchase) to evaluate various solutions.

For our computational study we solved 80 instances of the problem. We varied the weight w_c from 0.01 to 0.99 with $d = 2$ (40 problems) and $d = 3$ (40 problems) for two different fixed costs. On average the problem took 32 s of cpu time, with times ranging from 20 to 48 s. Based on our computational times it seems appropriate to solve this problem to obtain the optimal solution using a commercial package. Specialized implementation and heuristics may be necessary for larger problems if the computational times become prohibitive.

5.3 Optimal Assortment

To solve the retailer optimization problem (P1), we need estimates of fixed costs (K_j), contribution margins ($p_j - c_j$), and of w_c, the weight placed by the retailer on customer disutility relative to current period profits. We did not have access to real

cost and contribution data for the market for which consumer data were available. For the empirical illustration, we assume values of these parameters as follows. Retail contribution margins are assumed to be 30 % of the average retail price of the item. Thus, items can be ordered in terms of margin based on the average prices shown in Table 10.1. We examine two different levels of fixed costs in our illustrations: $1 per re-stocking period and $5 per stocking period. These levels of fixed costs ensure that at least one item is unprofitable to carry based on its primary demand. We explore the impact of varying w_c (over the space 0 to 0.99 in small steps) on the optimal assortment, profits and customer disutility.[12]

Case 1: Fixed Cost is $1 per item per stocking period

In Table 10.4 we show changes in the optimal assortment of items, customer disutility, and optimal profits as the weight on disutility in the objective function (w_c) is increased from 0 to 0.99. Note that items 2 and 6 are never included in the optimal assortment, regardless of the value of w_c, because of the pattern of first and second preferences discussed previously. When $w_c = 0$, the problem reduces to the pure profit maximization problem of a myopic retailer. Thus, the retailer should carry only those products whose contribution margin exceeds the fixed cost. The demand for a product, given an assortment, is the sum of its primary demand, and spillover demand from items not carried. The solution to the pure profit maximization problem is to carry four items (item numbers 1, 3, 5, and 7). Table 10.1 shows that products 1, 3, and 5 are the highest margin products (after products 6 and 2). Although item 4 has higher margin than item 7, item 7 is included in the optimal assortment instead of item 4 because of its large primary demand (354 consumers) relative to item 4 (88 consumers). When a weight of 0.03 is placed on disutility we find that item 4 is also included in the assortment now. As noted previously, item 4 has low primary demand, but is acceptable as a

Table 10.4 Optimal assortment and resulting disutility and profits (fixed cost = $1)

Weight on disutility (w_c)	Disutility	Profit	# customers not served	Optimal assortment
0.000	85.83	34.50	19	1, 3, 5, 7
0.010	85.83	34.50	19	1, 3, 5, 7
0.030	69.62	34.17	0	1, 3, 4, 5, 7
0.040	20.38	32.64	13	3, 7, 8
0.050	4.45	31.93	7	3, 4, 7, 8
0.200	2.27	31.60	0	3, 4, 5, 7, 8
0.300–0.990	0.00	30.72	0	1, 3, 4, 5, 7, 8

[12] For the illustration here we assume that the total market consists of the 1,097 consumers in our sample.

substitute by a large number of customers. Nineteen customers who were previously not served at all now find an acceptable product to buy. Moreover, with this assortment profits are slightly lower, but disutility is significantly reduced. This suggests that profit as a function of assortment carried is quite flat near the maximum. The introduction of a second criterion (i.e., disutility) into the objective function helps us to select the assortment that delivers close to maximum profits while reducing disutility. If customer disutility influences future store traffic and hence long-run profits, the results presented help the decision maker balance short-run with long-run profits.

As w_c is increased further, we find that the number of items in the optimal assortment decreases and then increases. At $w_c = 0.040$ the optimal assortment shrinks from {1,3,4,5,7} to {3,7,8}. The inclusion of item 8 is probably explained by its large primary demand (352 customers), which implies that when it is omitted from the assortment, large disutility is incurred. Further, half of the customers who prefer item 5 find item 8 acceptable. At $w_c = 0.050$ the optimal assortment expands to include item 4 once again. At $w_c = 0.30$ the optimal assortment expands to include all six products, other than items 2 and 6.

Note that we observe two kinds of non-monotonicities in the optimal behavior with increases in w_c. One, the number of items in the optimal assortment expands and then shrinks. Two, certain items (such as 4 and 1) enter the optimal assortment, then get dropped, and then get re-included. Such non-monotonic behavior of the optimal assortment reinforces the need for a decision support model for retail assortment decisions.

Case 2: Fixed Cost is $5 per item per stocking period

In Table 10.5 we show the optimal assortment and associated profits and disutility. Note that in the pure profit maximization case, 155 customers are not served and disutility incurred is quite high. Placing a weight of 0.03 on disutility expands the optimal assortment to include product 8 in addition to items 3 and 7. As a consequence, profits drop. However, the number of customers served increases significantly and disutility drops sharply.

A distinguishing feature of the optimal assortment in Case 2, relative to Case 1, is that with increase in w_c the number of items in the optimal assortment always increases. Furthermore, once an item enters the optimal assortment it stays in the

Table 10.5 Optimal assortment and resulting disutility and profits (fixed cost = $5)

Weight on disutility (w_c)	Disutility	Profit	# customers not served	Optimal assortment
0.000	95.92	22.72	155	3, 7
0.010	95.92	22.72	155	3, 7
0.030	20.38	20.64	13	3, 7, 8
0.300	4.45	15.93	7	3, 4, 7, 8
0.700–0.990	0.00	6.72	0	1, 3, 4, 5, 7, 8

Table 10.6 A three-product example

Customer	Utility		
	Product 1	Product 2	Product 3
1	5	2	1
2	5	2	1
3	1	2	5
Fixed cost	2	2	2
Margin	1.4	1.4	1.4

assortment with increases in w_c. We conjecture that the high fixed cost may cause such monotonic behavior of the optimal assortment.

Results in the $d = 3$ case are entirely consistent with the results for the $d = 2$ case with some differences that are intuitive. For reasons of space we do not show detailed results. At each level of w_c, we find that optimal profits are at least as large in the $d = 3$ case since consumers are assumed to be more willing to substitute to less-preferred products. As a result, the spillover demand to any product from items not carried is no lower in this case than in the $d = 2$ case. Further, disutility is at least as large in the $d = 3$ case. When the fixed cost per item is \$1, the optimal assortment changes non-monotonically with increases in w_c. When the fixed cost is \$5, on the other hand, the optimal assortment changes monotonically.

To deduce further inferences, we ran the model for both cases of fixed costs considered previously ($K = 1$ and 5) and equal margins across all products, set equal to average margin of eight products using depths $d = 2$ and $d = 3$. The optimal solutions exhibited monotone changes to the optimal assortment for all w_c values. While this is true for our particular data set, we are able to construct a three-product, three-customer instance to provide a counter-example (see data in Table 10.6) for this monotone behavior.

In this counterexample, we find that when $w_c = 0$, the optimal profits are 2.2, the disutility is 9, and the optimal assortment has only product 2. As w_c grows to 0. 1379, the assortment consists of product 1 only, and for higher values of w_c the optimal assortment consists of products 1 and 3.

The results show that it is very hard to predict the structure of the optimal assortment, especially when we consider a data-driven problem setting.

6 Summary, Extensions, and Future Work

We propose a model for the optimal assortment and stocking decisions for retail category management. In particular, we address the question of rationalization of the retail assortment, i.e., determining the optimal subset of items to retain from the set of items currently carried. We assume, based on empirical evidence reported in the literature, that consumers are willing to partially substitute less preferred items if their preferred items are not available. We also assume that consumers are

heterogeneous in their intrinsic preferences for items and in their price sensitivities, an assumption strongly supported empirically.

We propose that the appropriate objective function for a far-sighted retailer should include not only short-term profits but also a penalty for the disutility incurred by consumers who do not find their preferred items in the available assortment. The rationale for including such a penalty is that dissatisfied consumers are less likely to return to the store in the future. We propose a measure for disutility that recognizes differences between consumers in their intensity of dissatisfaction.

The retailer problem is formulated as an integer programming problem. We show that the problem is large but can be solved efficiently to obtain an optimal solution. We demonstrate an empirical application of our proposed model using household scanner panel data for eight items in the canned tuna category. Our results indicate that the inclusion of the penalty for disutility in the retailer's objective function is informative in terms of choosing an assortment to carry. We find that customer disutility can be significantly reduced at the cost of a small reduction in short term profits.

An immediate extension of the current work is to develop heuristics to solve the optimization problem since problem sizes in categories with a large number of items may be very large and computational times to find optimal solutions might be prohibitive. Furthermore, we realize that there is uncertainty due to errors in the utility function parameter estimates, which our optimization model assumes to be fixed. The problem formulation can be modified to allow for uncertain parameter estimates and use a stochastic programming approach to solve the assortment problem.

The approach described in this chapter is an illustrative first-step that attempts to close some of the modeling gaps in the literature. As outlined in the introduction, the complete assortment planning problem needs to consider several other factors. Next we discuss briefly several directions to extend the proposed model in future research.

1. *Shelf Space Constraints:* Typically, retailers have shelf space constraints which limit the amount of stock that can be carried within a category. These constraints can be incorporated within the context of our problem (P1). A complexity that now arises is the occurrence of stock-outs. Since customers have heterogeneous preferences for items, the dynamics of their arrival process also needs to be accounted for.
2. *Incorporating Demand Uncertainty*: In the current model, we assumed that utilities of each consumer segment are deterministic. In fact, from the retailer's perspective utilities are stochastic. Including stochastic utilities results in a mixed-integer stochastic programming problem.
3. *The Pricing Problem*: The basic formulation outlined in this chapter can be extended to study the joint pricing and assortment decisions. However, maximization over prices makes (P1) a non-linear optimization problem which can be solved using procedures outlined in Adams and Sherali (1990), for example. Alternately, heuristic procedures could be explored.
4. *The Display Effect or the Effect of Facings on Sales*: The literature on shelf space management has been concerned with the relationship between shelf space allocations and sales due to the influence of product display on demand.

The number of facings allocated to an item also determines the quantity stocked of this item (usually an integer multiple of the number of facings). Thus, the problem of determining the optimal assortment and inventory is inter-related with the shelf-space allocation problem. Extending the model presented in this chapter to incorporate the display effect presents two challenges: one, the problem of measuring the effect of product display on demand, and two, the optimization problem changes considerably since we will now have to decide on number of facings which will be an integer variable.

5. *Joint Fixed Costs*: Product lines for a retailer typically consist of several SKU's being supplied by the same manufacturer or wholesaler. Therefore, multiple products in a category may require common resources (contact, vendor management, etc.). The Dobson and Kalish (1993) formulation assumes independent fixed costs, and therefore it can overstate the fixed costs associated with incremental introduction of products that share fixed costs with incumbent products. In case of shared fixed costs, a firm can take the savings available into account when introducing products that require common resources. One approach is to define product classes, similar to manufacturing classes used by Morgan et al. (2001). We hypothesize that inclusion of common fixed costs (relative to the assumption of independent fixed costs) will increase the number of products offered, profits, as well as consumer satisfaction.

Acknowledgements We are grateful to the A.C. Nielsen Company for generously providing the data used in this paper, to Edward Malthouse for his help with setting up the data, to Qiang Liu for help with data analysis, and to Pradeep Chintagunta, Maqbool Dada and Yehuda Bassok for valuable comments on an earlier version of the paper.

Appendix

Proof of Proposition 4.1. Without loss of generality, we will illustrate this for the general case rather than the special case of fixed depth of search d.

First consider the constraint (10. 6a″). The constraint for $j = 1$ will be

$$1 - (x_{s2} + x_{s3} + \ldots + x_{sK} + x_{s0}) \geq y_1$$

However, from (10.6c) we know that

$$x_{s1} + x_{s2} + \ldots + x_{jK} + x_{s0} \leq 1$$

Actually, given a "no purchase" option, the above is an equality; i.e.,

$$x_{s1} + x_{s2} + \ldots + x_{jK} + x_{s0} = 1$$

Using this we rewrite $1 - (x_{s2} + x_{s3} + \ldots + x_{sK} + x_{s0}) \geq y_1$ as simply $x_1 \geq y_1$. Similarly, we can write (10. 6a″) for $j = k$ as

$$x_{s1} + x_{s2} + \ldots + x_{sk} \geq y_k.$$

Using this, for any arbitrary customer segment s that prefers K products in the ordinal order (without loss of generality) the constraint sets (10. 6a′) and (10. 6a″) are

(10. 6a′)		(10. 6a″)	
$V_{s1}x_{s1} + V_{s2}x_{s2} + \cdots + V_{sk}x_{sk} \geq V_{s1}y_1$	$(1')$	$x_{s1} \geq y_1$	$(1'')$
$V_{s1}x_{s1} + V_{s2}x_{s2} + \cdots + V_{sk}x_{sk} \geq V_{s2}y_2$	$(2')$	$x_{s1} + x_{s2} \geq y_2$	$(2'')$
⋮		⋮	
$V_{s1}x_{s1} + V_{s2}x_{s2} + \cdots + V_{sk}x_{sk} \geq V_{sk-1}y_{k-1}$	$((k-1)')$	$x_{s1} + x_{s2} + \cdots + x_{sk-1} \geq y_{k-1}$	$((k-1)'')$
$V_{s1}x_{s1} + V_{s2}x_{s2} + \cdots + V_{sk}x_{sk} \geq V_{sk}y_k$	(k')	$x_{s1} + x_{s2} + \cdots + x_{sk} \geq y_k$	(k'')

Consider normalized constraint (1′) and (1″):

$$x_{s1} + \left(\frac{V_{s2}}{V_{s1}}\right)x_{s2} + \cdots \left(\frac{V_{sk}}{V_{s1}}\right)x_{sk} \qquad \text{and} \qquad x_{s1} \geq y_1.$$

Since $(1'')$ and $(1')$ are identical in x_{s1} dimension and $(1')$ has $k-1$ extra variables (degrees of freedom), constraint $(1'')$ is tighter than constraint $(1')$. Using similar arguments one can show that constraints $(2'')$ to $((k-1)'')$ will be tighter than $(2')$ to $((k-1)')$. Constraint (k'') may be identical to (k'). The argument can be repeated for other segments. Thus problem (P1) with (10. 6a″) is a tighter formulation than (P1) with (10. 6a′).

To see that relaxation of x_{sj} still leads to an integer solution, first consider $(1'')$. If $y_1 = 0$, then $x_{s1} = 0$ using (10.6c). If $y_1 = 1$, then $x_{s1} = 1$. Now consider $(2'')$. Suppose $y_1 = 0$. If $y_2 = 0$ then $x_{s2} = 0$; otherwise $(y_2 = 1)$, $x_{s2} = 1$. However, if $y_1 = 1$, then (10.6b) ensures that $x_{s2} = 0$. Following this argument, we can show that x_{sj} is integer. ■

References

Adams, P. W., & Sherali, H. D. (1990). Linearization strategies for a class of zero–one mixed integer programming problems. *Operations Research, 38*, 217–226.

Arrow, K., Karlin, S., & Scarf, H. (1958). *Studies in the mathematical theory of inventory and production*. Stanford: Stanford University Press.

Bassok, Y., Anupindi, R., & Akella, R. (1997). Single period multi–product inventory models with substitution. *Operations Research, 47*, 632–642.

Bawa, K., & Shoemaker, R. W. (1987). The effects of a direct mail coupon on brand choice behavior. *Journal of Marketing Research, 24*, 370–376.

Boatwright, P., & Nunes, J. C. (2001). Reducing assortment: An attribute based approach. *Journal of Marketing, 65*, 50–63.

Borin, N., Farris, P., & Freeland, J. (1994). A model for determining retail product category assortment and shelf space allocation. *Decision Sciences, 25*(3), 359–384.

Borle, S., Boatwright, P., Kadane, J. B., Nunes, J. C., & Shmueli, G. (2005). Effect of product assortment changes on customer retention. *Marketing Science, 24*(4), 612–622.

Broniarczyk, S. M., Hoyer, W. D., & McAlister, L. (1998). Consumers' perceptions of the assortment offered in a grocery category: The impact of item reduction. *Journal of Marketing Research, 35*, 166–176.

Bucklin, R. E., & Gupta, S. (1992). Brand choice, purchase incidence, and segmentation: An integrated modeling approach. *Journal of Marketing Research, 29*(9), 201–215.

Bultez, A., Gijsbrechts, E., Naert, P., & Vanden Abeele, P. (1989). Asymmetric cannibalism in retail assortments. *Journal of Retailing, 65*(2), 153–192.

Bultez, A., & Naert, P. (1988). S.H.A.R.P.: Shelf allocation for retailer's profit. *Marketing Science, 7*(3), 211–231.

Carpenter, G., & Lehmann, D. R. (1985). A model of marketing mix, brand switching and competition. *Journal of Marketing Research, 22*, 318–329.

Chintagunta, P., Jain, D. C., & Vilcassim, N. J. (1991). Investigating heterogeneity in brand preferences in logit models for panel data. *Journal of Marketing Research, 28*, 417–428.

Chong, J., Ho, T.-H., & Tang, C. (2001). A modeling framework for category assortment planning. *Journal of Manufacturing and Service Operations Management, 3*(3), 191–210.

Cornuejols, G., Fisher, M., & Nemhauser, G. (1977). Location of bank accounts to optimize float: An analytic study of exact and approximate algorithms. *Management Science, 23*(8), 789–810.

Corstjens, M., & Doyle, P. (1981). A model for optimizing retail space allocations. *Management Science, 27*(7), 822–833.

Coughlan, A. T., Anderson, E., Stern, L. W., & El-Ansary, A. I. (2006). *Marketing channels*. Englewood Cliffs, NJ: Prentice Hall.

Dobson, G., & Kalish, S. (1988). Positioning and pricing a product line. *Marketing Science, 7*(2), 107–125.

Dobson, G., & Kalish, S. (1993). Heuristics for pricing and positioning a product–line using conjoint and cost data. *Management Science, 39*, 160–175.

Emmelhainz, M. A., Stock, J. R., & Emmelhainz, L. W. (1991). Consumer responses to retail stockouts. *Journal of Retailing, 67*(2), 139–147.

Erlenkotter, D. (1978). A dual-based procedure for uncapacitated facility location. *Operations Research, 26*(6), 992–1009.

Food Marketing Institute. (1993). *Variety or duplication: A process to know where you stand*. Washington, D.C.: The Research Department, Food Marketing Institute.

French, S., & Ruiz-Diaz, F. (1983). A Survey of multi–objective combinatorial scheduling. In S. French, et al. (Eds.), *Multi–objective decision making*. New York: Academic.

Gaur, V., & Honhon, D. (2006). Assortment planning and inventory decisions under a locational choice model. *Management Science, 52*(10), 1528–1543.

Gruen, T., Cortsen, D. S., & Bharadwaj, S. (2002). *Retail out-of-stocks: A worldwide examination of extent, causes and consumer responses*. Grocery Manufacturers of America.

Harris, B., & McPartland, M. (1993). Category management defined: What it is and why it works. *Progressive Grocer, 72*(9), 5–8.

Imai, K., & van Dyck, D. A. (2005). A Bayesian analysis of the multinomial probit model using the marginal data augmentation. *Journal of Econometrics, 124*(2), 311–334.

Kalish, S., & Nelson, P. (1991). A comparision of ranking, rating, and reservation price measurement in conjoint analysis. *Marketing Letters, 2*(4), 327–335.

Kamakura, W. A., & Russell, G. (1989). A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research, 26*, 379–390.

Kok, A., & Fisher, M. L. (2004). Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research, 55*(6), 1001–1021.

Krishna, A. (1992). The normative impact of consumer price expectations for multiple brands on consumer purchase behavior. *Marketing Science, 11*(3), 266–286.

Lee, H. L., & Nahmias, S. (1994). Single product single location models. In S. Graves, A. R. Kan, & P. Zipkin (Eds.), *Logistics of production and inventory. Handbook in operations research and management science*. Amsterdam: North–Holland.

Little, J. D., & Shapiro, J. (1980). A theory for pricing nonfeatured products in supermarkets. *Journal of Business, 53*(3), S199–S209.

Mahajan, S., & van Ryzin, G. (2001). Stocking retail assortments under dynamic consumer substitution. *Operations Research, 49*, 334–351.

McBride, R., & Zufryden, F. S. (1988). An integer programming approach to optimal product line selection. *Marketing Science, 7*, 126–140.

McCullogh, R., & Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics, 64*, 207–240.

Miller, C., Smith, S. A., McIntyre, S. H., & Achabal, D. D. (2006). *Optimizing retail assortments for infrequently purchased products*. Working Paper, Retail Management Institute, Santa Clara University.

Morgan, L. O., Daniels, R. L., & Kouvelis, P. (2001). Marketing/manufacturing trade-offs in product line management. *IIE Transactions, 33*, 949–962.

Nielsen Marketing Research (1992). *Category management: Positioning your organization to win*. Lincolnwood, IL: NTC Business Books.

Pentico, D. (1974). The assortment problem with probabilistic demands. *Management Science, 21*, 286–290.

Pentico, D. (1988). A discrete two-dimensional assortment problem. *Operations Research, 36*(2), 324–332.

Schiller, Z., Burns, G., & Miller K. L., (1996). Marketing: Making it simple. *Business Week*, 9 September.

Sloot, L., Fok, D., & Verhoff, P. C. (2006). The short- and long-term impact of an assortment reduction on category sales. *Journal of Marketing Research, XLIII*, 536–548.

Smith, S., & Agrawal, N. (2000). Management of multi–item retail inventory systems with demand substitution. *Operations Research, 48*, 50–64.

Swait, J., & Louviere, J. (1993). The role of the scale factor in the estimation and comparision of multinomial logit models. *Journal of Marketing Research, 30*, 305–314.

Urban, G. L., Johnson, P. L., & Hauser, J. R. (1984). Testing competitive market structures. *Marketing Science, 3*(2), 83–112.

van Dijk, A., van Heerde, H. J., Leeflang, P. S. H., & Wittink, D. R. (2004). Similarity based spatial methods to estimate shelf-space elasticities. *Quantitative Marketing and Economics, 2*, 257–277.

van Ryzin, G., & Mahajan, S. (1999). On the relationship between inventory costs and variety benefits in retail assortments. *Management Science,45*, 1496–1509.

Chapter 11
Optimizing Retail Assortments for Diverse Customer Preferences

Stephen A. Smith

1 Introduction

Assortment selection is one of the most important and difficult decisions that retailers face. Assortments are typically chosen subjectively, often before any sales have been observed for some candidate products. Compared to pricing or advertising decisions, assortment decisions are more difficult to adjust later on. For multi-featured items such as consumer electronics and durable goods, the large number of product options, together with limited display space and financial constraints all contribute to the complexity of this decision. Consumer preferences for the various product attributes may also be heterogeneous, which requires assessing tradeoffs between the products that appeal to diverse customer segments. Because of these complexities, intuitively chosen retail assortments are likely to be suboptimal.

This paper develops an operational methodology for selecting optimal retail assortments based on an underlying multinomial logit (MNL) choice model for each customer's selection of product and retailer. A formulation is developed for optimizing the retailer's expected profit across customers with heterogeneous preferences. The formulation can also include a variety of additional merchandising constraints, such as display space, price point coverage or brand offerings.

Choice models have been successfully applied in consumer package goods to predict customers' response to assortment changes, based on observing repeat purchase behavior. The increased use of the Internet as a shopping guide for more complex, less frequently purchased products provides an opportunity to

S.A. Smith (✉)
Department of Operations Management and Information Systems,
Leavey School of Business, Santa Clara University, 500 El Camino Real,
Santa Clara, CA 95053, USA
e-mail: ssmith@scu.edu

N. Agrawal, S.A. Smith (eds.), *Retail Supply Chain Management*,
International Series in Operations Research & Management Science 223,
DOI 10.1007/978-1-4899-7562-1_11

obtain detailed preference information for broader classes of merchandise. A commercial data base of consumer preferences for attributes and features of DVD players, which was obtained through interactive Internet sessions, is used to illustrate the methodology. Consumer surveys or past buying behavior of individuals might also be used as alternative sources for the preference information needed for this assortment optimization methodology.

The methods in this paper provide a basis for several strategic retailer decisions including: (1) determining the optimal set of SKUs to offer and their estimated selling proportions; (2) how the retailer's relative market strength affects the contents of the optimal assortment; (3) how changing the contents of the assortment affects the probability that customers choose a given retailer and (4) how the customers' preference structure affects the optimal assortment and the corresponding expected profits. In analyzing our sample data set, it was found that accounting for preference heterogeneity and customers' use of consideration sets both had significant impacts on the retailer's expected profits.

1.1 Literature Review

Kok et al. (2015) provide a comprehensive survey of recent papers in retail assortment planning, and thus this paper's literature review will focus on a few papers that are particularly relevant for the optimization model developed here. Several recent papers have developed models for assortment optimization based on a newsvendor type model for inventory cost. van Ryzin and Mahajan (1999), Cachon and Gurhan Kok (2007) and Cachon et al. (2005) use a multinomial logit (MNL) model in which customers have homogeneous expected utilities. In Mahajan and van Ryzin (2001), customers are heterogeneous with regard to utility and their paper explicitly models the substituted demand that results from random stockouts of the retailer's inventory, but optimizing the assortment requires solution heuristics that are based on the set of possible inventory trajectories over the season. Guar and Honhon (2006) used a Lancaster type of model of substitution for products distributed along a single attribute dimension, and analyzed the impacts of static and dynamic substitution under this preference structure. Honhon et al. (2010) consider assortment optimization with stockout based substitution for more general deterministic preference structures. This leads to a dynamic programming formulation, for which they develop solution heuristics. Rusmevichientong and Topaloglu (2012) consider a generalized version of the MNL in which the model parameters are random, and show that the optimal assortments satisfy the nested set properties that hold for the MNL choice model with fixed parameters. Sauré and Zeevi (2013) develop a retail assortment optimization model that incorporates learning through experimentation with alternative assortments, and study the tradeoff between gaining information and maximizing current revenue. Smith and Agrawal (2000) used a probability of substitution matrix across products to optimize assortments in combination with an approximate

newsvendor inventory model. Miller et al. (2010) provide a method for optimizing assortments for infrequently purchased products and compare the results for several simple assortment selection heuristics. Kok and Fisher (2007) develop a heuristic for optimizing the allocation of shelf facings and inventory levels for a supermarket based on a particular substitution structure that also considers stockouts. Chong et al. (2001) developed a more general hierarchical market model for retail assortment planning for repeat purchase items, but due to the complexity of the resulting objective function, used a local improvement heuristic for optimization.

Only two of the above papers address the issue of retailer choice. Cachon et al. (2005) investigates how three different consumer models for the value of additional search at alternative retailers can affect the optimal assortment. Cachon and Gurhan Kok (2007) develop a more general category management model based on the retailer choice probabilities obtained from the nested logit model, but require mean utilities that are homogeneous across customers.

Product line optimization models have used mathematical programming formulations to solve a related problem. In this setting, a manufacturer decides which set of products to produce, where each potential product is viewed as a collection of adjustable product attributes. Chen and Hausman (2000) considered product line selection based on the MNL choice model, with homogenous customer preferences. Green and Krieger (1985), McBride and Zufryden (1988), Dobson and Kalish (1988, 1993) and Kohli and Sukumar (1990) consider heterogeneous customer utilities, but assume deterministic product choices. Green and Krieger treat discrete price options as product attributes, as is done in this paper, while Dobson and Kalish treat product prices as separate decision variables. With the exception of Chen and Hausman, these mathematical programming formulations are computationally difficult to solve, in part because they assume strict utility maximization by customers. Some product line selection papers developed solution heuristics (Kohli and Sukumar 1990; Dobson and Kalish 1993) or suggested clustering of customer preferences to reduce the problem size (Green and Krieger 1985) so that iterative search methods can be applied. These product line optimization methods do not model retailer choice, nor do they include inventory management costs.

1.2 Summary of Results

This paper provides an operational assortment optimization model that includes general heterogeneous consumer preferences as well as the customer's choice of retailer within the MNL framework. It is shown that the input parameters required for modeling product choice and retailer choice can be estimated separately, which facilitates their use in an operational model for assortment optimization. Assuming homogeneous mean utilities, van Ryzin and Mahajan (1999) showed that the optimal assortments form nested sets as the assortment size increases. Rusmevichientong and Topaloglu (2012) extend this result for the case of unknown MNL parameters. For heterogeneous customer utilities and competing retailers, this

chapter shows that this nested set property no longer holds, but that nested optimal assortment sets do occur for two limiting cases: (1) a monopoly retailer and (2) perfect competition among retailers. An optimization formulation is developed, which can include linear retailer constraints on the contents of the assortment, such as brand coverage and display space limitations. Finally, a commercial data base of preferences for DVD players is analyzed to illustrate the sensitivity of the expected profit and optimal assortment to the customer preference structure. The results for this data set illustrate the importance of including preference heterogeneity and customers' use of considerations sets in assortment optimization, as well as the sensitivity of the retailer's profit to assortment size.

2 Model Description

This paper focuses on the assortment decision for a particular retailer r, whose objective is to maximize the expected profit over a fixed time period, e.g., the Fall season. It is assumed that other retailers do not react competitively to this retailer's decisions. The retailer's assortment is defined by a binary vector $y = y_1, y_2, \ldots, y_n$, where $y_j = 1$ if the retailer's assortment includes product j and 0 otherwise. Then let

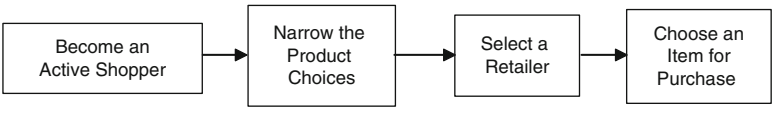
$$D_j(y) = \text{the random demand for product } j,$$

which depends on y as well as other factors that affect demand. We now develop a choice model that determines the probability distribution for $D_j(y)$.

2.1 Modeling the Consumer's Purchase Decision

First, suppose that customers are classified according to n distinct customer types indexed by $i = 1, \ldots, n$. It is assumed that customers of the same type assign the same expected values to various choice alternatives, but their actual purchase decisions also reflect individual random variations.

Actual purchases are the result of a sequential process that can be diagramed as follows:



The choice decisions in each of these steps can be described in terms of the iPACE model for retail shopping decisions that has been developed in the marketing literature, where iPACE stands for information, Price, Assortment, Convenience and Entertainment, (see e.g., Hanson and Kalyanam 2006, Chap. 13) By

becoming an active shopper, the customer is sufficiently interested in the product category to gather information. Using a variety of sources, which may include both Internet research and store visits, customers assess their utilities for the available products and the relative values of purchasing from the alternative retailers. This process allows a customer to narrow the set of choices to a "consideration set" of products. The customer selects a retailer based on the retailer's assortment, as well as the assessed convenience and entertainment values of shopping at that retailer. Finally, the customer makes a product selection from the choice set, which is defined as the intersection of the consideration set and the chosen retailer's assortment. Although this description is sequential, these decisions do not necessarily need to be made in any specific order. For example, the customer might choose the most preferred product first, and then select the retailer from which to purchase. The key assumption is that the combination of the utility of the retailer and utility of the chosen product jointly determine the customer's decision. This chapter assumes that these decisions are made normatively by customers, based on maximizing expected utility.

From the perspective of a particular retailer r, the customer may also choose the "no purchase" option for two reasons: (1) no product in the consideration set has positive net value, i.e., the choice set is empty or (2) the combined value of shopping and purchasing from this particular retailer's assortment either does not exceed the product's price, or is less than the combined value obtained from another retailer.

The mathematical models for each of these steps can be summarized as follows. The assortment decision is made for a fixed period of time, e.g., one season, and the time dependent parameters correspond to the length of this season. A random number N_i of customers of type i will become "active shoppers," i.e., they will gather information and make a purchase decision this season for this product category. We assume that N_i is a Poisson random variable with rate parameter λ_i. For the N_i shoppers, define

$$q_{ij}(y) = \mathrm{P}\big\{\text{customer type } i \text{ chooses product } j \text{ from this retailer} \big| \text{assortment } y\big\}.$$

This implies that $D_j(y)$, the random demand for product j defined previously, has a Poisson distribution with mean

$$\mu_j(y) = \sum_i \lambda_i q_{ij}(y). \tag{11.1}$$

The remaining customer decisions, which determine $q_{ij}(y)$, are based on the following utility model. The underlying choice model is a multinomial logit (MNL) in which customer i's combined utility for product j and retailer r is a random variable of the form

$$U_{ij}^{r} = U_{ij} + V_{ir} + \varepsilon_{ijr}, \tag{11.2}$$

where ε_{ijr} = Gumbel distributed error terms with mean 0 and scale parameter $\xi_{i,}$,
U_{ij} = the expected (net) utility obtained from purchasing product j,
V_{ir} = the additional utility obtained by purchasing from retailer r.

For this paper's analysis, the product price is included in U_{ij} as a fixed attribute, rather than a decision variable. For many retailers, this is justified based on operational practice. At the individual product level, tactical pricing decisions such as temporary markdowns are typically made by the retailer later on during the selling season, as part of promotional and advertising activities. Strategic pricing decisions, such as how to price relative to competitors, are typically made less frequently and at a higher level than just one product category. For assortment planning purposes, the product price is therefore the estimated average price for the season. A combined model that simultaneously optimizes product prices and the retail assortment is conceptually superior to separate decision models, but it cannot feasibly include all the other aspects of customers' purchasing decisions that are analyzed here.

Additive MNL models of the form (11.2) are frequently used for two dimensional choice decisions. (See, e.g., Ben Akiva and Lerman 1985 for further discussion.) In the context of this application, the error terms ε_{ijr} can capture both the customer's imprecise knowledge of his or her own utilities, as well as the retailer's imperfect knowledge of customers' utilities. It is common practice to rescale the utilities for each customer i so that the scale parameters $\xi_i = 1$ for all i. This is possible because dividing all utilities with subscript i by the same scalar ξ_i does not change which utility is the maximum for customer i. That is, probability statements about the maximum utility for customer i are not affected by this rescaling.

2.1.1 Narrowing the Product Choices

Narrowing the product choices is a "prescreening" step that does not change the fundamental structure of the underlying logit model. When there are many product alternatives to consider, marketing researchers have found that customers typically use some criteria to narrow their choices to a "consideration set" of products, which are then investigated in more detail. (See, e.g., Roberts and Lattin 1991; Andrews and Srinivasan 1995; Siddarth et al. 1995). In a normative framework, customer i would form a consideration set by eliminating all products with expected utility less than some threshold u_i, where the threshold is based on his or her cost of considering additional alternatives. Thus we define

u_i = customer i's minimum acceptable expected utility for considering a product,

$X_{ij} = 1$ if $U_{ij} \geq u_i$ and 0 otherwise, for all i, j.

Consideration sets can have a significant impact on the assortment optimization, as the numerical analysis in Sect. 3 illustrates.

2.1.2 Determining $q_{ij}(y)$

The definition of conditional probability implies that

$$\begin{aligned} q_{ij}(y) &= \mathrm{P}\{\text{customer } i \text{ purchases product } j \text{ from retailer } r \,|\, y\} \\ &= \mathrm{P}\{\text{customer } i \text{ purchases product } j \,|\, \text{purchases from retailer } r,\ y\} \\ &\quad {}^{*}\mathrm{P}\{\text{customer } i \text{ purchases from retailer } r \,|\, y\} \end{aligned} \tag{11.3}$$

This equation does not necessarily imply that the customer chooses the retailer first, but this decomposition allows a separable estimation of the required model parameters, as will be discussed later.

Given that customer i selects retailer r's assortment for a purchase, his or her choice set is defined as the intersection of the consideration set and retailer r's assortment, i.e.,

$$S_{ri} = \left\{ j \,|\, y_j X_{ij} = 1 \right\}, \quad \text{for all } i.$$

Given any choice set S_{ri}, the probability of selecting item $j \,\varepsilon\, S_{ri}$ is the standard MNL probability, which in this case is

$$P\{\text{customer } i \text{ purchases product } j \,|\, \text{chooses retailer } r, y\} = \frac{e^{U_{ij}}}{\sum_{k \in S_{ri}} e^{U_{ik}}}. \tag{11.4}$$

Ben Akiva and Lerman (1985, p. 282) show that the maximum utility that customer i obtains from the choice set S_{ri} has a Gumbel distribution, with mean

$$V_{ir}^{*} = \ln\left(\sum_{j \in S_{ri}} e^{U_{ij}} \right),$$

and the same scale parameter as the individual utilities. Thus, the total utility of purchasing from retailer r's assortment is Gumbel distributed with mean $v_{ir} = V_{ir} + V_{ir}{}^{*}$. The analogous result holds for all other retailers' assortments, which we index by ρ. Therefore, the maximum utility that customer i could obtain from shopping at other retailers also has a Gumbel distribution with mean

$$v_{io} = \ln\left(\sum_{\rho \neq r} e^{V_{i\rho} + V_{i\rho}^{*}} \right),$$

and the same scale parameter $\xi_i = 1$ as the individual utilities. This allows the retailer choice probability to be written as a binary logit probability

$$f_i = P\{\text{customer } i \text{ selects retailer } r | y\} = \frac{e^{v_{ir}}}{e^{v_{ir}} + e^{v_{io}}} = \frac{\sum_{j \in S_{ri}} e^{U_{ij}}}{e^{a_{ir}} + \sum_{j \in S_{ri}} e^{U_{ij}}} \text{ with } a_{ir} = v_{io} - V_{ir}. \tag{11.5}$$

The second fraction results if we multiply top and bottom by $\exp\{-V_{ir}\}$. From this point onward, we focus on the particular retailer r and simply write a_i for a_{ir}.

Combining the two probabilities in (11.3) using the assortment y for retailer r and the X_{ij} for customer i to define the choice set S_{ri}, we obtain the formula

$$q_{ij}(y) = \frac{y_j X_{ij} e^{U_{ij}}}{e^{a_i} + \sum_k y_k X_{ik} e^{U_{ik}}}, \tag{11.6}$$

after cancelling the term $\sum_{j \in S_{ri}} e^{U_{ij}}$. A key result in (11.6) is that a_i is a constant that is independent of retailer r's assortment decision y.

The size of a_i indicates the relative strength of retailer r's competitors for customer type i. The value of a_i can be obtained in various ways. One method is to assume that customer i knows the contents of all the retailers' assortments and chooses the best retailer by maximizing the total utility as described above. Alternatively, the customer might simply decide whether to continue shopping at other retailers based on an estimated value a_i, which corresponds to the estimated maximum utility improvement obtained from other retailers' products, plus the improvement in value obtained by buying from an alternative retailer versus buying from retailer r. For assortment optimization using (11.6), retailer r does not need to know which behavioral model applies to customer i, since a_i is simply a parameter to be estimated, as discussed below.

Kahn and Lehmann (1991) and others have suggested adding terms to V_{ir} to capture the additional customer value associated with properties of the assortment that increase its "breadth," such as the total number of products or the number of brands offered. The structure of the optimization model in this chapter does not allow these additional variables to be included in the retailer's objective function. But features such as the total number of products or the number of brands in the assortment can be included as constraints for the assortment optimization model, with their corresponding values being added as constant terms to V_{ir}. This allows a sensitivity analysis to be done with respect to these assortment parameters.

2.1.3 Estimation and Empirical Testing

An estimate for a_i can be obtained using (11.5) from the observed fraction f_i of customers of type i who choose retailer r for *any particular given* assortment. Assuming that f_i can be obtained approximately from market research data for the current assortment, we can solve for the corresponding a_i as follows

$$a_i = \left(\frac{1}{f_i} - 1\right) \sum_{j \in S_{ri}} e^{U_{ij}}.$$

This formula requires utility estimates for each product, which can be obtained from (11.4) as discussed previously. Thus, (11.4) and (11.5) allow the $\{U_{ij}\}$ and $\{a_i\}$ to be estimated separately, and they could in fact be obtained from different assortments.

Purchasing behavior for consumer package goods based on multi-stage logit models has been studied empirically for a variety of model forms. For example, Cintagunta (1993) provides a summary of articles that include empirical studies of three stages of consumer purchase decision making: (a) whether or not to purchase from this retailer (b) item choice from a retailer and (c) purchase quantity. See also Roberts and Lattin (1997) for a literature review. In forecasting demand for consumer package goods, "purchase incidence," which is defined as the probability that the customer makes a shopping trip to a given retailer that results in a purchase from the category, plays a role that is similar to retailer choice in this paper. [See, e.g., Bucklin and Lattin 1991 for a discussion of using the binary logit model for purchase incidence.]

2.1.4 Elasticity Comparisons

Formula (11.6) shows that adding another product to the assortment *increases* the probability that customer i purchases from this retailer, but *decreases* the probability that each of the original products in the assortment is selected. The magnitudes of these effects depend on a_i, as shown below.

Let $Q_i(y) = \mathrm{P}\{\text{customer } i \text{ purchases from this retailer}\}$, where

$$Q_i(y) = \sum_j q_{ij}(y) = \frac{P_i(y)}{e^{a_i} + P_i(y)}, \quad \text{with } P_i(y) = \sum_j y_j X_{ij} e^{U_{ij}}.$$

Interpreting partial derivatives as changes in y_k from 0 to 1, we can define the two elasticities

$$\frac{1}{Q_i(y)} \frac{\partial Q_i(y)}{\partial y_k} = \frac{e^{a_i} X_{ik} e^{U_{ik}}}{P_i(y)[e^{a_i} + P_i(y)]} \quad \text{and} \quad \frac{1}{q_{ij}(y)} \frac{\partial q_{ij}(y)}{\partial y_k} = -\frac{X_{ij} e^{U_{ij}}}{[e^{a_i} + P_i(y)]^2}.$$

The first and second elasticities show, respectively, that:

1. The percentage increase in total sales to customer i from adding product k is *greater* when a_i is *larger*.
2. The percentage of cannibalization of product j's sales due to adding product k is *smaller* when as a_i is *larger*.

Taken together, these results imply that including additional products in the assortment is more advantageous to the retailer when the retailer's competition is stronger.

2.2 Retailer's Assortment Optimization

The profit function $\Pi_j(D_j(y))$ for each product j is based on a newsvendor type model. In general, a fixed cost

$$F_j = \text{the fixed cost of stocking product } j$$

should also be included. The expected profit $\Pi(y)$ for the planning period as a function of y can therefore be written as the sum of the expected profits for the various products

$$\Pi(y) = \sum_j \left\{ E\left[\Pi_j\left(D_j(y)\right)\right] - F_j y_j \right\}.$$

[It should be noted that even though the random variables $\Pi_j(D_j(y))$ are not independent, their expectations are still additive.] A more general optimization problem can be defined if there are nonlinear cost interactions between the products, but that formulation will not be developed in this chapter.

The newsvendor expected profit for product j for a fixed time period as a function of the assortment y can be written as

$$E\left[\Pi_j\left(D_j(y)\right)\right] = \max_{S_j} \left\{ m_j \mu_j(y) - c_{uj} E\left[D_j(y) - s_j\right]^+ - c_{oj} E\left[s_j - D_j(y)\right]^+ \right\} \quad (11.7)$$

where $E[x]^+$ denotes the expected value of $\max\{0, x\}$ and

$s_j =$ the base stock level for product j for the time period
$m_j =$ unit profit margin for product j
$\mu_j(y) =$ expected demand during the time period $= E[D_j(y)]$
$c_{uj} =$ "understock" cost per unit $c_{oj} =$ "overstock" cost per unit

The financial input quantities can be calculated in the usual way, *i.e.*,

$m_j =$ selling price − unit cost,
$c_{uj} =$ shortage loss − unit cost
$c_{oj} =$ unit cost − salvage value.

From (11.6), we see that $y_j = 0$ implies $D_j(y) = 0$ with probability 1, which implies that the expected profit is 0. That is, there is no specific shortage cost c_{uj} that results from not including a given item in y, but there is a net loss of expected utility for the retailer's assortment, which increases the likelihood that the customer will choose another retailer. This is because when a customer's most preferred item is missing, the customer either substitutes another item from this retailer's assortment or chooses another retailer. The demand that results from substitutions for items not in the retailer's assortment is captured in $\mu_j(y)$. Substitutions from stockouts are ignored, as discussed below. From the standpoint of this retailer r, the probability of choosing another retailer is lumped together with the "no purchase" option.

Using the newsvendor critical ratio formula, the optimal base stock level s_j^* satisfies

$$s_j^* = \arg\min_S \left\{ s \middle| P\{D_j(y) \le s\} \ge \alpha_j = \frac{c_{uj}}{c_{uj} + c_{oj}} \right\}.$$

The overstock cost c_{oj} above can have a variety of interpretations. For continuing products that will be offered in subsequent seasons, it is the unit holding cost for the season, while for "seasonal" products, it is the unit cost minus the expected salvage value per unit for any excess inventory at the end of the season.

There are various fixed costs F_j that can be associated with stocking items in a product category. For larger items such as furniture, it is common to display one unit in the store and hold additional inventory elsewhere, for example. In this case, F_j would include the required floor space for display. For smaller items, there may be a shelf facing with one item viewable, and the remaining items stored behind it. In both these cases, F_j would include the fixed cost of the required display space in the store when the item is in the assortment.

2.2.1 Incremental Demand Arising from Substitution

Kok et al. (2015) define two kinds of substitution-based demand: (1) assortment based substitution in which a customer switches to another product when a more preferred product is not carried in the assortment and (2) stockout-based substitution in which the customer substitutes another product if a more preferred alternative is in the assortment, but it is out of stock. This chapter captures assortment-based substitution through the MNL choice model discussed previously, but it ignores stockout-based substitution. Some recent papers have modeled stockout-based substitutions, but this generally leads to complex optimizations, and thus solution heuristics are required. Mahajan and van Ryzin (2001) and Guar and Honhon (2006) and Honhon et al. (2010) assume that customers maximize utility over the items that are currently available, i.e., they treat the retailer's assortment as dynamic. These approaches are quite general, but require heuristic solutions for

most customer preference structures. The other assortment optimization models discussed previously in the literature review have either not treated this stockout-based substitution or have bounded its effects.

This chapter assumes that the customer chooses the retailer based on the complete assortment y, and that product demands which encounter stockouts of products in the retailer's assortment become lost sales. Thus the demand arising from stockout-based substitutions is ignored. Smith and Agrawal (2000) argue using bounds, that the absolute percentage error in expected demand that results from ignoring demand from stockout based substitutions is bounded by $(1-\alpha)(1-L)$, where α is the target service level and L is the probability that the customer is unwilling to substitute. For retailers that set high service level targets for most products during the normal selling season, this bound implies that stockout substitutions will rarely occur. The stockout based demand needs to be counted only if the customer is willing to switch to another product from the same retailer. Given that alternative retailers exist for many items, customers who choose another retailer instead of substituting a different product will be correctly captured by the lost sales assumption. Also, when the service level is defined as the probability of no stockout using the normal distribution, the fraction of demand served is typically larger than the service level. For example, for the normal distribution with $P\{\text{no stockout}\} = \alpha = 0.9$, approximately 96 % of demand will be served, and with $\alpha = 0.95$ approximately 98 % of demand will be served before a stockout occurs. The error corresponding to the unserved demand is further reduced by eliminating those customers who do not substitute another product from this retailer. Thus, for retail products that have high service levels such as 0.9 or 0.95, it seems reasonable to ignore substitution demand arising from stockouts.

2.2.2 Two Variants of the Objective Function

Products that may have purchase quantities larger than one can be handled in a variety of ways. One method is to use a compound Poisson distribution for demand, in which customers arrive according to a Poisson process and then select their purchase quantities randomly. For example, Poisson arrivals with a purchase quantity selected from a logarithmic distribution result in a negative binomial distribution for total demand during any fixed period. Smith and Agrawal (2000) used the negative binomial distribution and found that a linear approximation to the newsvendor objective function worked well in that case. Other papers on assortment optimization (e.g., van Ryzin and Mahajan 1999; Mahajan and van Ryzin 2001; Guar and Honhon 2006) have used a normal approximation for demand to obtain a newsvendor expected profit function.

When there are time based holding costs, it may be advantageous for retailers to restock more frequently than once per season. This feature can be added to the newsvendor model (11.7), provided that the assortment does not change in midseason. If there is an additional cost $h =$ unit holding cost for one restocking

period, a cost term of the form $0.5h\left[s+\left|s-D_j(y)\right|^{+}\right]$ is subtracted from the objective function. The critical ratio stock level formula still holds, where c_{oj} is replaced by $c_{oj}+h$ and c_{uj} is replaced by $c_{uj}-0.5\,h$. The costs c_{oj} and c_{uj} may also be allowed to vary by time period.

2.2.3 A Linear Approximation for the Objective Function

It can be verified by numerical calculation that for common ratios of profit margin to overstock and understock costs, the newsvendor expected profit function (11.7) is approximately linear in the expected demand $\mu_j(y)$ for the Poisson distribution. That is, when the various costs are held fixed and expected demand increases, the target service level remains constant and the safety stock increases in such a way that the sum of the terms in (11.7) increases approximately linearly as a function of the mean.

For the Poisson demand distribution, this approximation is illustrated for a range of parameter ratios in Fig. 11.1. To simplify the graph, all $F_j=0$ and all profits have been divided by c_u. That is, when all cost parameters are expressed as multiples of c_u, the graphs can be expressed as (expected profit)/c_u, which implies that the only required variables are the service level α and the mean demand. Using linear regression, the R^2 values for all the linear fits to the points in this figure are at least 0.998.

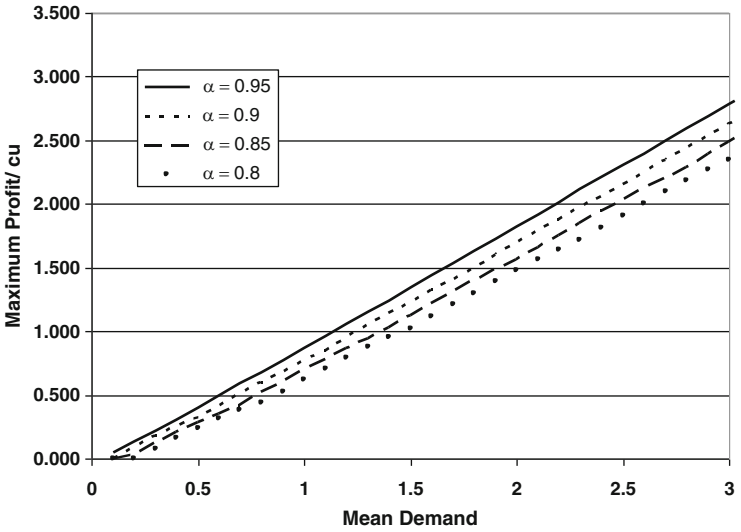


Fig. 11.1 Maximum profit vs. mean demand m = cu, co = cu(1/α−1)

The linear approximation implies that are constants π_j and b_j derived from the slope and intercept of the regression line for product j such that the expected profit can be approximated as follows

$$E\left[\Pi_j\left(D_j(y)\right)\right] \approx \pi_j \mu_j(y) - y_j\left(b_j + F_j\right).$$

In general, it appears that the quality of the fit improves as the mean increases and as the service level α increases. When $F_j = 0$, Fig. 11.1 shows that the b_j values are positive. This is because the expected profit becomes negative for low enough mean demand, but in these cases $y_j = 0$ will be optimal.

Using $C_j = b_j + F_j$ to combine the constants b_j with the fixed costs F_j, and recalling that $\mu_j(y) = \sum_i \lambda_i q_{ij}(y)$, the retailer's approximate objective function can therefore be written as

$$\Pi^*(y) = \sum_j \pi_j \sum_i \lambda_i q_{ij}(y) - \sum_j C_j y_j. \tag{11.8}$$

This objective can be maximized with respect to y, subject to various constraints such as display space or brand representation in the assortment.

2.3 Properties of the Optimal Assortment

When customers' utilities are Gumbel distributed with homogeneous means, van Ryzin and Mahajan (1999) showed that the optimal assortments form nested sets. This case corresponds to $U_{ij} = U_j$ for all i in this paper's notation. That is, if S^K is the best assortment of size K, then $S^K \subseteq S^{K+1}$ for all K. With nonhomogeneous means U_{ij}, however, this property no longer holds, as demonstrated by the following counterexample. Let $\lambda_i = 1$ and $\exp(a_i) = 10$ for all i and consider the following matrix of $\exp(U_{ij})$ values

		Products		
		1	*2*	*3*
Customers	*1*	1,000	2	1,000
	2	1,000	1,000	2
	3	2	1,000	2
	4	2	2	1,000

Let the unit profits for the three products be 10, 9, 9 respectively. Clearly, the best single product is Product 1. But it can be seen from the table of expected profits below that the best two products are 2 and 3.

	y		Expected Profit
0	1	1	35.6
1	1	0	31.0
1	0	1	31.0

Thus, although Product 1 is the best single product, it is not part of the best set of two products.

Nested set properties do hold for two limiting cases, however. First, let us consider the case in which $a_i = a$ for all i and a is very large. Then rewrite $\Pi^*(y)$ as

$$\Pi^*(y) = e^{-a}\sum_{i,j}\lambda_i\pi_j\left(\frac{y_jX_{ij}e^{U_{ij}}}{1+e^{-a}P_i(y)}\right) - \sum_j C_jy_j. \tag{11.9}$$

As a becomes sufficiently large, the term in parenthesis approaches $y_{ij}X_{ij}e^{Uij}$. Thus, if the products are ordered so that

$$\pi_1\sum_i \lambda_iX_{i1}e^{U_{i1}} - C_1 \geq \pi_2\sum_i \lambda_iX_{i2}e^{U_{i2}} - C_2 \geq \ldots, \tag{11.10}$$

then the optimal assortments will be {1}, {1, 2}, ... for a sufficiently large. This implies that there is an optimal product ordering for the assortment, if the retailer's competition is sufficiently strong, even when consumer preferences are heterogeneous. In microeconomic terms, this might be called the "perfectly competitive" case.

A second special case arises when $\exp(a_i)$ approaches 0 for all i. In this case, the retailer is effectively a monopolist, since any consumer who purchases will choose this retailer. For the case in which $X_{ij} = 1$ for all i, j, every product in the retailer's assortment is in every customer i's choice set. Thus, the optimal strategy for a monopoly retailer is to rank products in order of profitability, based on ranking the expected profits as follows

$$\pi_1\sum_i \lambda_i - C_1 \geq \pi_2\sum_i \lambda_i - C_2 \geq \ldots. \tag{11.11}$$

But if some $X_{ij} = 0$, this property may not hold, because some customers may not consider the retailer's most profitable product and thus would not choose it. *Thus, with considerations sets, there may be no specific nested set property when exp(a_i) approaches 0, for all i.*

2.3.1 Sensitivity to the Retailer's Market Strength

To illustrate the difference in the two rankings (11.10) and (11.11), let us consider an example with 5 customer types and 20 products, where the utilities U_{ij} were generated by taking samples from a uniform distribution on [0, 2]. Let all $X_{ij} = 1$

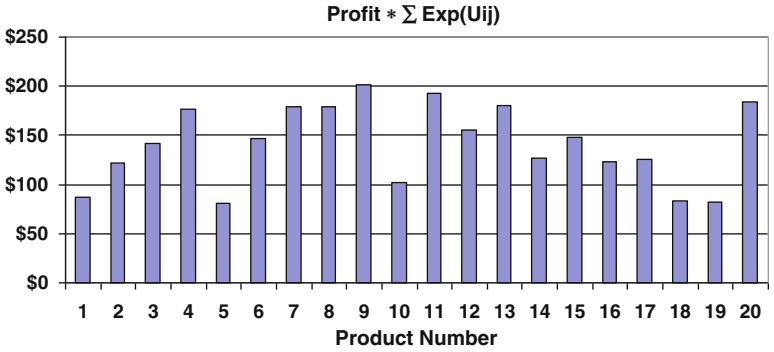


Fig. 11.2 Profitability calculations for 20 randomly generated products

and all $\lambda_i = 1$ in this example. The 20 products are assigned gradually decreasing unit profits π_j : \$10.00, \$9.90, \$9.80, . . ., \$8.10 and fixed costs $C_j = 0$ for all j. Thus, for the case in which the retailer's competitive position is very strong, the products' ranking is based on (11.11), which implies that the products would be ranked in order of the unit profits, 1, 2, 3, ... Therefore, for a retailer with a dominant market position, the optimal assortment of K products is $\{1, 2, \ldots, K\}$.

On the other hand, for a retailer in a weak competitive position, the product rankings are based on the rankings in (11.10). The calculated results for (11.10) are illustrated in Fig. 11.2.

The height of the bars in Fig. 11.2 shows that the expected values for this case are quite different from those that would produce the ranking of 1, 2, 3, ... determined by (11.11). For example, the top 5 products based on ranking the values in Fig. 11.2 are: {9, 11, 20, 13, 7}.

2.4 Solving the Optimization Problem

If the total number of products is small, optimal assortments can be obtained by an exhaustive search, but this becomes more difficult for larger numbers of products. Based on the structure of the problem, certain products may be eliminated from the assortment *a priori*, which reduces the problem size. Substituting the definition (11.6) of $q_{ij}(y)$ into (11.8), the objective function can be written as

$$\text{Max } \Pi^*(y) = \sum_{j\geq 1} y_j\{\pi_j r_j(y) - C_j\} \quad \text{with } y_j = 0,1 \text{ for all } j \geq 1,$$

$$\text{where } r_j(y) = \sum_i \lambda_i \left(\frac{y_j X_{ij} e^{U_{ij}}}{e^{a_i} + \sum_{k\geq 1} y_k X_{ij} e^{U_{ij}}} \right) \tag{11.12}$$

For any y such that $y_k = 0$, define

$$\Delta_k r_j(y) = r_j(y + e_k) - r_j(y), \quad \text{where } e_k = \text{ the unit vector with kth element } = 1.$$

It can be verified that

$$\text{If } y_k = 0, \text{ then } \Delta_k y_j \left[r_j(y) - C_j \right] \leq 0 \text{ for all } j \neq k.$$

This has the implication that if $\pi_k r_k(e_k) - C_k \leq 0$ for any k, then $y_k = 0$ must hold. That is, $y_k = 1$ cannot be optimal since y_k could be changed to 0 and all terms in the objective function will improve or stay the same. This observation can used to eliminate some products before searching on y. However, it appears that an exhaustive search over the remaining 0,1 variables is required to optimize the assortment.

Retailer imposed constraints, such as the number of products must be at least K, or at least one product of Brand B must be included, can be added as linear constraints on y. For example, if the assortment must include at least one product of Brand B, define the logical inputs

$$I_{Bj} = 1 \text{ if product } j \text{ is of brand } B, \text{ and 0 otherwise.}$$

Then the brand constraint is of the form

$$\sum_j y_j I_{Bj} \geq 1 \quad \text{for band } B.$$

We can also include a display space constraint of the form

$$\sum_j d_j y_j \leq D, \text{ where}$$

$d_j =$ the space required for product j
$D =$ total available display space for this category.

These additional constraints also reduce the number of alternatives to be searched.

3 Illustrative Application for a DVD Player Data Base

This section illustrates the application of the optimization model to a set of customer utilities derived from a conjoint analysis of actual consumer Internet responses. The preference data were collected through the ***Active Decisions'*** *Active Buyers Guide Sales Assistant* website. [See www.activedecisions.com. This company has been acquired by Knova Systems, who plan to offer conjoint utility encoding as a

consulting service.] Visitors to *activebuyersguide.com, yahoo.com* and other e-commerce sites completed an interactive survey to elicit their preference tradeoffs for product attributes. These preferences are defined so as to be *independent* of the specific set of products in the market. Product utilities were then derived from additive conjoint analysis of 2,213 customer responses for the DVD player category. That is, each customer's net utility for a particular product was calculated as the sum of his or her "part worths" for the attributes of that product, including the price. (See Green and Srinivasan 1978; Cattin and Wittink 1982; Wittink and Cattin 1989 for discussions of conjoint analysis methods. The conjoint analysis of this data was performed by *Active Decisions* and the author is indebted to them for sharing their results).

The utility values were then normalized by dividing each utility U_{ij} by customer i's maximum utility to obtain

$$S_{ij} = \frac{U_{ij}}{\max\limits_{k\in\Omega} U_{ik}} \quad \text{for all } \; i,j.$$

After this normalization, it was assumed that $\xi_i = 1$ for all i. Consideration sets based on utility thresholds can then be defined as a fixed fraction θ of each customer's maximum utility over all products. That is,

$$u_i = \theta \max_{j\in\Omega} U_{ij}, \quad \text{where } \; \Omega = \text{the set of all products in the market.}$$

Thus, $X_{ij} = 1$ if and only if $S_{ij} > \theta$.

Assortment optimization for this example was done for the case of "large" a_i, i.e., the retailer's competitive position is weak. Thus, the optimal assortments will form nested sets according to (11.10), as discussed previously. Because of the highly competitive nature of the DVD player market and because this retailer was not a dominant player in consumer electronics, this assumption seemed appropriate. However, the database had no data available on retailer preference so this assumption could not be tested.

3.1 Comparing the Model's Predictions to a Retailer's Sales Data

In order to test the predictive accuracy of utilities in the data base and the MNL choice models, we obtained data on the observed selling proportions for an assortment of 30 DVD payers offered by a major retail chain. These selling proportions were compared to those predicted by the MNL choice model fitted to the product attribute utilities in the DVD Player data base. The actual selling proportions of the products ranged from 0.2 to 16 %. [There were 117 different DVD player products at the time the data set was collected, and the retailer data was obtained for the same time period.] A variety of θ values were tested to obtain the correlations and the R-square values are shown below in the table below.

3.1.1 Actual vs. Predicted Selling Proportions for 30 Products

θ	Correlation	R-square
0	69 %	47 %
0.9	78 %	60 %
0.95	79 %	62 %
1.0	72 %	53 %

This table indicates that the fit is reasonably good for all θ values, but the accuracy improves somewhat when customers are assumed to use moderately restrictive consideration sets. Further investigation also revealed that most of the error in these predictions resulted from over-predictions for three products, which the retailer reported were unavailable in some stores. This test supports the use of the utilities in the data base, and also suggests a fairly high θ value such as $\theta = 0.9$ or 0.95 may be appropriate for this data set.

3.2 *Comparing the Expected Revenue of the Retailer's Assortment vs. the Optimal Assortment*

The objective function in (11.12) was then applied to the set of 117 DVD player products available at that point in time to determine the optimal assortment of 30 products. For the optimization, it was assumed that each of the 2,213 respondents to the online survey represents a customer segment of equal size, *i.e.*, the λ_i were assumed to be equal for all i. The fixed costs C_j were set to zero and the product prices from the DVD Player data base were used to compute the expected revenue from a given assortment. Since the revenue comparisons will be done on a percentage basis, it is not necessary to know the actual number of buyers per segment. For percentage calculations with $\lambda_i = \lambda$ for all i, the λ will cancel out of the profit comparisons. Therefore, for the case of "very large" a_i, the objective function in (11.12) can be maximized by substituting a linear objective function that is similar to the ranking calculation in (11.10),

$$\begin{aligned} &\text{Max } \Pi_0(y) = \sum_{j \geq 1} y_j \pi_j X_{ij} e^{U_{ij}}, \\ &\text{subject to } y_j = 0, 1 \text{ for all } j \geq 1 \text{ and } \sum_{j=1}^{n} y_j = 30. \end{aligned} \tag{11.13}$$

The optimal assortments were then determined for various values of $\theta = 0.9$, 0.95 and 1.0, which are captured by changes in the X_{ij}. The table below compares the percentage improvements achieved by the optimal assortment over the retailer's current assortment, for the various θ choices.

θ	Revenue improvement	Common products
0.9	169 %	11 (37 %)
0.95	185 %	9 (30 %)
1.0	208 %	7 (23 %)

The revenue improvements in this table are optimistic because they assume that each customer i's buying behavior exactly matches the MNL model. However, even recognizing this, it appears that using the MNL-based optimal assortment with consideration sets has substantial potential to improve this retailer's revenues.

3.3 The Impact of Customer Preference Structure

The analysis above is based on the use of both consideration sets and heterogeneous customer market segments. To test the impact of these structural assumptions, we focus on three sensitivity questions:

1. *What is the impact of including customer preference heterogeneity in determining optimal assortments?*
2. *How does customers' use of consideration sets impact the optimal assortments and expected profits?*
3. *How does the expected profit increase with assortment size, i.e., how does the optimal assortment size depend on the fixed costs of offering additional products*

3.3.1 Customer Heterogeneity

To examine the role of customer preference heterogeneity in developing the optimal assortment, optimal assortments for homogeneous preferences were generated by replacing the S_{ij} with "average" values S_j, which equal the average S_{ij} value over all customer types i. The expected profits for these optimal assortments were then compared to the profits for the optimal assortment with heterogeneous preferences S_{ij} in Fig. 11.3.

The potential revenues of the two optimal assortments converge when essentially all positive utility products are carried by the retailer. However, for assortment sizes 10–30 that are relevant to most retailers, the optimal assortments for heterogeneous preferences result in profits almost twice as large. Examining the contents of the assortments produced by the two methods found only about 5 % common items in the assortments of sizes 5–30. Thus, for this data set, ignoring customer heterogeneity has significant financial consequences and major impacts on the optimal assortment.

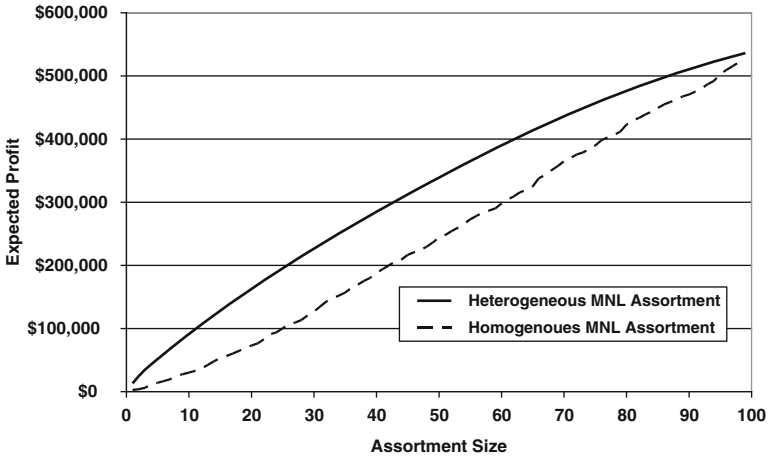


Fig. 11.3 Including preference heterogeneity in assortment optimization

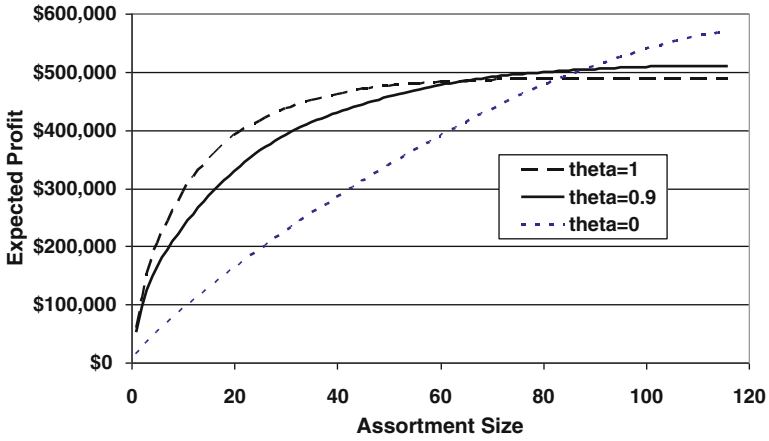


Fig. 11.4 The effect of consideration sets on expected profit

3.3.2 The Use of Consideration Sets

To analyze the impact of consideration sets, the optimal assortments for $\theta = 0$, 0.9 and 1.0 are compared in Fig. 11.4, where $\theta = 0$ corresponds to "no consideration sets." Figure 11.4 shows that when customers use consideration sets and the retailer uses this information correctly in developing the optimal assortments, a substantial increase in expected profit results for typical assortment sizes. For assortments in the 5–10 item range, the $\theta = 0.9$ or 1.0 cases yielded two to three times the profit of the optimal assortment without consideration sets.

Consideration sets allow the retailer to use a more focused assortment. When customers use consideration sets, the retailer can achieve 80–90 % of the maximum possible profit with an assortment sizes of only about 30 items, while these assortment sizes can achieve only about 50 % of the maximum without consideration sets. For $\theta = 1$, all customers can receive their first choice product with an assortment size of 66, but for $\theta = 0$ additional products always increase expected sales.

The shape of the curves in Fig. 11.4 also determines the impact of the fixed costs C_j on the optimal assortment size. For an assortment of size 30, for example, the slopes of the lines are approximately, \$3,500, \$5,000 and \$6,000, respectively, which correspond to the marginal benefits of an additional product. [These dollar figures correspond to one purchase by each of the 2,213 active shoppers in the category. This level of sales would correspond to an aggregate across multiple stores.] Thus, consideration sets allow high fixed costs to be justified for small numbers of products, but tend to limit the optimal assortment size as the number of products increases.

It was assumed in Fig. 11.4 that the optimal assortment was determined for the correct θ value in each case. But since customers' behavior with regard to consideration sets may be difficult to predict, it is interesting to consider the impact of incorrect assumptions about consideration sets. This calculation is illustrated in Fig. 11.5, where the optimal assortment for $\theta = 0$ was used when the correct value was $\theta = 0.9$, and vice versa.

This shows that if customers form consideration sets based on $\theta = 0.9$, the optimal assortment for $\theta = 0$ results in a reduction in expected profit of 12–50 % for assortments in the range of 10–30. On the other hand, if customers do not use consideration sets to prescreen the products, i.e., $\theta = 0$ is correct, the optimal assortment for $\theta = 0.9$ results in a 10–20 % reduction in expected profit. Thus, for this data set, the less risky alternative is to assume that customers do use considerations sets.

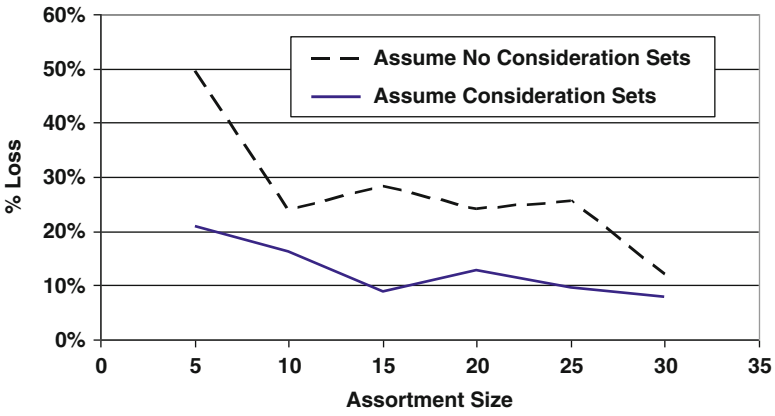


Fig. 11.5 Percentage profit loss for incorrect consideration set assumption with $\theta = 0, 0.9$

4 Summary and Conclusion

This chapter has developed an operational model for assortment optimization, based a multinomial logit choice model with general heterogeneous customer preferences. The structure of the model allows the required input parameters for product choice and retailer choice to be estimated separately from product sales and retailer market shares. These estimates can be based on observed consumer choices for previous assortments, which need not be optimal. The linear approximation of the newsvendor cost function assumes that temporary stockouts result in lost sales, which restricts the model's use to retailers or categories of products with relatively high service levels. However, this assumption leads to a closed form objective function that captures the impact of the assortment on both retailer choice and product choice. While the optimal assortments may no longer form nested sets for heterogeneous preferences, it is shown that the special cases of perfect competition and retailer monopoly do lead to different sequences of optimal nested sets, and it is illustrated how the optimal assortment transitions between these two extremes as the retailer's market share increases.

The optimization model can accommodate a variety of additional retailer constraints. For example, it may be important to: (1) require that certain top brands be represented in the assortment; (2) provide some level of assortment stability across time for customers; (3) stay within a given display space constraint; or (4) carry products with the full range of price points to promote the image of a category killer. The analysis of the DVD player data base illustrated the decreasing marginal benefits associated with increasing assortment size and also the sensitivity of the optimal assortment to the input assumptions regarding the customer choice process. Including customer heterogeneity had significant impacts on both the optimal assortments and the expected profits. Consideration sets, which have been studied in the context of modeling customer choice, but have not previously been included in assortment optimization, were found to strongly influence the optimal assortment for the DVD player data base. This analysis supports the importance of using a consumer choice model that includes heterogeneous preferences and consideration sets in obtaining optimal assortments. The sensitivity analysis also illustrates the potential profit improvement for additional selling effort designed to influence customers' product choices.

There are a number of promising avenues for future research. Clustering customers into fewer classes can reduce the problem size and lead to shorter computation times for the general competitive case. Analytical methods for choosing the best customer clusters for a given database of utilities could therefore extend the applicability of the optimization model. Clusters based on customers' preferences for product attributes, as opposed to individual product utilities, may lead to clusters that are more stable over time. Better optimization approaches that exploit the specific structure of the assortment problem may also exist. It is hoped that this chapter will also lead to additional research on the development of decision support systems for assortment planning that implement this optimization model for choosing assortments, taking into account both product choice and retailer choice.

Acknowledgement The author is grateful to Dale Achabal, Kirthi Kalyanam, Shelby McIntyre and Chris Miller for many valuable discussions and to Active Decisions, Inc. for providing the data base that was used for testing the optimization model. This research was partially supported by the Retail Workbench Research and Education Center at Santa Clara University.

References

Andrews, R. L., & Srinivasan, T. C. (1995, February). Studying consideration effects in empirical choice models. *Journal of Marketing Research*, 32, 30–41.

Ben Akiva, M., & Lerman, S. (1985). *Discrete choice analysis*. Cambridge: MIT Press.

Bucklin, R., & Lattin, J. (1991, Winter). A two state model of purchase incidence and brand choice. *Marketing Science*, 10, 24–39.

Cachon, G., Terwiesch, C., & Yi, X. (2005). Assortment planning in the presence of consumer search. *Manufacturing and Service Operations Management, 7*(4), 330–346.

Cachon, G., & Gurhan Kok, A. (2007). Category management and coordination in retail assortment planning in the presence of basket shopping consumers. *Management Science, 53*(6), 934–951.

Cattin, P., & Wittink, D. R. (1982, Summer). Commercial use of conjoint analysis: A survey. *Journal of Marketing, 46*, 44–53.

Chen, K. D., & Hausman, W. H. (2000). Technical note - mathematical properties of the optimal product line selection problem using choice-based conjoint analysis. *Management Science, 46* (2), 327–332.

Cintagunta, P. K. (1993). Investigating purchase incidence, brand choice, and purchase quantity decisions of households. *Marketing Science, 12*, 184–208.

Chong, J.-K., Ho, T.-H., & Tang, C. (2001). A modeling framework for category assortment planning. *Manufacturing and Service Operations Management, 3*(3), 191–210.

Dobson, G., & Kalish, S. (1988). Positioning and pricing a product line. *Marketing Science, 7*(2), 107–125.

Dobson, G., & Kalish, S. (1993). Heuristics for positioning and pricing a product line using conjoint and cost data. *Management Science, 39*(2), 160–175.

Green, P. E., & Krieger, A. M. (1985). Models and heuristics for product line selection. *Marketing Science, 4*(1), 1–19.

Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research, 5*(2), 103–123.

Guar, V., & Honhon, D. (2006). Assortment planning and inventory decisions under a locational choice model. *Management Science, 52*(10), 1528–1543.

Hanson, W., & Kalyanam, K. (2006). *Internet marketing and e-commerce*. Cincinnati, OH: Southwestern College Publishing.

Honhon, D., Guar, V., & Seshadri, S. (2010). Assortment planning and inventory management under stockout based substitution. *Operational Research, 58*(5), 1364–1379.

Kahn, B. E., & Lehmann, D. R. (1991, Fall). Modeling choice among assortments. *Journal of Retailing*, *67*, 274–299.

Kohli, R., & Sukumar, R. (1990). Heuristics for product-line design using conjoint analysis. *Management Science, 36*(12), 1464–1478.

Kok, A. G., & Fisher, M. (2007). Demand estimation and assortment optimization under substitution: Methodology and application. *Operational Research, 55*(6), 1001–1021.

Kok, A. G., Fisher, M., & Vaidyanathan, R. (2015). Assortment planning: review of literature and industry practice. In N. Agrawal & S. Smith (Eds.), *Retail supply chain management* (2nd ed.). New York: Kluwer Academic Publisher.

Mahajan, S., & van Ryzin, G. (2001). Stocking retail assortments under dynamic substitution. *Operational Research, 49*(3), 334–351.

McBride, R. D., & Zufryden, F. S. (1988). An integer programming approach to the optimal product line selection problem. *Marketing Science, 7*(2), 126–140.

Miller, C. M., Smith, S. A., McIntyre, S., & Achabal, D. (2010). Optimizing and evaluating retail assortments for infrequently purchased products. *Journal of Retailing, 86*(2), 159–171.

Rusmevichientong, P., & Topaloglu, H. (2012). Robust assortment optimization in revenue management under the multinomial logit choice model. *Operational Research, 60*(4), 865–882.

Roberts, J. H., & Lattin, J. M. (1991, November). Development and testing of a model of consideration set composition. *Journal of Marketing Research*, *28*, 429–440.

Roberts, J. H., & Lattin, J. M. (1997, August). Consideration: Review of research prospects and future insights. *Journal of Marketing Research*, *34*, 406–410.

Sauré, D., & Zeevi, A. (2013). Optimal dynamic assortment planning with demand learning. *Manufacturing and Service Operations Management, 15*(3), 387–404.

Siddarth, S., Bucklin, R., & Morrison, D. (1995, August). Making the cut: Modeling and analyzing choice set restriction in scanner panel data. *Journal of Marketing Research*, *32,* 255–266.

Smith, S. A., & Agrawal, N. (2000). Management of multi-item retail inventories systems with demand substitution. *Operational Research, 48*, 50–64.

van Ryzin, G., & Mahajan, S. (1999). On the relationship between inventory costs and variety benefits in retail assortments. *Management Science, 45*, 1496–1509.

Wittink, D. R., & Cattin, P. (1989). Commercial use of conjoint analysis: An update. *Journal of Marketing, 53*(3), 91–96.

Chapter 12
Multi-location Inventory Models for Retail Supply Chain Management

A Review of Recent Research

Narendra Agrawal and Stephen A. Smith

1 Introduction

Research on multi-level inventory systems is critical to retail supply chain management. Multi-level systems are commonly observed in most retail environments, where regional distributions centers (warehouses) stock products to replenish inventory at the retail stores. There is a rich and vast literature in the field of operations management that focuses on the design and management of multi-echelon inventory systems, which can be applied to retailing. Even so, a variety of open problems remain, and this continues to be a fruitful area for researchers. While more than two echelons are also observed in practice, most retailers now prefer to move toward the simpler, two-echelon systems. Such structures are common even in pure play "E-tailers," such as Amazon.com. Amazon.com started with the idea of owning no distribution centers at all, and relying on direct shipments of books from publishers to customers for demand fulfillment. However they now manage a small number of distribution centers, and use a combination of direct shipments from vendors and shipments from their warehouses for demand fulfillment. Traditional "bricks and mortar" retailers today also face the problem of designing inventory management systems for items that are purchased through their Internet sales channels, in combination with normal store replenishment.

This review paper covers a subset of the research on this topic. Because of the vastness of the literature on multi-level inventory systems, we felt it was important to limit the scope of our survey in a meaningful way. First, we restrict our attention to papers after 1993, and refer the reader to the reviews in other papers for articles prior to 1993. For example, Axsater (1993a), Federgruen (1993), and Nahmias and

N. Agrawal (✉) • S.A. Smith
Department of Operations Management and Information Systems, Leavey School of Business, Santa Clara University, Santa Clara, CA 95053, USA
e-mail: nagrawal@scu.edu

N. Agrawal, S.A. Smith (eds.), *Retail Supply Chain Management*,
International Series in Operations Research & Management Science 223,
DOI 10.1007/978-1-4899-7562-1_12

Smith (1993) contain excellent reviews of the work up to that point. We discuss some of the earlier articles that provide foundations for results that we are presenting, or were not included in the reviews listed above. Second, we omit papers on certain model formulations that are not typical of retail inventory management. For example, we exclude the literature on serial systems, since they are not representative of typical retail chains, and are a special case of general multi-location multi-echelon systems. Also excluded are papers that assume deterministic demand, since demand uncertainty is a key aspect of most retail systems.

Finally, we focus our attention primarily on periodic review systems. Most retail chains today employ technologies such as point-of-sale (POS) scanner systems that provide real time access to sales and inventory data. Consequently, in principle, continuous review models could be an appropriate construct for these retail systems. However, two issues limit the practical applicability of this assumption. First, due to contracts with vendors and shipping companies, shipments occur primarily on a pre-specified schedule, and often a variety of items are delivered simultaneously. Second, despite the real time access to sales information, the ERP databases and inventory allocation algorithms are typically updated periodically. Thus, strictly speaking, inventory decisions must be made by planners according to predefined cycles. Consequently, periodic review systems are a better representation of the inventory management systems used by most retailers. For the sake of completeness, in the appendix we briefly present the formulation of some continuous review models along with a few key references.

The rest of the paper is organized as follows: We begin by discussing the key modeling issues in Sect. 2. In Sect. 3, we present the general formulation for periodic review inventory model, and review the relevant literature. Key conclusions and opportunities for further research are discussed in Sect. 4. The continuous review model is discussed briefly in the Appendix.

2 Modeling Issues

2.1 The Key Decision

The fundamental decision to be made in two-echelon retail inventory systems is the appropriate division of inventory between the central (warehouse) location, and each of the retail stores.[1] Clearly, more inventory at the retail stores provides a higher service level to customer demand, but this also increases costs associated with carrying the inventory. The holding cost is higher at stores, due to increased shrinkage and because space in retail stores is typically more costly than warehouse space.

[1] Earlier papers used the term "retailers" to refer to individual retail locations, while more recent papers have used the term "stores." In this paper, we will use the term stores or retail stores for the lowest echelon level in the inventory system.

Higher costs also result from transporting additional items to stores, which increases the product's value. Also, immediate distribution of a large proportion of the inventory to stores makes it difficult to address subsequent inventory imbalances across stores, because lateral shipments between stores are not part of normal replenishment. That is, keeping additional inventory at the warehouse offers the advantage of risk pooling, since inventory can be directed to those stores that need it most. This can potentially reduce overall inventory investments and costs. However, the resulting shipment delays may adversely affect customer service levels. This type of risk pooling has been referred to as the *depot effect*. The other advantage of having a warehouse is the possibility of risk pooling over the length of the replenishment lead time from the external supplier. This is sometimes referred to as the *joint replenishment effect*. In other words, while replenishment orders placed by the warehouse take into account actual demands at the retail stores, the actual decision to allocate this inventory to stores can be delayed until the replenishment order is received. The additional demand information gained during this lead time can be used to make more efficient inventory decisions. Note that this benefit can be realized even if the warehouse holds no inventory.

2.2 Modeling Demand

The Poisson distribution is often used to model retail store demand, using a probability function of the form

$$P\{\text{Demand} = k\} = e^{-\lambda}\lambda^k/k! \qquad k = 0, 1, 2, \ldots$$

with mean = variance = λ. The Poisson distribution is a particularly attractive assumption for modeling demand in continuous review systems because it requires only a single parameter (λ), and the resulting analysis is more tractable.

When mean demand per period is large, the normal distribution can be used to approximate the Poisson. To model discrete demand, the discrete probabilities can be approximated by

$$P\{\text{Demand} = k\} = \Phi\left(k + 0.5\middle|\mu, \sigma\right) - \Phi\left(k - 0.5\middle|\mu, \sigma\right) \qquad k = 0, 1, 2, \ldots$$

where $\Phi(x|\mu,\sigma)$ = normal cumulative distribution with mean μ and variance σ^2.

Some empirical studies of retail data (e.g., Agrawal and Smith 1996) have found that retail demands are more variable than the Poisson distribution, which has a fixed variance to mean ratio of one. There are some practical reasons why actual demand may have higher variance than would be predicted by a Poisson distribution. Random variations may occur in the underlying Poisson arrival rate due to the weather, competitors' promotions, or special events that are not captured by the inventory system's forecasts. Second, customers whose purchases are Poisson arrivals may introduce additional variability by purchasing multiple items of the

same kind. The normal distribution can accommodate more variation, by selecting a larger variance, but the empirical analysis mentioned above found that the normal distribution fit low demand items poorly because it assigns probability to negative values and because it is symmetric about its mean.

This suggests that a compound Poisson distribution or a negative binomial distribution may provide a better choice for modeling retail store demand. In particular, the negative binomial can be generated either from a Poisson distribution whose parameter λ has a gamma distribution, or from a compound Poisson with a geometrically distributed purchase quantity. Agrawal and Smith (1996) found that the negative binomial fit the store level demand data better than either the Poisson or normal distributions. The negative binomial distribution with parameters N and p has the following discrete probability function:

$$P\left(D = k \middle| N, p\right) = f_k(N,p) = \binom{N+k-1}{N-1} p^N (1-p)^k,$$
$$0 < p < 1, \quad N > 0, \quad k = 0, 1, \ldots$$

where the cumulative probability distribution is

$$F_k(N,p) = \sum_{j=0}^{k} \binom{N+j-1}{N-1} p^N (1-p)^j.$$

The mean and variance are

$$\mu = N\left(\frac{1}{p} - 1\right), \text{ and } \sigma^2 = N\left(\frac{1-p}{p^2}\right).$$

The ratio of the variance to the mean is 1/p, which is greater than one and can be arbitrarily large. This makes the negative binomial distribution particularly attractive for retailing applications that have high demand variability.

Other assumptions for modeling retail demand include the Gamma (Bradford and Sugrue 1990), Gumbel (Lariviere and Porteus 1999), and the general exponential family of distributions (Agrawal and Smith 2012).

We also note that the majority of papers assume that demand at different locations is independently distributed. There are a few exceptions that allow correlations across stores or across time, which are described later in this chapter.

Finally, in any store level model, it is important to specify assumptions regarding the treatment of excess demand at the stores. Primarily for analytical tractability, most papers assume that unmet demand is backordered, not lost. While backordering is common for some classes of expensive retail items, excess demands for most department store and grocery items result in lost sales to another retailer, or possibly substitution of another item in the store. Backordering can serve as a good approximation to the lost sales case, provided that the inventory service level at the store is sufficiently high.

A few researchers have assumed lost sales for unmet store demands. Because of the complexity of modeling lost sales, these papers generally assume that the latest store demand information is available with zero delay prior to store replenishment. This zero delay assumption is generally correct in today's retail environment, since electronic data interchange (EDI) can provide essentially continuous communication of demand information across locations, and stores are typically replenished after hours, when no sales are occurring. But the lost sales case is significantly more complex analytically than the backorder case. With lost sales, the inventory level at any time t depends on all the individual demands and replenishments that have occurred previously, while in the backorder case, computing the inventory level requires knowledge of only the total demand over the previous periods. That is, in the backorder case, the inventory level at time t ($IL(t)$) follows from the well known relationship between inventory position (IP(t)) and total demand during the lead time ($D(t-L,\ t)$), where $IL(t) = IP(t-L) - D(t-L,\ t)$. Therefore, knowledge of the actual demand or order placed in every period is not needed to determine the inventory level in a given period. This does not hold for lost sales, adding significant complexity to the analysis.

2.3 Lead Times

Two types of lead times are relevant in such systems. The first is the replenishment lead time at the warehouse for orders placed with external suppliers. Since most researchers assume no capacity constraints on the supplier, these lead times may be assumed to be constant. Exceptions are papers that explicitly model production capacity constraints. We briefly mention this literature later. The second lead time is for orders placed by retail stores at the warehouse. This consists of two components—the shipment time, which is generally assumed to be constant (but may vary across locations), and the lead time due to shortage delays at the warehouse, which is random. Consequently, the effective lead time at the stores, i.e., the sum of the two components is always stochastic due to the possibility of stockouts at the warehouse. It is also a function of the specific allocation rules at the warehouse when shortage occurs. Thus, determining the store lead time distribution is a key analytical challenge.

2.4 Allocation Policies Used at the Warehouse

How the warehouse allocates inventory among competing store demands in shortage situations is a critical determinant of the complexity of multi-location inventory models. It also affects the service level and the cost structure for the retail stores. Conceptually, researchers have considered four different policies for what the warehouse does with the inventory it receives from the external supplier (McGavin et al. 1993). The first policy is essentially a "pass-through," where the warehouse holds no stock, but allocates and ships it to the stores as soon as stock is

received from the supplier(s). This is similar to the cross-docking policy that is practiced at many retail warehouses today. The second policy, called the equal interval policy, attempts to balance the stores' inventory at regular intervals. The third policy is called a two-interval policy, where the warehouse makes two shipments during the period between consecutive replenishments from the supplier. The final policy is called as the virtual allocation policy, where units of inventory at the warehouse are reserved for specific demands as they occur at the retail stores. This essentially imposes a first come first served discipline on demand fulfillment. We will discuss the modeling implications of each of these policies in the next section.

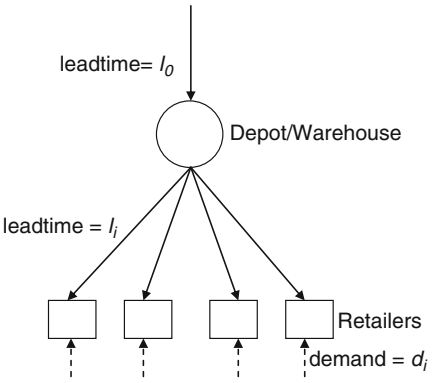
3 The General Periodic Review Inventory Model

Consider a single-item, discrete-time, two-echelon system, where the top echelon consists of a depot (also referred to as the warehouse) which supplies a collection of N retail stores, numbered $1,..,N$, with l_0 and l_i corresponding to the lead times for the depot and the retail outlet i respectively. Random demand occurs in each period at each retail store, with

$D_i(t, t+s) =$ the total demand at location i during periods $t, \ldots, t+s$, and

$$D_0(t, t+s) = \sum_{i=1}^{N} D_i(t, t+s)$$

is the system wide demand during the same period. We let $D_i^{(l)}$ and $D_0^{(l)}$ be the l-period demand at retailer i and the warehouse with cumulative distribution functions $F_i^{(l)}$ and $F_0^{(l)}$ respectively. Unmet demand is backlogged at the retailer, with a penalty cost of p_i per unit backordered and h_0 and $(h_0 + h_i)$ are the inventory holding costs assessed on ending inventory at the depot and the retailer i, respectively.



In each period, we define the following sequence of events:

1. Current period's ordering and shipment decisions are made.
2. Shipments are received.
3. Demand occurs.
4. Holding and penalty costs are assessed based on ending inventory levels.

Define $I_i(t)$ as the echelon stock (stock on hand plus in transit to and on hand at successor points minus backorders from external customers) at location i at the beginning of any period t just after the receipt of a shipment, and $\hat{I}_i(t)$ as the corresponding value at the end of the period t. Define $\hat{I}_i(t) = \hat{I}_i^+(t) - \hat{I}_i^-(t)$. Then $\hat{IP}_i(t)$ and $IP_i(t)$ are the echelon inventory positions just before and after ordering (at the depot) or shipment (if i is a retailer), where echelon inventory position is the echelon stock level plus all orders in transit to that location.

At the end of any period t, the total cost for the whole system, which includes holding and penalty costs, can be expressed as

$$\begin{aligned} & h_0\left(\hat{I}_0(t) - \sum_j \hat{I}_i(t)\right) + \sum_j (h_0 + h_i)\hat{I}_i^+(t) + \sum_j p_i \hat{I}_i^-(t) \\ & \quad = h_0\hat{I}_0(t) + \sum_i \left(h_i\hat{I}_i(t) + (h_0 + h_i + p_i)\hat{I}_i^-(t)\right). \end{aligned}$$

Then, using the notation

$$C_0(t) = h_0\hat{I}_0(t), \text{ and } C_i(t) = h_i\hat{I}_i(t) + (h_0 + h_i + p_i)\hat{I}_i^-(t).$$

The total cost is equal to:

$$C_0(t) + \sum_{i=1}^{N} C_i(t).$$

The expected system costs then depend on the *ordering decision* at the warehouse (which raises the inventory position $IP_0(t)$ of the system to, say, y_0), and on how shipment quantities for retail stores are determined, i.e., the *allocation decision*. Let the corresponding inventory positions at the retailers be denoted by $y_1, \ldots, y_N$. The first decision determines the expected cost at a warehouse at the end of period $(t + l_0)$, and limits the amount to which the aggregate echelon inventory positions of the retail stores can be raised in period $(t + l_0)$. The later decision is particularly relevant in case of shortage situations. These decisions are not independent, which makes the overall optimization problem challenging. So, the upper limit on the aggregate echelon inventory position of the stores can be specified as

$$\sum_{i=1}^{N} IP_i(t + l_0) \leq y_0 - D_0(t, t + l_0 - 1).$$

Obviously, these decisions influence the cost at echelon i at the end of period $(t + l_0 + l_i)$. Therefore, the effect of decisions made in period t, $C(t)$, is

$$C(t) = C_0(t + l_0) + \sum_{i=1}^{N} C_i(t + l_0 + l_i).$$

Thus, for any given ordering policy, the expected long-run average cost is given as

$$\lim_{T\to\infty}\frac{1}{T}E\left[\sum_{t=0}^{T-1}\sum_{i=0}^{N}C_i(t)\right]=\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}E[C(t)].$$

Minimization of the long run average expected value of this function is the overall objective in the two echelon system.

3.1 Solution Methodologies

Determining optimal strategies for general two echelon systems remains difficult. Consequently, most papers use approximations. While some papers make use of relaxation techniques to obtain bounds on the true costs or profits, others impose specific restrictions on the class of inventory policies and then determine the optimal policy within that class. In all cases, the issue of inventory allocation must be addressed carefully.

The form of the optimal solution can be characterized in special cases. One way of rationing, called as the *myopic allocation method*, allocates the echelon stock of the warehouse at the beginning of period $(t + l_0)$ such that the sum of the expected costs at the stores in period $(t + l_0 + l_i)$ is minimized, without regard to later periods. A relaxation of this problem allows the quantities allocated to stores to be negative (by ignoring the constraint that the retail stores' inventory positions must be greater than at the beginning of period $t + l_0$). This is called as the *balance assumption*. The key advantage of the balance assumption is that the echelon stock (sum of the total inventory in the system) suffices to determine the warehouse ordering decision. Further, it also makes the myopic allocation policy optimal. The drawback is that this approach gives up the risk pooling advantage associated with holding stock back at the warehouse. In any case, the balance assumption underestimates the total costs since it is a relaxation. However, absent these assumptions, it turns out that base stock policies are not optimal for such systems (Clark and Scarf 1960). Van Donselaar and Wijngaard (1987), Eppen and Schrage (1981) and Federgruen and Zipkin (1984a) discuss the consequences of making this assumption in detail. These early papers consider special cases of the problem: for example, Eppen and Schrage (1981) consider a two echelon model with identical retailers and a depot that doesn't carry any stock. Jackson (1988) extends the Eppen and Schrage model to allow the warehouse to carry stock, while Jackson and Muckstadt (1989) allow non-identical retailers, but with identical cost parameters. Federgruen and Zipkin extend the Eppen and Schrage model to include non-identical retailers, non-stationary demand, and (*s*,*S*) ordering at the warehouse, but they determine their allocation policies under the assumption that the warehouse is stock-less. Jonsson and Silver (1987) also assume that the warehouse is stock-less, but extend the Eppen and Schrage model to include the possibility of a single, complete

redistribution of inventory between the retailers in the period before the end of any review cycle for the warehouse. Erkip et al. (1990) consider a model like Eppen and Schrage (1981) but allow demand correlation across retailers as well as time. Chen and Zheng (1994) develop lower bounds for costs, based on a cost allocation mechanism, for serial, assembly and distribution systems. Our system is an example of their distribution system.

McGavin et al. (1993) model a system with identical retailers, zero lead times for shipments from the warehouse to each retailer, centralized control and periodic replenishment at the warehouse. The overall stock allocation consists of four decisions: the number of withdrawals from the warehouse stock (which is an opportunity to allocate inventory to retailers), the time between these withdrawals, the quantity withdrawn, and the division of the withdrawn stock to each retailer. The first three decisions are set when the warehouse is replenished and the last one depends on retailer inventories. In particular, they model two opportunities for allocating stock from the warehouse to the retailers, which need not be equally spaced between warehouse replenishments. They seek to determine the effective timing of these two instances and the allocated quantities, so as to minimize lost sales per retailer. This assumption of lost sales makes this paper's contribution a significant departure from the majority of the literature in this stream of work. However, as noted before, this requires the retailer lead time to be zero. They show that the best allocation policy is one that balances retailer inventories (i.e., maximizes the minimum retailer inventory). Heuristic policies are developed assuming that the number of retailers is infinitely large, and are numerically tested in the finite retailer case. In particular, they test the 50/25 heuristic, where the first interval is 50 % of the replenishment cycle and the second withdrawal quantity is 25 % of the replenishment cycle's mean demand. The resulting analysis suggests the insight that the choice of the withdrawal quantity and division of inventory may matter more than the number of withdrawals.

Ahire and Schmidt (1996) consider a mixed continuous and periodic review system with one warehouse and multiple, non-identical retailers. While the retailers follow a continuous review (r, Q) policy, the warehouse follows a periodic review policy (with review period T). At the warehouse, the review period is divided into equally spaced intervals, where at each such point, a group of identical retailers (say, within a geographic zone) are reviewed. Each such zone, however, is reviewed only once per review cycle. The implication of this setup is that the retailer system is equivalent to a (nQ, r, T) system. The lead time consists of a deterministic component, the shipping lead time from the warehouse, and a stochastic component, due to possible shortages at the warehouse (however, order splitting is not allowed), and due to the fact that their orders are only reviewed periodically. Thus, an order may have to wait for anywhere from 1 to T periods before it is even reviewed by the warehouse. Results from Little's Law are used to approximate the shortage delays. Retailer demand is assumed to be Poisson, while the warehouse demand is approximated by a normal distribution, whose parameters are computed. The resulting approximations for financial and operational performance metrics compare well to those obtained through simulation.

Graves (1996) considers a general distribution network following a periodic review, order up to policy at each location. Under the assumption that each location orders at pre-set and known times, he specifies a *virtual allocation* policy where a unit at the supply location is committed/reserved for each unit demanded at the time of the occurrence of the demand. This assumes that the warehouse has real time visibility into the retail demand. Shipments, however, occur only at the next appropriate time after order receipt. The committed units can not be used to satisfy any other order. Unmet demand at the warehouse is backordered and satisfied in a first-come-first-served manner. Independent demand occurs at each retail location following a Poisson process, and excess demand is backordered. Since the order interval is present and excess demand is backordered, each location orders an amount that equals the total demand since the last order. The analysis requires the characterization of the run-out time, the time at which the warehouse runs out of inventory to allocate to the retail sites. The demand at the warehouse is approximated with a Negative Binomial distribution, whose moments can be determined. Various performance metrics can then be quantified using this approximation. Diks and de Kok (1998) model a general N-echelon divergent system where every location can hold stock, and determine policies that minimize long run average costs.

This idea of pre-set, staggered schedules for ordering is also considered in Chen and Samroengraja (2000). In a one-warehouse, multi-retailer model, where retailers are identical, and face i.i.d. demands, they assume that the warehouse follows a periodic review (s, S) policy to receive shipments from a source of unlimited supply with lead time L. The warehouse orders are based on its local inventory position. Between consecutive warehouse ordering epochs, the retailers, whose ordering points are pre-set and equally staggered with groups of retailers ordering at each such epoch, place orders, following base-stock policies with a common order up to level. Two different allocation policies are evaluated. The first, called past priority allocation (PPA) backlogs the unmet demand from a retailer, and fills it in a first-come-first-served manner from the inventory at the warehouse. However, actual shipment occurs only at the next epoch when the retailer places an order with shipment lead time l. The second policy, called current priority allocation (CPA) gives priority to the current order and backorders for the retailer designated to order in a given period. Thus, under PPA, the warehouse may carry inventory earmarked for a retailer while it denies inventory to orders from other retailers. In the second case, some retailers may be backlogged for several consecutive periods while others get replenished. The PPA model lends itself better to exact analysis. Solutions for this formulation are obtained through an approximation procedure. The CPA model is harder to evaluate exactly, but simulation studies indicate that the optimal policies are close to those under the PPA regime. Unlike in the Graves (1996) paper where inventory at the warehouse is committed to demands as they occur, here, the allocation decision is delayed until the retailer actually places an order. Their derivation of the exact cost function in the PPA case is based on a different accounting scheme. Warehouse holding costs occurring in period $(t+L)$ are charged to period t. For retailers, in period t, they charge the total holding and

backorder cost over the next N periods (N is the number of ordering epochs within each warehouse cycle) for the retailer designated to order in that period. The exact calculation under the CPA method is difficult since the distribution of a retailer's inventory position at any time depends not only on the inventory position L periods ago, but also on the exact pattern of deliveries from the outside supplier.

Continuing in the spirit of generalization, Axsater et al. (2002) allow the retailers to be non-identical. The warehouse holds stock and orders from an external supplier in multiples of a given batch size, receiving shipments after a fixed lead time. Lead times for shipping to retailers is constant, but can vary by retailer. Instead of the balance assumption, they consider the virtual assignment rule, where the inventory ordering decision at the warehouse accounts for all retailer inventory positions and assigns inventory to retailers as soon as orders are placed. The final inventory allocation, however, is made only upon the arrival of the replenishment. This is a more restrictive policy that overstates costs. Instead of the myopic allocation policy, they consider a two-step allocation policy, which allows some inventory to be retained at the warehouse. Essentially, at the beginning of each period, the remaining time until the next ordering opportunity is assumed to consist of two intervals, the second one being a single period, at which point reallocation can be done again. An optimization methodology is developed under these assumptions and the results are found to compare very favorably with the case of balance assumption and myopic allocation.

Under the balance assumption, Dogru et al. (2013) establish the convexity of the cost function for the infinite horizon case and discrete demand case, which implies the existence of optimal policies that are base stock policies. They also characterize newsvendor inequalities that must be satisfied by the optimal solutions. For example, for the special case of identical retailer holding and penalty costs at the retailers, and under the myopic allocation and balance assumptions, the well known *critical fractile* solution yields the optimal stocking policy for each location.

3.2 Batch Ordering

The use of batch ordering policies imposes additional complexities on the model since the demand at the warehouse is no longer a simple convolution of the retailers' individual demands. Further, if the retailers follow a periodic review policy, a retailer's order consisting of multiple batches may have to be split across multiple shipments. Of course, the issue of allocation of scarce warehouse inventory remains. Analytically, the key challenge is to determine the distribution of the retailers' replenishment lead time, which consists of both the shipping time (constant) and additional delays due to shortages at the warehouse. Two approaches have been used in the literature for this purpose. One is to evaluate when a batch is ordered by the retailer relative to when the warehouse orders it (as in Svoronos and Zipkin 1988). The second is to evaluate when a batch is ordered by the warehouse relative to when the retailer orders it. In cases with a single warehouse, the later approach is more tractable. This is the approach used in the following two papers.

Cachon (1999) considers a one warehouse N (non-identical) retailer model where the retailers as well as warehouse follow (R, nQ) policies. Retailers follow a periodic review policy with period T, but the ordering process is balanced in the sense that a fixed number N/T of retailers order every period. Unmet demand is backordered, and partial fulfillment is allowed. Retailer orders are randomly shuffled upon receipt, and fulfilled in a first-come-first-serve manner. Exact expressions are derived for costs, as well as demand variability at the warehouse. The key result is that the warehouse demand variability decreases due to balancing (rather than synchronizing retailer orders, where all retailers order simultaneously). Further, under a balanced system, increasing the length of the review period T and decreasing the order batch size also helps lower the supplier's demand variability. However, these strategies may not necessarily decrease total supply chain costs, since they might actually increase the retailers' ordering or inventory costs.

Cachon (2001a) considers a similar model but with identical retailers, and where each location reviews and orders in each period. All locations follow a batch ordering policy. Demand is stochastic and discrete. Average inventory and backorder levels and fill rates are evaluated exactly at each location. Safety stock requirements are determined exactly at the retailers, but approximately at the warehouse.

3.3 Lost Sales

All papers described thus far assume that unmet demand is backordered, McGavin et al. 1993. Another exception is Nahmias and Smith (1994), which focuses on a one warehouse multi retailer system, and assumes that a given fraction of unmet retailer demand is lost. Order up-to policies are used at the retailers, and the replenishment lead time from the warehouse is assumed zero. The warehouse also uses an order up to policy with zero lead times. The length of the review period at the warehouse is a multiple of the retailer's review period, and the stock levels are such that shortages only occur in the mth period within any cycle. This assumption, along with that of zero lead times, is necessary to lend tractability to the model.

In contrast to most other papers, they assume that the demand at the retailers follows a negative binomial distribution, which has been shown to fit retail data well (Agrawal and Smith 1996) because the variance to mean ratio is often larger than one. Since the warehouse supports many stores, the warehouse demand can be approximated by a normal distribution. Exact expressions are derived for the average inventory level and lost sales at stores and the warehouse. Representative retail data is used to illustrate the results and generate managerial insights. For example, they show that the benefits of holding stock at the warehouse depend upon item characteristics—items with low optimal service levels at stores derive the most benefit by holding the majority of the stock at the warehouse. Increasing the

frequency of store delivery can also reduce costs, especially for items that require high optimal service levels at stores.

Anupindi and Bassok (1999) quantify the benefit of centralizing stocks in a single warehouse, two-retailer setting, where a fixed fraction, $1-\alpha$, of unmet demand at the retailers is lost. The remaining customers look for the product at the other retailer. They too assume zero lead times for shipments to retailers. Each retailer faces an independent demand (with known distribution), buys from the warehouse at a unit cost w and sells it to their customers for a price p. Since they consider a stationary, infinite horizon model, the problem boils down to a single period newsvendor-type problem. In the simplest case where $\alpha = 0$, i.e., all unmet demand is lost, they show that centralization does not necessarily increase sales. This depends upon the nature of the demand distribution, as well as the value of the critical fractile. For example, for demand with a normal or exponential distribution, centralization leads to higher sales, while for a Uniform distribution, this happens only if the critical fractile has a value less than 0.77.

In the general case when $\alpha > 0$, the solution corresponds to a Nash equilibrium. They find that the expected total profits for the retailers are greater when stocks are centralized. However, the total sales are greater in the centralized case only if α is smaller than a certain threshold. The manufacturer/warehouse will prefer the centralized case only if α is smaller than a threshold (one interpretation for α in their model is the fraction of customers that, when unsatisfied at a local retailer due to stockouts, search for the goods at other retailers). Interestingly, even the total supply chain profit may decrease due to centralization in some cases. This happens when α is larger than some threshold value, which in turn is a function of the wholesale price w. These insights apply even when coordinating contracts are used. Thus, the main insight from this analysis is that while conventional wisdom dictates that costs decrease (and profits increase) under centralized systems due to risk pooling benefits, this benefit may not result for all parties in the supply chain.

3.4 Decentralized Environments (Quantifying the Value of Information Sharing)

The discussion thus far assumed that the entire supply chain was under central control, and information about all locations was available to the central decision maker. This assumption is not appropriate when the entities at the different echelons operate independently. When decisions are made so as to optimize local incentives, the overall supply chain performance may not be optimal. The consequences of the resulting actions by the supply chain participants include the well known bullwhip effect, as discussed in Lee et al. (1997a, b).

In an early paper, Eppen (1979) showed that in a multi-location model with normal and correlated demand, the total holding and penalty costs are lower in a centralized system than in a decentralized system. This result was later generalized

for other distributions in Chen and Lin (1989) and Stulman (1987), and to include inter-node transportation costs in Chang and Lin (1991).

Recently, however, spurred by the advances in information technology and software solutions, explicitly quantifying the potential value of information sharing in supply chains has been the subject of a number of papers. For example, Cachon and Fisher (2000) quantify this value in the case of a single warehouse multi-retailer environment. The retailers are identical, and use periodic review batch ordering policies. Retailers order periodically, in batches of a given size Q, and receive shipments after a fixed lead time. The warehouse also orders in multiples of Q, and receives its orders from an external supplier after a constant lead time. Inventory is allocated using a batch priority rule, where each batch order is assigned a priority, and shipments are done in the order of priority. By comparing the total supply chain costs with and without information sharing, they conclude that the value of information sharing is rather limited, 2.2 % on average. However, the benefit from shorter lead times and smaller batch sizes was nearly 20 % each. The explanation they offer is that demand information only matters when the retailer inventory levels are very low, since otherwise, they don't need to place orders. However, this is precisely when retailers actually place orders, so essentially, the demand data is already captured in the order information.

Lee et al. (2000) quantify the value of information sharing, albeit in a one warehouse one retailer supply chain. In contrast to the earlier papers which assume the demand is independent and identical across time, they assume that demand at the retailer is auto-correlated [AR(1)], such that

$$D_t = d + \rho D_{t-1} + \varepsilon_t,$$

where $d > 0$, $-1 < \rho < 1$, and ε_t is normally distributed with mean zero and standard deviation of σ. Both locations order every period in a periodic review system, with fixed lead times for shipments to each location. Unmet demand at the retailer is backordered, while at the warehouse excess demand is met with a special order placed at an external supplier at an additional cost. They assume that the manufacturer bears the full cost of guaranteeing supply to the retailer. They characterize the retailer's ordering process, which becomes the demand process for the manufacturer. In the case of no information sharing, the manufacturer only receives the retailer's orders. In the case of information sharing, the manufacturer also receives information about actual demand, which allows him to obtain the value of the error term ε_t, thereby lower demand variability. Since the manufacturer bears the full cost of assuring supply, the retailer's inventory costs remain unchanged with information sharing. However, information sharing leads to lower inventory levels as well as lower costs for the manufacturer. Further, they show that the benefit of information sharing is greater when the auto-correlation or demand variance is high. This analysis is complicated by the fact that when demand is auto-correlated, exact expressions for average inventory levels cannot be derived. Consequently, they make use of approximations for the retailer's and manufacturer's inventory levels.

Chen (1998) also quantifies value of information, but in a serial system with continuous review policies. They report cost benefits in the range of 2–9 %. Gavirneni et al. (1999) also consider a serial system (one warehouse, one retailer), but extend the model to the case where the manufacturer's capacity is limited. By comparing the base case to one in which the manufacturer obtains information about the retailers' demand distribution and inventory policy parameters, they are able to quantify the value of information. They find that the value of information is more compelling when end item demand is not very variable, when the retailer's $(S - s)$ is not very large or very small, or, when supplier's capacity is large. Aviv and Federgruen (1998) also consider the benefits resulting from sharing demand forecasts, also with limited supplier capacity.

3.5 Lateral Pooling

There is a large body of research that focuses on the issue of lateral pooling, also referred to as transshipments. In practice, this is rarely done for low-ticket items, since the cost and time involved in repackaging leftover inventory, shipping it to another location, and unpacking it again can easily wipe out the margins. However, for bigger ticket items, like electronics, expensive jackets and suits, and automobiles, this practice is common. Obviously, the presence of an information technology solution that provides information about inventory levels is a prerequisite for this system. One stream of research on transshipments addresses the problem in the context of repairable items. In the interest of staying focused on the retail environment, we will not review this literature, but instead direct the interested reader to Cohen et al. (2006), Muckstadt (2004), Axsater (1990) and Lee (1987), and the references contained therein. A more recent review of the literature can be found in Paterson et al. (2011).

Since the other locations serve as a backup location from which to fill unmet demand, albeit at some cost, this alters the penalty incurred due to shortages. Similarly, since there is the possibility of selling excess inventory to other locations, it alters the salvage value. Depending upon the cost of transshipment and the terms of the exchange, a retail location may, in some conditions, find it profitable to transfer its inventory to another location even when it has its own demand to meet. Clearly, each location will need to determine rules for when is it appropriate to give up its inventory. In any case, the inventory stocking policy must be modified. A second factor to consider is whether the stocking decisions are made centrally, or in a decentralized manner. In the later case, a game theoretic formulation is necessary to determine the optimal inventory ordering and allocation rules to appropriately model the incentives for each party. This results from the externality created due to decentralized decision making—larger inventory carried by one location could lower the stockout cost for others. Similarly, lower inventory levels at one location make it more economical for another location to dispose of its excess inventory. An important source of

distinction between papers on this topic is whether the redistribution of stock occurs *after* or *before* demand is realized.

We begin with the former category first. Early works on this topic include Krishnan and Rao (1965) and Karmarkar and Patel (1977). Both assume identical costs at retailers, an assumption later generalized by Tagaras (1989). Robinson (1990) formulates the problem for an arbitrary number of non-identical retailers, and shows the optimality of order up to policies. However, analytical solutions can be determined only for the case of identical retailers, or when there are only two retailers. Consequently, Monte Carlo simulation has been used to solve the general case. All these papers assume zero replenishment and shipment lead times. This assumption leads to the result of "complete pooling" (Tagaras 1989), which implies that if transshipment is economically viable, then it is optimal for each location to make its excess inventory available for lateral shipments, *i.e.,* there is no reason for holding inventory back at any location. This logic, *a priori*, may not hold if the replenishment lead times are non-zero. This factor is the focus of Tagaras and Cohen (1992), which we discuss next due to its generality.

Tagaras and Cohen (1992) model a multi-period, one-warehouse, two-retailer locations system, where demands occur independently at the retail locations. Shipments from the warehouse to retailer i arrive after L_i periods. Order-up-to policies are followed by each retailer, who faces a unit holding cost c_{hi} on the ending inventory OH_i as well as shortage cost c_{pi} on the backorders BO_i. Additionally, there is a unit lateral shipment cost c_{ij} incurred for the X_{ij} units shipped from i to j. The transshipment policy is determined by whether the inventory level (or inventory position) at the shipping location i is above a threshold level r_i, and target inventory level t_j, (or inventory position) at the receiving location j, which must not be exceeded after transshipment. Four transshipment policies are thus generated. The first two involve on-hand inventory level as the criteria. In the first case, transshipment occurs only if a location faces a shortage (i.e., $t_i = t_j = 0$). Under the second policy, transshipment can take place even if there are no shortages (i.e., $t_i = r_i = 0, i = 1,2$). Obviously, $r_i = r_j = 0$ implies complete pooling in this case. The third and fourth policies are similar to these two, except that the triggers are inventory positions. The objective is to determine order quantities Q_i that minimize total expected costs, as given by:

$$E(C) = \sum_{1}^{2} \left\{ c_i E(Q_i) + c_{hi} E(OH_i) + c_{pi} E(BO_i) + \sum_{j=1, j\neq i}^{2} c_{ij} E\left(X_{ij}\right) \right\}.$$

Exact analysis of this formulation is mathematically intractable. Consequently, search procedures are used to determine optimal solutions. They also derive heuristics based on the assumption of zero lead times. The key finding is that the complete pooling policy always dominates, as was the case when lead times are zero. In other words, hedging, by holding back inventory, or transshipping in anticipation of shortages is not optimal. Also, the heuristics were found to be near-optimal. These results are extended to the case where the transshipment lead times are non-zero in Tagaras and Vlachos (2002).

Archibald et al. (1997) also consider a two-location model, but assume that unmet demand at a location can be met either through transshipment from the other location, or through an emergency shipment from the supplier (no warehouse is assumed). The demand distribution is assumed to be Poisson. A Markov chain formulation is developed to characterize the optimal policies, which are shown to be of the order up to type. The model is then extended to the case of multiple items with constraints on the amount of inventory that can be carried at any location.

Herer et al. (2006) generalize Robinson (1990) to include more general cost structures, and develop an optimization approach that is guaranteed to converge, as compared to Robinson's heuristic which does not provide such a guarantee. They too assume zero lead times, show optimality of order up to policies, which are computed using Infinitesimal Perturbation Analysis. The transshipment quantities are determined by solving a linear programming formulation.

Bertrand and Bookbinder (1998), on the other hand, consider a general, periodic review model for the case where the redistribution decision is made *before* demand realization. They consider a model with multiple non-identical retailers that are supplied by a warehouse. The warehouse does not carry any stock, but allocates it to stores on the basis of their inventory levels so as to minimize total costs. In the period immediately before the end of the cycle (after which the warehouse orders again), inventory can be redistributed so as to minimize shortage in the last period. The assumption is that shortages primarily occur in the last period in any cycle. The redistribution decision is determined using a greedy heuristic. The optimal policies, and the corresponding costs and service level are determined using simulation, since any analytical treatment is intractable. Similar assumptions were made earlier in Jonsson and Silver (1987), but the objective was to minimize the total number of stockouts.

Anupindi and Bassok (1999), which was discussed earlier, model interactions between retailers when transshipments are possible. Similarly, Rudi et al. (2001) consider interactions between retailers in a game-theoretic setting, although their work is based on ideas contained in earlier papers by Parlar (1988) and Lippman and McCardle (1997). In the later two papers, in case of stockouts, it is the customer demand that is directed to the other location. This is different from the currently assumed scenario more relevant to us where products are transferred (albeit at a cost). Nonetheless, the modeling mechanics are similar. Rudi et al. (2001) consider the interactions between two firms, each modeled as a newsvendor within a single period framework. They assume that transshipment occurs *after* demand is realized, and the number of units exchanged from location i to location j is

$$T_{ij} = \min\left\{ (D_j - Q_j)^+, (Q_i - D_i)^+ \right\}.$$

A unit cost is incurred for each unit shipped, and a unit price is charged that varies by shipping location. The resulting profit functions follow in a straightforward manner from the newsvendor methodology. They characterize the optimal decision in the centralized as well as the decentralized cases by solving for the Nash

equilibrium. The pricing decision is also evaluated. Extending this approach to the case of more than two locations is complicated by the specific construction of the schedule of transshipment prices and costs.

Anupindi et al. (2001) develop a more generalized framework for the analysis of decentralized distribution systems. They assume N retailers who face stochastic demands and hold stocks locally and/or at one or more central locations. An exogenously specified fraction of any unsatisfied demand at a retailer could be satisfied using excess stocks at other retailers and/or stocks held at a central location. The operational decisions of ordering inventory and allocation of stocks and the financial decision of allocation of revenues/costs must be made in a way consistent with the individual incentives of the various independent retailers. They develop a "coopetitive" framework for the sequential inventory and allocation decisions. They define *claims* that allow them to separate the ownership and the location of inventories in the system. For the cooperative shipping and allocation decision, they develop sufficient conditions for the existence of the core of the game. For the inventory decision, they develop conditions for the existence of a pure strategy Nash Equilibrium. They show that there exists an allocation mechanism that achieves the first-best solution for inventory deployment and allocation, and develop conditions under which the first best equilibrium will be unique.

Dong and Rudi (2004) include the consequences of lateral shipments between retailers on the warehouse/manufacturer in their study. However, they do so in a single period setting with identical retailers. Recall that Anupindi and Bassok (1999) solved only the two retailer case. They analyze the case where the manufacturer is a price taker as well as one where he is a price setter (i.e., a Stackelberg leader). Following an analysis in a newsvendor type setting, they find that the benefit of transshipment is no longer guaranteed, rather it depends upon the parameters of the problem.

In an interesting paper, Zhao et al. (2005) formulate the problem faced by a network of decentralized retailers who stock inventory of a common item (they consider this problem in the context of a spare parts dealer network). Each location follow an (S, K) type policy. S denotes the order up to level while K denotes a threshold rationing level such that inventory will be shared with the other dealer only if the inventory level exceeds the threshold. Higher values of K imply that smaller portions of inventory are available for sharing. While demand occurs independently at each location, this possibility of inventory sharing changes the cost structure. Thus, each location needs an incentive to share inventory. Otherwise, it might find it profitable to retain inventory to satisfy future demand (understandably, the complete pooling result does not always hold in the decentralized setting). This manufacturer can either provide incentives for sharing, or subsidize the cost of sharing the inventory. The consequences differ. The first incentive induces the locations to lower their threshold rationing levels instead of increasing their stocking levels. The second induces them to lower their stocking levels, which results in lower service levels. Thus, from the manufacturer's point of view, a combination of such incentives may be best.

3.6 Fashion Products

The majority of the papers discussed thus far model environments in which the product being managed is a basic, replenishable item. In contrast, there is a smaller literature that explores issues relevant to the management of fashion products in large, multi-echelon retail chains. Fashion products tend to have very short selling seasons, with replenishment lead times that may be substantially longer than the length of the selling season. Consequently, these environments differ in that the retailer may have a very limited number of opportunities (often one or two) to place inventory in stores, and demand uncertainty tends to be large. At the same time, for many fashion forward retailers, sales from such products form the bulk of revenues.

For single retail location environments, the problem can be modeled in a straightforward manner using the well known newsvendor formulation. Extensions to the case of multiple locations, but with only a single opportunity to position the retail inventory, are fairly straightforward too. However, the problem is more complicated when there are multiple locations, limited inventory on hand, and more than one opportunity to stock stores. Multiple stocking opportunities also offer the possibility of forecast updates based on observed sales.

Fisher and Rajaram (2000) consider a demand model, with different store types. They consider the problem of determining the optimal set of test stores to stock prior to the beginning of the selling season. Using sales histories of comparable products from a prior season, they cluster the stores in the chain deterministically using a store similarity measure and then choose one test store from each cluster. Then, in the test period, inventory is placed in the test stores so that demand can be observed, from which, regression is used to estimate sales for the season. They use linear regression to estimate forecasts for season sales. Test stores are obtained deterministically by considering only the prior season sales.

Agrawal and Smith (2012) develop a two period inventory decision model for seasonal items at a retail chain with non-identical stores. As is typical in such scenarios, they assume that store demands can be correlated across the chain, and across the two time periods. At the beginning of the second period, demand forecasts and inventory policies can be revised, based on the observed demands in the first period. They develop a generalized Bayesian inference model assuming that the store demand distributions share a common unknown parameter. They also develop a two stage optimization methodology to determine the total order quantity, as well as the initial and revised store stocking policies for the two periods, taking into account the fact that store stocking policies in the first period affect the demand information that is collected. If many stores are stocked in the first period, better information about demand may be possible, but fixed costs associated with stocking stores, especially at low-volume ones, can lower profits. Additionally, ordering and inventory allocation decisions made in the first period also affect the amount of inventory that will be available for stores in the second period. To reduce the state space of this problem, they develop a normal approximation for the excess inventory left over at the end of the first period, which greatly simplifies the analysis.

By comparing the performance of the system under different supply chain flexibility arrangements, they develop counterintuitive insights regarding the magnitude of benefits resulting from (1) using updated demand information to modify store inventory levels and the set of stores that are stocked in mid-season (internal flexibility), and (2) flexible supply arrangements that allow the total replenishment quantity to be adjusted in mid-season (external flexibility). They find that the value from store adjustment can be significant even without learning (i.e., the ability to update demand forecasts based on observed sales) or external flexibility. The incremental value of external flexibility can also be significant, but only if it is accompanied by learning. On the other hand, the value of learning alone was small without either external flexibility or store adjustment capability. Thus, internal flexibility (store adjustment and learning) increases the value of external ordering flexibility.

3.7 Transportation Issues

A closely related problem in multi-location systems is that of determining optimal policies and routes for scheduling vehicles to deliver products to the various retailers in the network. The well known joint replenishment problem is also a part of this stream of work. This area represents a substantial body of research, and we will not review it in this paper. However, we will briefly point to some of the papers, and encourage the interested readers to follow the references therein.

Papers that focus on the joint replenishment problem when demand is deterministic include Jackson et al. (1985), Anily and Federgruen (1991), Federgruen and Zheng (1992), Vishwanathan and Mathur (1997), Speranza and Ukovich (1994) and Bramel and Simchi-Levi (1995). Papers that consider stochastic demands include Balintfy (1964) (can order, must order, order up to levels in a continuous review setting); Silver (1981) and Federgruen et al. (1984) (determining can-order policies); Atkins and Iyogun (1988) (periodic review policies for coordinated replenishments); Pantumsinchai (1992) (heuristics for Q, S policies for multiple items); Viswanathan (1997) ((T, s, S) policies); Pryor et al. (1999) (single item with transportation set up costs), and Cachon (2001b) (single store but multiple items, capacitated vehicles).

There are also many papers that consider vehicle routing along with inventory costs, but the few among these that allow for stochastic demand include Federgruen and Zipkin (1984b), McGavin, et al. (1993), Adelman and Kleywegt (1999) and Reinman et al. (1999).

3.8 Additional Issues

While the focus of the papers discussed thus far was primarily on cost minimization, another approach to system design may be driven by service level targets. For this type of problem, de Kok (1990) assumes that the depot does not carry any stock

and imposes a service level target at the retail locations. This model is extended in Verrijdt and de Kok (1995) for more general N-echelon networks, and in de Kok et al. (1994) to allow the depot to hold stock as well. Diks and de Kok (1998) derive newsvendor equalities for such systems under continuous demand.

In an interesting paper, Erkip et al. (1990) consider a multi-echelon model with multiple retail outlets whose demands may be correlated with each other and also across time, but do not consider forecast revision as demand data becomes available. They model demand at retailer j in period t as

$$d_{jt} = R_j \hat{D}_t L_t + \varepsilon_{jt},$$

where R_j is the average fraction of chain-wide demand at store j, $\hat{D}_t$ is the forecasted chain-wide demand, L_t is the normally distributed (with unit mean) *index variable* for period t, and ε_{jt} is the normally distributed (with zero mean) random forecast error at store j. The index variable parameter, common to all stores, is assumed to be an autoregressive process of order one. This is what induces correlation across stores and time. To lend tractability to their analysis, they need to assume that the coefficient of variation of demand at each store is equal. This assumption, along with the allocation assumption at the warehouse allows them to derive newsvendor type cost minimizing solutions for the problem.

While allocation policies are clearly important in the papers discussed above, this issue is also the subject of other papers developed in the context of assembly/production systems. In this case, when multiple products require the same common component, the available stock of components needs to be allocated in shortage situations. Similarly, in single location problems where there are multiple "classes" of demand, some allocation mechanism must be designed. Comparing these settings to distribution systems, it is clear that in both these cases, the inventory dynamics at the retail locations are not relevant, but the problem of inventory allocation is similar to that faced by the warehouse in our model. Without reviewing in detail, we list some of the papers in this category for the sake of completeness: Collier (1982), Baker et al. (1986), Gerchak and Henig (1986), Gerchak et al. (1988), Ha (1997) and Agrawal and Cohen (2001).

In papers discussed thus far, the locations of the various facilities was given. However, this may very well be a decision if the objective is to design (or redesign) a firm's supply chain network. This is the subject of investigation in Berman et al. (2012). They consider the joint problem of choosing the location of the DCs, assignment of retailers to DCs, and setting inventory policies at the retail locations. Using approximations for the cost average functions at the retail locations, the problem is formulated as a non-linear integer program, and a Lagrangian relaxation method is developed and tested to solve the problem.

Finally, for versions of our problem that include capacity constraints, i.e., capacitated production/distribution systems, see Glasserman and Tayur (1994) and Rappold and Muckstadt (2000), and the references therein.

4 Conclusions

The reviews presented in this paper as well as earlier ones clearly show that much has been accomplished in the area of designing and managing multi-location retail supply chain structures. However, our collaboration with a number of prominent retail chains has identified several of practical issues that have yet to be examined in any detail. The brief description of these issues that follows here is by no means an exhaustive list, and the interested reader should append this list to the other open questions discussed in many of the papers that we have reviewed here.

The trend towards micro-merchandising presents the first set of opportunities. Since local consumer preferences vary by location, retailers are attempting to customize their product assortments and model stocks to such local needs. However, this requires investing in mechanisms and methodologies that can allow retailers to determine what such differences are, and how best to let inventory policies be influenced by such information. Correlations between demand across stores and across time add additional complexity to such decisions in general. Agrawal and Smith (2012) present one approach for addressing this problem. This work can be generalized to include multiple products, multiple planning periods, and the potential to use pricing as yet another instrument for supply chain flexibility.

As we move from planning of one product to multiple products that form an assortment, practical considerations relating to product packaging become important. Products often move in supply chains in the form of pre-packs. For example, for an apparel retailer, a pre-pack might consist of one red, two black, and one grey t-shirt. Such pre-packs may also contain products corresponding to different sizes. Designing such pre-packs is critical to supply chain efficiency. Obviously, smaller pre-packs maximize the ability of stores to match supply and demand cost effectively. However, larger pre-packs minimize packaging and material handling costs throughout the supply chain. They also result in the possibility of shipping more units than are really needed at stores. When retail stores vary greatly in their sales rates, the problem of pre-pack design assumes even greater complexity.

While the mathematical models described in this paper have the ability to make unique inventory decisions at the store level, in practice, for large chains with thousands of stores, managing such a large number of policies is prohibitive. Consequently, stores are often grouped into a manageable number of categories (e.g., 4–10), such that the same policy can be implemented within a category. While mathematically suboptimal, the practical advantages are substantial. However, this raises the interesting question of how best to specify such categories, particularly considering store differences across geographies and product categories.

Pricing and markdown strategies in retail chains are yet another rich area of research. The majority of papers we have discussed here ignore the pricing decision. Most pricing papers that we are aware of are single location models. How best to determine pricing and inventory policies simultaneously across chains is an important research topic for retailing.

Finally, no discussion of the retail industry can be complete without recognizing the tremendous opportunities afforded by multi-channel formats, where retailers attempt to access customers using the traditional store, plus the Internet and catalog channels. Retailers vary greatly in their capabilities to deliver their products and services in this manner, and few appear to have realized any potential supply chain synergies from jointly optimizing such formats. This, we hope, will be a topic that researchers in the area of supply chain management will explore in the coming years.

Appendix: Continuous Review Inventory Systems

Many of the results in this research area, particularly for centrally controlled continuous review systems, grew out of the METRIC approximation derived in the seminal work done by Sherbrooke (1968). Consider a one-warehouse multi-retailer system where inventory is managed using a one-for-one $(S-1, S)$ inventory policy. Further, let the demand distribution at each retailer i be independent and Poisson (λ_i). Then, it follows that the demand faced by the warehouse is Poisson $(\lambda_0 = \Sigma_{i=1..N}\ \lambda_i\ )$. Using Palm's theorem, it then follows that the number of outstanding orders at the warehouse has a Poisson distribution with mean $\lambda_0\ L_0$, where L_0 is the replenishment lead time at the warehouse. Then, for a given order up to level S_0, expressions for expected backorders (B_0), waiting time (W_0) as well as inventory levels (I_0) can be derived as follows:

$$E(B_0) = \sum_{j=S_0+1}^{\infty} (j - S_0)\frac{(\lambda_0 L_0)^j}{j!}\exp\big(-\lambda_0 L_0\big),$$

$$E(W_0) = E(B_0)/\lambda_0,$$

$$E(I_0) = \sum_{j=0}^{S_0-1} (S_0 - j)\frac{(\lambda_0 L_0)^j}{j!}\exp\big(-\lambda_0 L_0\big).$$

While the actual lead time is random, the average lead time for retailer orders now equals the shipping lead time plus the average delay time due to shortages at the warehouse. The problem is that the random replenishment lead times for retailers are not independent, since they all depend upon the inventory situation at the warehouse. The METRIC approximation ignores this correlation, and replaces the random lead time with its expected value. This allows results similar to the ones for the warehouse to be derived for the retailers as well. Thus, cost expressions can be derived and optimized.

Exact expressions can be obtained by characterizing the steady state distributions of inventory levels. While the previous papers focused on characterizing the distribution of the retailer lead times, an alternate approach was taken by Axsater (1990) to develop an exact evaluation methodology for the costs directly. In particular, he observed that any unit ordered by facility i will be used to fill the

S_i-th unit of demand at this facility following that particular order, where S_i is the order up to level. Therefore, the distribution of the time elapsed between an order and the occurrence of the unit of demand that it will satisfy will have an Erlang (λ_i, S_i) distribution, with the following density function:

$$g_i^{S_i}(t) = \frac{\left(\lambda_i^{S_i} t^{S_i - 1}\right)^j}{(S_i - 1)!} \exp(-\lambda_i t).$$

Now, conditioning on the delay at the warehouse (which also has an Erlang distribution similar to the one above), cost expressions for that unit can be derived (consisting of holding and backordering costs). Axsater derived a recursive procedure for evaluating the resulting costs. Thus, this method primarily focuses on keeping track of costs associated with arbitrary supply units.

Such procedures and results become ineffective when we consider general systems where one-for-one policies are replaced by batch ordering policies (R, Q) due to fixed ordering costs. In this case, the demand arising from retailers is no longer Poisson, but Erlang instead. Consequently, the demand process at the warehouse is the sum of N Erlang processes, which is more complicated to analyze.

This generalization is considered in Axsater (1993b), where the author considers a one warehouse multi-retailer inventory system, with N *identical* retailers facing independent Poisson demand. However, all locations are allowed to order in batches using a *(R, Q)* policy, and the policies at the warehouse are defined in terms of retailer batches. Lead times are assumed to be constant. Unmet demand is assumed to be backordered, and costs include proportional holding as well as backordering costs. The basic idea stems from a similar observation in Axsater (1990). In this case, a sub-batch ordered at the warehouse will fill the $(R_w + 1)$th subsequent order for a retailer batch at the warehouse. Of course, this will happen after a random number of system demands. The costs are then derived by conditioning on which subsequent demand triggers an order. Exact as well as approximate evaluation procedures are derived.

Following a similar logic, in Axsater (1997), the results are further generalized to a two-level inventory system with one warehouse N retailers and constant lead times (transportation times), but where the retailers face *different compound Poisson* demand processes. All facilities apply continuous review ***echelon*** stock (R, Q) policies and backorder unmet demands. They provide a method for exact evaluation. Note however that echelon stock based policies may not always dominate installation stock based policies.

The third approach to solving such problems is based on characterizing the steady state distribution of inventory levels. For example, Graves (1985) fitted a two parameter Negative Binomial distribution to the number of outstanding orders for the basic METRIC model. In a similar manner, Chen and Zheng (1997) consider a one warehouse N retailer system where the retailers face different but independent compound Poisson demands, lead times are fixed, and orders are restricted to be batches of some specified lot size. They too assume *installation* stock based replenishment policies. For the case of simple Poisson demands, exact results are

possible. The inventory level at the warehouse can be determined easily, since its echelon inventory position has a uniform distribution. The distribution of the inventory level at the retailer locations is more complicated, for which the authors determine an exact procedure. For the case of compound Poisson demand, approximate evaluation methods are derived.

References

Adelman, D., & Kleywegt, A. (1999). *Price directed inventory routing*. Working paper, University of Chicago, Chicago, IL.

Agrawal, N., & Cohen, M. A. (2001). Optimal material control in an assembly system with component commonality. *Naval Research Logistics, 48*, 409–429.

Agrawal, N., & Smith, S. A. (1996). Estimating negative binomial demand for retail inventory management with unobservable lost sales. *Naval Research Logistics, 43*, 839–861.

Agrawal, N., & Smith, S. A. (2012). Optimal inventory management for a retail chain with diverse store demands. *European Journal of Operational Research, 225*, 393–403.

Ahire, S. L., & Schmidt, C. P. (1996). A model for a mixed continuous-periodic review one warehouse, N retailer inventory system. *European Journal of Operational Research, 92*, 69–82.

Anily, S., & Federgruen, A. (1991). Capacitated two-stage multi-item production/inventory model with joint set up costs. *Operations Research, 39*(3), 443–455.

Anupindi, R., & Bassok, Y. (1999). Centralization of stocks: Retailers vs. manufacturer. *Management Science, 45*(1), 178–191.

Anupindi, R., Bassok, Y., & Zemel, E. (2001). A general framework for the study of decentralized distribution systems. *Manufacturing and Service Operations Management, 3*(4), 349–368.

Archibald, T. W., Sassesn, S. A. E., & Thomas, L. C. (1997). An optimal policy for a two depot inventory problem with stock transfer. *Management Science, 43*(2), 173–183.

Atkins, D., & Iyogun, P. (1988). Periodic versus "can-order" policies for coordinated multi-item inventory systems. *Management Science, 34*(6), 791–796.

Aviv, Y., & Federgruen, A. (1998). *The operational benefits of information sharing and vendor managed inventory (VMI) programs*. Working paper, Washington University, St. Louis, MO.

Axsater, S. (1990). Modeling emergency lateral transshipments in inventory systems. *Management Science, 36*, 1329–1338.

Axsater, S. (1993a). Continuous review policies for multi-level inventory systems with stochastic demand. In S. C. Graves, A. H. G. Rinnooy Kan, & P. H. Zipkin (Eds.), *Handbooks in operations research and management science* (Logistics of production and inventory, Vol. 4, pp. 175–197). Amsterdam: Elsevier Science Publishing Company B.V.

Axsater, S. (1993b). Exact and approximate evaluation of batch ordering policies for two-level inventory systems. *Operations Research, 41*(4), 777–785.

Axsater, S. (1997). Simple evaluation of echelon stock (R, Q) policies for two-level inventory systems. *IIE Transactions, 29*, 661–669.

Axsater, S., Marklund, J., & Silver, E. A. (2002). Heuristic methods for centralized control of one-warehouse, N-retailer inventory systems. *Manufacturing and Service Operations Management, 4*(1), 75–97.

Baker, K. R., Magazine, M. J., & Nuttle, H. L. (1986). The effect of commonality on safety stock in a simple inventory model. *Management Science, 32*, 982–988.

Balintfy, J. (1964). On a basic class of multi-item inventory problems. *Management Science, 10*, 287–297.

Berman, O., Krass, D., & Tajbaksh, M. M. (2012). A coordinated location-inventory model. *European Journal of Operational Research, 217*, 500–508.

Bertrand, L. P., & Bookbinder, J. H. (1998). Stock redistribution in two-echelon logistics systems. *Journal of the Operations Research Society, 49*, 966–975.

Bradford, J. W., & Sugrue, P. K. (1990). A Bayesian approach to the two-period style-goods inventory problem with single replenishment and heterogeneous Poisson demands. *The Journal of the Operational Research Society, 41*(3), 211–218.

Bramel, J., & Simchi-Levi, D. (1995). A location based heuristic for general routing problems. *Operations Research, 43*, 649–660.

Cachon, G. P. (1999). Managing supply chain demand variability with scheduled ordering policies. *Management Science, 45*(3), 843–856.

Cachon, G. P. (2001a). Exact evaluation of batch-ordering inventory policies in two-echelon supply chains with periodic review. *Operations Research, 49*(1), 79–98.

Cachon, G. (2001b). Managing a retailer's shelf space, inventory and transportation. *Manufacturing and Service Operations, 3*(3), 211–229.

Cachon, G. P., & Fisher, M. L. (2000). Supply chain inventory management and the value of shared information. *Management Science, 46*(8), 1032–1048.

Chang, P. L., & Lin, C. T. (1991). On the effects of centralization on expected costs in a multi-location newsboy problem. *Journal of Operational Research Society, 42*, 1025–1030.

Chen, F. (1998). Echelon reorder points, installation reorder points, and the value of centralized demand information. *Management Science, 44*(12), S221–S234.

Chen, M. S., & Lin, C. T. (1989). Effects of centralization on expected costs in a multi-location newsboy problem. *Journal of Operational Research Society., 40*, 597–602.

Chen, F., & Samroengraja, R. (2000). A staggered ordering policy for one-warehouse multi-retailer systems. *Operations Research, 48*(2), 281–293.

Chen, F., & Zheng, Y. S. (1994). Lower bounds for multi-echelon stochastic inventory systems. *Management Science, 40*(11), 1426–1443.

Chen, F., & Zheng, Y. S. (1997). One-warehouse multiretailer systems with centralized stock information. *Operations Research, 45*(2), 275–287.

Clark, A. J., & Scarf, H. (1960). Optimal policies for a multi-echelon inventory problem. *Management Science, 40*, 1426–1443.

Cohen, M. A., Agrawal, N., & Agrawal, V. (2006). Achieving breakthrough service delivery through dynamic asset deployment strategies. *Interfaces, 36*(3), 259–271.

Collier, D. A. (1982). Aggregate safety stock levels and component part commonality. *Management Science, 28*, 1296–1303.

de Kok, A. G. (1990). Hierarchical production planning for consumer goods. *European Journal of Operations Research, 45*, 55–69.

de Kok, A. G., Lagodimos, A. G., & Siedel, H. P. (1994). *Stock allocation in a two-echelon distribution network under service constraints*. Working Paper. Department of Industrial Engineering and Management Science, Eindhoven University of Technology, EUT 94-03.

Diks, E. B., & de Kok, A. G. (1998). Optimal control of a divergent multi-echelon inventory system. *European Journal of Operations Research, 111*, 75–97.

Dogru, M. K., de Kok, A. G., & van Houtum, G. J. (2013). Newsvendor characterizations for one-warehouse multi-retailer systems with discrete demand under the balance assumption. *Central European Journal of Operations Research, 21*, 541–559.

Dong, L., & Rudi, N. (2004). Who benefits from transshipment? Exogenous vs. endogenous wholesale prices. *Management Science, 50*(5), 645–657.

Eppen, G. D. (1979). Effect of centralization on expected costs in a multi-location newsboy problem. *Management Science, 25*, 498–501.

Eppen, G., & Schrage, L. (1981). Centralized ordering policies in a multi-warehouse system with lead times and random demands. In L. B. Schwarz (Ed.), *Multi-level production/inventory control systems: Theory and practice* (pp. 51–67). Amsterdam: North-Holland.

Erkip, N., Hausman, W., & Nahmias, S. (1990). Optimal centralized ordering policies in multi-echelon inventory systems with correlated demands. *Management Science, 36*(3), 381–392.

Federgruen, A. (1993). Centralized planning models for multi-echelon inventory systems under uncertainty. In A. H. G. Rinnooy Kan & P. H. Zipkin (Eds.), *Logistics of production and inventory. Handbooks in operations research and management science* (Vol. 4, chap. 3, pp. 133–173). Amsterdam: Elsevier.

Federgruen, A., Groenevelt, H., & Tijms, H. (1984). Coordinated replenishments in a multi-item inventory system with compound Poisson demands and constant lead times. *Management Science, 30*, 344–357.

Federgruen, A., & Zheng, Y. S. (1992). The join replenishment problem with general joint cost structures. *Operations Research, 40*, 384–403.

Federgruen, A., & Zipkin, P. H. (1984a). Approximation of dynamic, multi-location production and inventory problems. *Management Science, 30*, 69–84.

Federgruen, A., & Zipkin, P. (1984b). A combined vehicle routing and inventory allocation problem. *Operations Research, 32*(5), 1019–1037.

Fisher, M. L., & Rajaram, K. (2000). Accurate retail testing of fashion merchandise: Methodology and application. *Marketing Science, 19*(3), 266–278.

Gavirneni, S., Kapuscinski, R., & Tayur, S. (1999). Value of information in capacitated supply chains. *Management Science, 45*(1), 16–24.

Gerchak, Y., & Henig, M. (1986). An inventory model with component commonality. *Operations Research Letters, 5*, 157–160.

Gerchak, Y., Magazine, M. J., & Gamble, A. B. (1988). Component commonality with service level requirements. *Management Science, 34*, 753–760.

Glasserman, P., & Tayur, S. (1994). The stability of a capacitated, multi-echelon production-inventory system under a base-stock policy. *Operations Research, 42*(5), 913–925.

Graves, S. C. (1985). A multi-echelon inventory model for a repairable item with one-for-one replenishment. *Management Science, 31*(10), 1247–1256.

Graves, S. C. (1996). A multi echelon inventory model with fixed replenishment intervals. *Management Science, 42*(1), 1–18.

Ha, A. (1997). Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Science, 43*(8), 1093–1103.

Herer, Y. T., Tzur, M., & Yucesan, E. (2006). The multilocation transshipment problem. *IIE Transactions, 38*, 185–200.

Jackson, P. L. (1988). Stock allocation in a two-echelon inventory system or what to do until your ship comes in. *Management Science, 34*, 880–895.

Jackson, P., Maxwell, W., & Muckstadt, J. (1985). The joint replenishment problem with power-of-two intervals. *IIE Transactions, 17*, 25–32.

Jackson, P. L., & Muckstadt, J. A. (1989). Risk pooling in a two-period multi-echelon inventory stocking and allocation problem. *Naval Research Logistics, 36*, 1–26.

Jonsson, H., & Silver, E. A. (1987). Analysis of a two-echelon inventory control system with complete redistribution. *Management Science, 33*(2), 215–227.

Karmarkar, U. S., & Patel, N. R. (1977). The one-period, N-location distribution problem. *Naval Research Logistics Quarterly, 24*, 559–575.

Krishnan, K. S., & Rao, V. R. K. (1965). Inventory control in N warehouse. *Journal of Industrial Engineering, 16*, 212–215.

Lariviere, M. A., & Porteus, E. L. (1999). Stalking information: Bayesian inventory management with unobserved lost sales. *Management Science, 45*(3), 346–363.

Lee, H. L. (1987). A multi-echelon inventory model for repairable items with emergency lateral transshipments. *Management Science, 45*(5), 633–640.

Lee, H. L., Padmanabhan, P., & Whang, S. (1997a). Information distortion in a supply chain: The bullwhip effect. *Management Science, 43*, 546–558.

Lee, H. L., Padmanabhan, P., & Whang, S. (1997b). Bullwhip effect in a supply chain. *Sloan Management Review, 38*, 93–102.

Lee, H. L., So, C., & Tang, C. S. (2000). The value of information sharing in a two-level supply chain. *Management Science, 46*(5), 626–643.

Lippman, S. A., & McCardle, K. F. (1997). The competitive newsboy. *Operations Research, 45*(1), 54–65.

McGavin, E. J., Schwarz, L. B., & Ward, J. E. (1993). Two-interval inventory allocation policies in a one-warehouse N-identical retailer distribution system. *Management Science, 39*(9), 1092–1107.

Muckstadt, J. A. (2004). *Analysis and algorithms for service parts supply chains* (Springer series in operations research and financial). Berlin: Springer Verlag.

Nahmias, S., & Smith, S. A. (1993). Mathematical models of retailer inventory systems: A review. In R. K. Sarin (Ed.), *Perspectives in operations management* (pp. 249–278). MA: Kluwer Academic Publishers.

Nahmias, S., & Smith, S. A. (1994). Optimizing inventory levels in a two-echelon retailer system with partial lost sales. *Management Science, 40*(5), 582–596.

Pantumsinchai, P. (1992). A comparison of three joint ordering inventory policies. *Decision Science, 23*, 111–127.

Parlar, M. (1988). Game theoretic analysis of the substitutable product inventory problem with random demands. *Naval Research Logistics, 35*, 397–409.

Paterson, C. G., Kiesmuller, R., & Teunter, K. G. (2011). A comparison of three joint ordering inventory policies. *European Journal of Operational Research, 210*, 125–136.

Pryor, K., Kapuscinski, R., & White, C. (1999). *A single item inventory problem with multiple setup costs assigned to delivery vehicles*. Working paper, University of Michigan, Ann Arbor.

Rappold, J. A., & Muckstadt, J. A. (2000). A computationally efficient approach for determining inventory levels in a capacitated multi-echelon production-distribution system. *Naval Research Logistics, 47*, 377–398.

Reinman, M., Rubio, R., & Wein, L. (1999). Heavy traffic analysis of the dynamic stochastic inventory-routing problem. *Transportation Science, 33*(4), 361–380.

Robinson, L. W. (1990). Optimal and approximate policies in multi-period, multi-location inventory models with transshipments. *Operations Management, 38*(2), 278–295.

Rudi, N., Kapur, S., & Pyke, D. F. (2001). A two location inventory model with lateral shipment and local decision making. *Management Science, 47*(12), 1668–1680.

Sherbrooke, S. C. (1968). Metric: A multi-echelon technique for recoverable item control. *Operations Research, 16*(1), 122–141.

Silver, E. (1981). Establishing reorder points in the (S, c, s) coordinated control system under compound Poisson demand. *International Journal of Production Research, 19*, 743–750.

Speranza, M. G. W., & Ukovich, W. (1994). Minimizing transportation and inventory costs for several products on a single link. *Operations Research, 42*(5), 879–896.

Stulman, A. (1987). Benefits of centralized stocking for the multi-center newsboy problem with first-come-first-serve allocation. *Journal of Operational Research Society, 38*, 827–832.

Svoronos, A., & Zipkin, P. (1988). Estimating the performance of multi-level inventory systems. *Operations Research, 36*(1), 57–72.

Tagaras, G. (1989). Effects of pooling on the optimization and service levels of two-location inventory systems. *IIE Transactions, 21*(3), 250–257.

Tagaras, G., & Cohen, M. A. (1992). Pooling in two-location inventory systems with non-negligible replenishment lead times. *Management Science, 38*(8), 1067–1083.

Tagaras, G., & Vlachos, D. (2002). Effectiveness of stock transshipment under various demand distributions and nonneglibible transshipment times. *Production and Operations Management, 11*(2), 183–198.

Van Donselaar, K., & Wijngaard, J. (1987). Commonality and safety stocks. *Engineering Costs and Production Economics, 12*, 197–204.

Verrijdt, J. H. C. M., & de Kok, A. G. (1995). Distribution planning for a divergent N-echelon network without intermediate stock under service restrictions. *International Journal of Production Economics, 38*, 225–243.

Vishwanathan, S., & Mathur, K. (1997). Integrating routing and inventory decisions in a one-warehouse multi-retailer multi-product distribution system. *Management Science, 43*(3), 294–312.

Viswanathan, S. (1997). Periodic review (s, S) policies for joint replenishment inventory systems. *Management Science, 43*(10), 1447–1453.

Zhao, H., Deshpande, V., & Ryan, J. K. (2005). Inventory sharing and rationing in decentralized dealer networks. *Management Science, 51*(4), 531–547.

Chapter 13
Manufacturer-to-Retailer Versus Manufacturer-to-Consumer Rebates in a Supply Chain

Goker Aydin and Evan L. Porteus

1 Introduction

Rebates are widely used as promotional tools. In this paper we investigate the effects of two kinds of rebates (from the manufacturer) on supply chains: retailer rebates and consumer rebates. Retailer rebates, also known as channel rebates, are payments from the manufacturer to the retailer based on the sales performance of the retailer. Taylor (2002) cites several examples of the use of retailer rebates, in industries that range from software to printers, from network hardware switching to automotive. Consumer rebates, which are no less widespread than retailer rebates, are payments from the manufacturer to the consumer upon the consumer's purchase of the manufacturer's product. Most everybody is familiar through personal experience with the use of consumer rebates in consumer electronics, automotive and food products industries. The magnitude of rebate offers can reach surprisingly large numbers: A *New York Times* article reports that $10 billion worth of consumer rebates were offered in 2002 (Millman 2003).[1] Although some consumers do not claim their rebates (especially when the rebate size is small), the number of claims for consumer rebates is not negligible either: In 1998 Young America Inc. was reported to mail out 30 million rebate checks a year on behalf of companies like PepsiCo Inc., Nestle SA and OfficeMax (Bulkeley 1998). More recent statistics also suggest that the rebate

[1] In some cases, retailers themselves offer rebates to consumers. It is possible that the amount $10 billion quoted in the article includes the rebates offered by the retailers themselves.

G. Aydin (✉)
Kelley School of Business, Indiana University, 1309 East Tenth Street, Bloomington, IN 47405, USA
e-mail: ayding@indiana.edu

E.L. Porteus
Graduate School of Business, Stanford University, Stanford, CA, USA

N. Agrawal, S.A. Smith (eds.), *Retail Supply Chain Management*,
International Series in Operations Research & Management Science 223,
DOI 10.1007/978-1-4899-7562-1_13

activity remained strong in recent years. For example, according to a phone survey conducted by Consumer Reports National Research Center in 2009, 70 % of consumers reported having claimed a rebate within the past 12 months.[2] Similarly, high rebate activity was reported in a survey conducted by Parago, a firm that runs rebate and reward programs for its clients. The company's 2010 survey found that 47 % of consumers had submitted a rebate within the past 12 months.[3]

It is likely that rebates will remain on the scene as online shopping becomes more popular and smart phones start to play a larger role in consumers' purchases. In fact, online shopping enables instantly redeemable rebates, which are more attractive to customers. For example, Parago's shopper behavior study for 2013 found that 83 % of customers agree that "when shopping online, a discount via rebate is attractive." Similarly, 80 % of customers agreed that "the ability to submit a rebate via a smart phone is attractive," and 75 % of customers said that they wanted to scan a barcode in-store for rebates on their phone.[4]

For both retailer and consumer rebates, there do exist different implementations. Retailer rebates can be paid for each unit the retailer sells to the end customer or only for units sold in excess of a target number (Taylor 2002). Here we focus on the former type. In our model, the manufacturer uses consumer rebates for the sole purpose of selling more to the retailer. Thus, we do not address the role they may have early in a product's life cycle to learn more about demand or later to increase demand for unintended excess inventories. Consumer rebates can be in the form of mail-in rebates or coupons. Moreover, there are different kinds of coupons; some can be instantly redeemed at the time of purchase and some can be used only the next time a product is purchased. Of course, the specifics of the rebate offer have an influence on how attractive consumers find the rebate and how many customers will redeem the rebate. Here we use a stylized model of consumer rebates. We assume that (all) consumers treat a rebate of \$1 as being equivalent to a price discount of $\$\alpha$ and will redeem their rebates with probability β, where $0 < \alpha \leq 1$ and $0 < \beta \leq 1$. Thus, if a consumer rebate of x is offered on a product with price p, then the *effective retail price* is $p - \alpha x$ and if y customers buy the product, then the expected number of claims will be βy. Note that consumers are homogeneous in regard to the parameter α and we do not explicitly model a customer's decision of whether to claim a rebate or not. We shall see that modeling consumer rebates at this aggregate level allows us to identify the roles of the *claim rate* β and the *effective fraction* α in splitting the supply chain profit between the retailer and the manufacturer.

While the values of both α and β are likely to depend on many factors, we expect that they will be similar for products in a given category. For example, according to a survey of AC Nielsen's Homescan Consumer Panel, 27.7 % of households that

[2] Source: http://www.consumerreports.org/cro/magazine-archive/september-2009/personal-finance/rebates/overview/rebates-ov.htm. La28/28/2013.

[3] Source: http://www.parago.com/2011/02/11/parago-announces-surging-rebate-activity-in-2010/. La28/28/2013.

[4] Source: http://www.slideshare.net/TheresaWabler/letsmakeadeal-21181087. La28/28/2013.

reported buying computer products said mail-in rebates were very important when they bought PCs, monitors, printers and peripherals; 35.7 % said they were somewhat influenced by rebates (Ricadela and Koenig 1998). The same article reports, however, that consumers are less influenced by rebates when purchasing software. This example suggests that the value of α depends to a large degree on the product category. The claim rate, on the other hand, is likely to depend on the size of the rebate itself. For example, an educational software vendor reports that 8–10 % of its customers claim $10 rebates, and the claim rate increases to 20 % for $20 rebates (Bulkeley 1998). Likewise, according to an estimate reported in the *US News & World Report*, "for pricey items with rebates worth $50, the redemption rate is below 50 %. On smaller items with rebates under $10, redemption rates are likely to be in the single digits" (Palmer 2008). Nevertheless, the rebate sizes tend to be similar within a product category and, hence, the product category seems to be a more important determinant of the claim rate than the size of the rebate. For example, in contrast to the software vendor who faced claim rates in the 10–20 % range, the now-defunct PC seller eMachines had a mail-in rebate program, which had seen a 70–90 % claim rate prior to its cancellation (Olenick 2002). In the case of new automotive purchases, where the rebates are even larger, the usual practice is for the rebate to be instantaneously redeemable at the time of purchase, which suggests that $\alpha = \beta = 1$. In summary, while consumer response to rebate offers may vary in the size of the rebate, much of this variation may be accounted for by the product category.

In order to compare and contrast the effects of the two rebate types on the supply chain, we consider a single-retailer, single-manufacturer supply chain selling a single product, and we analyze the equilibrium outcome under each rebate policy. (The decision of what rebate type to use is not endogenous to our model; instead, we analyze and compare the equilibria under each rebate type.) We assume that the wholesale price for the product is exogenously fixed. This assumption is mainly for tractability, but it is also an approximation of an environment where rebate offers constitute a further stage of decision making in a supply chain with a well-established wholesale price. The consumer demand for the product is stochastic and depends on the effective retail price. In the case of a retailer rebate, the effective retail price is simply the retail price, whereas in the case of a consumer rebate, the effective retail price is the retail price minus the effective fraction of the consumer rebate. We assume that the expected demand for the product is a function of the effective retail price, and the realized demand is a multiplicative random perturbation of that expected demand. The assumption of a multiplicative model is not without consequence; it implies that the coefficient of variation of demand is constant with respect to price.

Under either rebate policy, before the start of the single-period selling season, the retailer must determine the retail price, and the manufacturer needs to choose the size of the rebate (or rebates, if both rebate types are used in the supply chain) simultaneously. This simultaneous determination of the rebates and the retail price can be seen as approximating a negotiation process between the manufacturer and

the retailer in setting the terms of a rebate offer. Once the price and rebate(s) are announced, the retailer decides how many units of the product to purchase. The manufacturer builds that amount and delivers it to the retailer by the beginning of the selling season. At the end of the selling season, all unmet demands become lost sales, and leftover inventory is salvaged. This model would be particularly applicable to high-tech products where the short life cycle of the product can be modeled as covering a single season with a single ordering and pricing opportunity. The more replenishments take place during the life cycle of the product and the more price adjustments made, the more approximate our model becomes.

Of course, both retailer and consumer rebates provide the retailer with an incentive to stock more. However, the two rebates differ in how they achieve this result: Retailer rebates do so by increasing the retailer's margin on every unit sold, whereas consumer rebates do so by boosting the demand for the product. We find that, as expected (in equilibrium), when retailer rebates are present, the retailer will reduce the retail price (by an amount less than the rebate itself) to increase the sales volume of an item and collect a larger sum from the manufacturer in rebates, thereby passing on to the consumer some of the benefits it receives. On the other hand, a consumer rebate will induce the retailer to increase the retail price (by an amount less than the effective rebate) to take advantage of the boost in demand that arises from a consumer rebate, thereby sharing in some of the benefits offered to consumers. We show that the total supply chain profit always improves under retailer rebates, compared to no rebates. The same is true for consumer rebates, provided that the effective fraction (α) is larger than the claim rate (β). However, if $\alpha < \beta$, then total supply chain profit may suffer. We provide numerical examples to demonstrate that neither the retailer nor the manufacturer always prefers one particular kind of rebate to the other. In addition, our numerical examples suggest that, contrary to popular belief, it is possible for both firms to prefer consumer rebates even when all such rebates are redeemed.

In comparing the two rebate types, we find that the split of supply chain profits under consumer rebates depends critically on α and β. In particular, we obtain the following results:

- Under the consumer rebate equilibrium, the retailer's share of the supply chain profit will be $\frac{\alpha}{\alpha+\beta}$, and the manufacturer's $\frac{\beta}{\alpha+\beta}$. In other words, the profit will be divided so that the ratio of the retailer profit to the manufacturer profit will be α/β.
- The higher α is with respect to β (i.e., the higher consumers value the rebate relative to the rate at which consumers redeem them), the more attractive the consumer rebate becomes from the overall supply chain's perspective. Therefore, one can conclude that, everything else being equal, the more attractive the consumer rebate from the overall supply chain's perspective, the larger the retailer's share of the supply chain profit will be in equilibrium.
- Note that the retailer's share is increasing in α and decreasing in β, and the opposite is true for the manufacturer. Nevertheless, as we demonstrate through a numerical example, this does not mean that the retailer and the manufacturer are

at odds in terms of what α and β they prefer. It turns out that, under a consumer rebate equilibrium, both firms can prefer α to be larger and β to be smaller; even though the manufacturer's share of supply chain profits is smaller, the manufacturer gets more, because the increase in the supply chain profits more than compensates for the decrease in the share it gets.

In the next section, we review the related literature and compare our model to those in earlier research. Section 3 describes our model and discusses our results for the case where both rebate types are used simultaneously. In Sects. 4 and 5, we discuss our results when retailer rebates and consumer rebates are used in isolation. We provide a number of numerical examples in Sect. 6 to demonstrate some interesting equilibrium outcomes. We conclude in Sect. 7. All proofs are provided in the appendix.

2 Literature Review

The marketing and economics literature has investigated the use of consumer rebates. For example, Gerstner and Hess (1991, 1995) use a demand model where the consumer population consists of two segments; the size and reservation price of each segment is deterministic and known. The higher-end segment has a cost associated with redeeming a consumer rebate, reflecting the higher disutility price-insensitive customers have for claiming rebates. The supply chain is assumed to be serving only the higher-end segment in status quo. They examine how retailer rebates (called push price promotions) and consumer rebates (called pull price promotions) can be used to induce the retailer to serve the lower-end segment as well as the higher-end one, and how such promotions affect manufacturer and supply chain profits. Narasimhan (1984) offers a price discrimination argument to explain the use of consumer rebates. He considers a model where the firm offering the rebate is selling directly to the end consumer. In his model, a consumer need not redeem a rebate every time she purchases a product. He models the consumer's decision of how many rebates to use as a utility maximization problem, and shows that the more price-sensitive a customer, the more she engages in consumer rebates. Therefore, rebates result in the firm selling at a lower price to consumers who are more price sensitive. In this sense, the consumer rebate acts as a price discrimination device. Our model is less general than this stream of research because we do not model how individual consumers respond differently to rebate offers. Instead, we model the effect of rebates at the aggregate demand level, through the effective fraction parameter α and the claim rate parameter β. Our model is more general in the sense that we incorporate demand uncertainty and retail stock level decisions.

There is also a stream of research in marketing that considers the use of trade promotions; i.e., a discount in wholesale price offered by the manufacturer in order to induce the retailer to lower the retail price. Since the typical assumption of this research stream is that all demand is met (i.e., sales equals demand), such a discount

in wholesale price is equivalent to a retailer rebate. Most of the model-based work in this research stream involves multiple competing manufacturers, and the emergence of trade promotions is explained through the equilibrium of the game among these multiple manufacturers. In this setting, the manufacturer is assumed to be selling directly to the end consumers, and the role of the retailer is ignored. See, for example, Raju et al. (1990), Lal (1990) and Rao (1991). Our model has only a single manufacturer, but we add explicit consideration of a retailer, demand uncertainty, and the retailer's decision of the stock level.

There is another marketing research stream on trade promotions that considers manufacturers selling through a retailer. For example, Lal et al. (1996) consider an infinite horizon model where two identical manufacturers sell through a single retailer. Their customer population consists of three customers: one switcher and two loyals. In this model, trade promotions exist because the manufacturers compete for the switcher. Dreze and Bell (2003) consider a single-retailer, single-manufacturer setting where customer demand is a deterministic function of price. They compare the effects of two different contractual arrangements for trade promotions: off-invoice deals that correspond to a wholesale price discount and scan-back deals that correspond to retailer rebates. In this model, even though demand is deterministic, the retailer may choose to carry inventories to take advantage of a temporary promotional offer from the manufacturer. In our model, the reason a retailer chooses to carry inventories is due to demand uncertainty. We also emphasize how the rebates affect supply chain profits and the shares that the two firms get.

There is earlier work in the operations management literature that considers the role played by retailer rebates in the presence of operational concerns like inventory costs. Taylor (2002) considers retailer rebates in a model where demand is stochastic, but the retail price is exogenously given. He shows that retailer rebates paid for units sold beyond a target level can be used to achieve supply chain coordination. He also analyzes a model where the retailer can exert sales effort to influence demand. In this case, retailer rebates can still achieve coordination, but a returns policy should also be implemented. Using a more general model, Krishnan et al. (2004) focus on the use of retailer rebates in the presence of retailer efforts. Their main focus is finding coordinating contracts. Unlike these two, we do not model the retailer's sales effort; however, we consider a model with price-dependent stochastic demand, and retail price is endogenous to our model in that the retailer decides what price to charge. We do not seek to establish channel-coordinating mechanisms, but we do show that retailer rebates improve supply chain profits. We also compare the supply chain profit under retailer rebates with that under consumer rebates.

There is an extensive operations management literature on the price setting newsvendor problem, in which a retailer faces a single-period inventory and pricing problem with stochastic, price-dependent demand. See, for example, Petruzzi and Dada (1999) for a review with extensions. Our analysis benefits from Petruzzi and Dada (1999); in particular, Lemma 4(a) in the appendix is due to them. In their multiplicative model, they assume that demand is given by $ap^{-b}\epsilon$, where ϵ is a

random variable. In this demand model, the price elasticity of expected demand is constant. Our assumptions do not cover this specific model, but we do allow the (absolute) price elasticity of expected demand to be increasing in price, thereby complementing some of the existing structural results on the price setting newsvendor problem. Kalyanam (1996) finds empirical support for both constant and increasing price elasticity of demand. In this chapter, we use an inverse demand representation to write the retailer's and manufacturer's expected profit functions, which facilitates our analysis. (See the next section.) Aydin and Porteus (2008) study an inventory and pricing problem where a retailer sets the prices and inventory levels for an assortment of substitutable products, and they take advantage of a similar representation.

A closely related paper is by Chen et al. (2007), who consider the question of consumer rebates from an operations management perspective. As in our model, they consider a single-retailer, single-manufacturer supply chain where one-shot inventory and pricing decisions are made to satisfy price-dependent uncertain customer demand. Their consumer rebate is an exogenously fixed fraction of the wholesale price and the decision making is sequential: the manufacturer chooses the wholesale price first, and the retailer chooses the retail price second. Our wholesale price is exogenous but our consumer rebate is a decision variable. We add consideration of retailer rebates and our assumptions allow us to show how the claim rate and the effective fraction parameters affect the split of supply chain profits between the retailer and the manufacturer.

Since the initial publication of this chapter, several further contributions have been made to the literature on the role of consumer rebates in supply chain management. In a set of recent papers, Demirag and colleagues also compare two avenues available to manufacturers: offering rebates to consumers or offering incentives to retailers. Demirag et al. (2010) compare manufacturer-to-consumer rebates with manufacturer-to-retailer incentives, which take the form of a lump sum payment (in contrast to the manufacturer-to-retailer rebate in our model, which is a per-unit payment). The retailer uses this incentive to offer discounts to select customers, thus effectively achieving price discrimination among customers. By studying several scenarios (including both stochastic and deterministic demand models), the paper investigates which of the two schemes the manufacturer prefers. Demirag et al. (2011b) extend this work to the case with two manufacturers and two competing retailers. In a different vein, Demirag et al. (2011a) start with a model similar to ours, but they assume that the retailer is risk averse. They show that the rebate scheme preferred by the manufacturer does depend on the degree of retailer's risk aversion.

Another set of recent papers focuses on rebates paid to consumers only (whereas we study rebates paid to the retailer as well), but they allow consumer rebates to come from either the manufacturer or the retailer (whereas we allow consumer rebates to come from the manufacturer only). These papers model interactions in a two-stage supply chain using the Stackelberg equilibrium, where the manufacturer moves first, followed by the retailer. Cho et al. (2009) use a deterministic demand model, and they pay special attention to how the equilibrium depends on the fixed cost of adopting a

rebate initiative. Arcelus et al. (2012) and Geng and Mallik (2011) adopt a newsvendor setting to compare retailer-driven versus manufacturer-driven rebates. Both allow the redemption rate to be a function of the rebate size—this is a dependence we do not model. Arcelus et al. (2012) treat the wholesale price as endogenous, and they find conditions under which it is best for only the retailer to offer the rebate. Geng and Mallik (2011) treat the wholesale price as exogenous, and they show that the average effective price paid by consumers is higher in the presence of rebates.

Focusing on manufacturer-to-consumer rebates only, a few recent papers study how rebates play out in the presence of supply chain initiatives that restrict the retail price. For instance, Yang et al. (2010) study how manufacturer-suggested retail prices (MSRP) interact with rebates. In a similar vein, Khouja and Zhou (2010) study a supply chain where the manufacturer implements incentives that curb the retail price. They use a model where consumers are heterogeneous in the value they derive from a rebate. Their main result is that rebates are good for the supply chain as a whole, owing to the limits on retail price.

3 Consumer and Retailer Rebates Together

In this section, we describe our model when the manufacturer uses both retailer and consumer rebates, and we derive some preliminary results. The use of both rebates at the same time is quite common in the automotive industry, where retailer rebates are usually called dealer incentives and the consumer rebates are offered in the form of cashback allowances. In the following sections, we will focus on the cases where each rebate type is used in isolation, and the results developed in this section will apply to those special cases. Let r_R denote the retailer rebate and r_C the consumer rebate, each paid to their respective recipients for every unit the customer buys. Also, let p be the retail price of the product.

Let us first describe the demand model. First, the higher the consumer rebate the larger the stochastic demand will be. Therefore, the demand should be a function of r_C as well as p. Let $D(p, r_C)$ denote the stochastic demand for the product. We assume that consumers treat a \$1 rebate as the equivalent of an $\$\alpha$ price discount; i.e., consumers act as if the unit retail price they are paying is $p - \alpha r_C$. We will impose the following assumptions on the demand model:

(A1) $D(p, r_C) = f(p - \alpha r_C)\epsilon$,
(A2) ϵ is a strictly positive random variable with a strictly increasing failure rate (IFR),
(A3) $f(\cdot)$ is strictly decreasing, and $f(x) \to 0$ as $x \to \infty$, and
(A4) $\frac{f'(\cdot)}{f(\cdot)}$ is non-increasing.

The first assumption implies that the expected demand is a function of the retail price minus the effective consumer rebate (i.e., the price after rebate). (*A*1) and (*A*2) implicitly assume that ϵ is independent of price and any rebate. Thus, (*A*1) implies

that the coefficient of variation of demand for the product does not change with price. The requirement in (*A*2) that ϵ be IFR is not very restrictive as many probability density functions, including the normal and Weibull with shape parameter greater than one, satisfy this assumption. (For more on IFR distributions, see Barlow and Proschan 1965.) (*A*3) is a natural assumption that means the expected demand is decreasing in price. This assumption is violated only for very few luxury items. (*A*4) implies that the magnitude of the expected demand's elasticity to price is increasing in p; i.e., as price gets larger the percentage change in demand in response to a percentage change in price gets larger. (*A*4) is satisfied by many commonly used forms of price dependency. For example, it is easy to check that (*A*4) will be satisfied when expected demand is exponentially decreasing in price; i.e., $f(x) = e^{-ap}$, or when expected demand is linearly decreasing in price; i.e., $f(x) = a - bx$, or when expected demand is given by the logit demand model; i.e., $f(x) = \dfrac{exp(u_1 - x)}{exp(u_0) + exp(u_1 - x)}$.

We define the following notation:

w : unit wholesale price charged by the manufacturer

c : unit production cost

v : unit salvage value

$\Phi(x, p - \alpha r_C)$: cumulative distribution function (cdf) of $D(p, r_C)$

$\phi(x, p - \alpha r_C)$: probability density function (pdf) of $D(p, r_C)$

$\Phi_\epsilon(\cdot)$: cdf of ϵ

$\phi_\epsilon(\cdot)$: pdf of ϵ

We assume that $\Phi(x, p - \alpha r_C)$ is twice-continuously differentiable in both its arguments. Throughout the remainder of the paper, given a function g of vector x, we use $\nabla_i g(\tilde{x})$ to denote the partial derivative of $g(x)$ with respect to the ith component of x evaluated at $x = \tilde{x}$. Similarly, $\nabla^2_{ij} g(\tilde{x})$ and $\nabla^2_{ii} g(\tilde{x})$ denote the cross-partial and second partial of $g(x)$ at $\tilde{x}$, respectively.

Before the selling season starts, the retailer determines p, and, simultaneously, the manufacturer chooses r_R and r_C. The assumption of simultaneous decision making implies that one party in the supply chain is not particularly more powerful than the other, so one party cannot impose its respective decision on the other. We assume that all the parameters and distributions are known by both the retailer and the manufacturer.

Once the price and rebates are announced, the retailer chooses the stock level and the manufacturer then builds that amount, which is delivered to the retailer by the beginning of the selling season. After the selling season is over, the retailer will salvage the leftover inventory at unit salvage value of v. We assume that $w > c > v$ for the problem to make economic sense. In the presence of retailer rebates, there is the possibility that the retailer could misreport the amount of sales to collect larger rebates from the manufacturer. For example, the retailer could dump all the leftover

inventory and claim that it had been sold. While the existence of a salvage value alleviates this moral hazard problem, a complete avoidance of such misreporting of sales requires some form of possibly costly monitoring of retail sales. We return to this issue in Sect. 7.

The retailer's profit function is given by

$$\Pi_R(p,y,r_C,r_R) = (p+r_R)\left[\int_0^y x\phi(x,p-\alpha r_C)dx + y(1-\Phi(y,p-\alpha r_C))\right] + v\int_0^y (y-x)\phi(x,p-\alpha r_C)dx - wy. \tag{13.1}$$

Note that the optimal stock level for the product, $y^*(p,r_C,r_R)$, is given for each given retail price p, wholesale price w and rebates r_R and r_C as the critical fractile solution:

$$\Phi(y^*(p,r_C,r_R),p-\alpha r_C) = \frac{p+r_R-w}{p+r_R-v} \tag{13.2}$$

It is important to note how the two different kinds of rebates affect $y^*(p,r_C,r_R)$: the stock level chosen depends on the retailer rebate since the critical fractile itself is a function of the retailer rebate, whereas the consumer rebate affects the stock level through its impact on the demand distribution.

The retailer's profit function can be rewritten as the following induced profit function, obtained by substituting for $\Phi(y^*(p,r_C,r_R),p-\alpha r_C)$ in (13.1) (see, for example, Porteus 2002):

$$\Pi_R(p,r_C,r_R) = (p+r_R-v)\int_0^{y^*(p,r_C,r_R)} x\phi(x,p-\alpha r_C)dx, \tag{13.3}$$

where $y^*(p,r_C,r_R)$ is as defined by (13.2). Define the inverse demand function $z(p,r_C,\xi)$ as

$$\Phi(z(p,r_C,\xi),p-\alpha r_C) = \xi. \tag{13.4}$$

With this definition, $z(p,r_C,\xi)$ is the demand that corresponds to the ξ fractile of Φ, given the retail price p and consumer rebate r_C. We use this representation as it provides a more convenient way of dealing with the pricing problems to be solved. Using the inverse demand function, we can rewrite (13.2) as $y^*(p,r_C,r_R) = z(p,r_C,\frac{p+r_R-w}{p+r_R-v})$. Also, we can rewrite the retailer's induced profit function in (13.3) as

$$\Pi_R(p, r_C, r_R) = (p + r_R - v) \int_0^{\frac{p+r_R-w}{p+r_R-v}} z(p, r_C, \xi) d\xi. \tag{13.5}$$

The following proposition states our structural result on $\Pi_R(p, r_C, r_R)$.

Proposition 1 *Suppose (A1) through (A4) hold. Then, given r_C and r_R, there is a unique $p > w - r_R$ that optimizes the retailer's profit, and this unique p satisfies the first order condition (FOC) for $\Pi_R(p, r_C, r_R)$.*

The manufacturer's profit function is given by

$$\Pi_M(p, r_C, r_R) = (w - c) y^*(p, r_C, r_R) - (\beta r_C + r_R) \left[\int_0^{\frac{p+r_R-w}{p+r_R-v}} z(p, r_C, \xi) d\xi + \frac{w - v}{p + r_R - v} y^*(p, r_C, r_R) \right]. \tag{13.6}$$

The first term in (13.6) is the profit margin of the manufacturer multiplied by the number of units ordered by the retailer. The term in brackets is the expected sales. Note that the rebate the manufacturer pays per unit sold is the retailer rebate r_R, plus a fraction β of the consumer rebate r_C (since a fraction β of consumers claim their rebate). Therefore, the expected total rebate payment made by the manufacturer is $\beta\, r_C + r_R$ multiplied by the expected sales. The following proposition states some structural results on $\Pi_M(p, r_C, r_R)$:

Proposition 2 *Suppose (A1) through (A4) hold. Then, given p:*

(a) *Suppose r_R is fixed so that $p + r_R > v$. Then, either the manufacturer's profit is optimized at $r_C = 0$, or there exists a unique r_C that satisfies the FOC for $\Pi_M(p, r_C, r_R)$ and such r_C optimizes the manufacturer's profit.*

(b) *Suppose r_C and p are fixed. Then, either the manufacturer's profit is optimized at $r_R = 0$, or there exists a unique r_R that satisfies the FOC for $\Pi_M(p, r_C, r_R)$ and such r_R optimizes the manufacturer's profit.*

(c) *At any r_C and r_R such that $\nabla_2 \Pi_M(p, r_C, r_R) = \nabla_3 \Pi_M(p, r_C, r_R) = 0$, we have $\nabla_2 y^*(p, r_C, r_R)/\beta > \nabla_3 y^*(p, r_C, r_R)$.*

Parts (a) and (b) of the proposition establish that the manufacturer's profit is well-behaved in the rebates. We cannot rule out the possibility that the manufacturer's profit will be decreasing in the retailer or the consumer rebate. Therefore, the manufacturer's optimal solution may involve a zero rebate. To understand part (c), note that increasing the retailer rebate by \$1 costs the manufacturer \$1 for every unit sold to consumers, while increasing the consumer rebate by \$1 costs only \$$\beta$ for every unit sold to consumers. Thus, part (c) says that, given a pair of rebates that is a candidate for the manufacturer's optimal solution, the marginal increase in units

sold to the retailer, per manufacturer's effective (at-risk) cost of a rebate-dollar, is higher for consumer rebates than retailer rebates.

Let $\Pi_{SC}(p, r_C, r_R)$ be the profit of the supply chain for a given retail price p, consumer rebate r_C and retailer rebate r_R. Note that $\Pi_{SC}(p, r_C, r_R) = \Pi_R(p, r_C, r_R) + \Pi_M(p, r_C, r_R)$ where $\Pi_R(p, r_C, r_R)$ and $\Pi_M(p, r_C, r_R)$ are as defined by (13.5) and (13.6), respectively. The following proposition states how the supply chain profit will be split between the two parties under an equilibrium solution.

Proposition 3 *Suppose (A1) through (A4) hold. Furthermore, suppose that a pure-strategy Nash equilibrium exists for the game between the retailer and the manufacturer. Let* $\tilde{p}$ *be an equilibrium retail price, and* $\tilde{r}_R$ *and* $\tilde{r}_C$ *the corresponding equilibrium rebates. The stock level that arises under this equilibrium is given by* $y^*(\tilde{p}, \tilde{r}_C, \tilde{r}_R)$ *where* y^* *is given in* (13.2) *. Under this equilibrium, if* $\tilde{r}_C > 0$*, then* $\frac{\Pi_R(\tilde{p}, \tilde{r}_C, \tilde{r}_R)}{\Pi_M(\tilde{p}, \tilde{r}_C, \tilde{r}_R)} = \frac{\alpha}{\beta}$.

As we will see later on, this particular division of the supply chain profit under an equilibrium solution is due to the use of the consumer rebate, and, as stated in the proposition, will be true whenever the equilibrium consumer rebate is (strictly) positive. The key assumption that leads to this interesting result is that the demand uncertainty is multiplicative. We will discuss the rationale behind this result in detail when we discuss the use of consumer rebates in isolation. Also, this constant-split property allows us to conclude that, even when multiple Nash equilibria (with strictly positive consumer rebates) exist, there is one equilibrium that is preferred by both parties to all other equilibria, and the equilibrium preferred by both parties is the one under which the supply chain profit is at its highest among all other equilibria. If one could argue that our model captured the first order issues addressed in the automotive industry, where it is plausible to assume that both α and β are equal to one (due to the large sums involved in cashback allowances), one could say that the rebates would lead to dividing the channel profits evenly between the manufacturers and the dealers.

In the next two sections, we will consider the cases that arise when either only retailer rebates or only consumer rebates are used.

4 Retailer Rebate Only

The *retailer rebate game* is the game between the retailer and the manufacturer in the previous section with the restriction that $r_C = 0$. We will continue to use the same notation as before, replacing r_C with zero where necessary. The structural results on the profit functions of the manufacturer and the retailer (adapted for $r_C = 0$) will carry over directly from the previous section. In addition, the following proposition states how the optimal decision of one player changes with the decision of the other one.

Proposition 4 *Suppose (A1) through (A4) hold and* $r_C = 0$. *Let* $p^*(r_R)$ *be the optimal price chosen by the retailer as a response to a given* r_R *and* $r_R^*(p)$ *the optimal retailer rebate chosen by the manufacturer as a response to a given* p. *Then:*

(a) $-1 \leq \dfrac{dp^*(r_R)}{dr_R} < 0.$

(b) $-1 < \dfrac{dr_R^*(p)}{dp} \leq 0.$

(c) *There exists a unique Nash equilibrium for the retailer rebate game.*

The first part of the proposition above implies that when the manufacturer offers an additional \$1 rebate to the retailer for every unit sold, the retailer will decrease the selling price of the product, but the price discount will be less than \$1. Therefore, the retailer rebate results in some savings being passed on to the customer. Likewise, when the retailer reduces the price of the product by \$1, the manufacturer will increase the rebate paid to the retailer, but by less than \$1. The following proposition summarizes our results in this setting.

Proposition 5 *Suppose (A1) through (A4) hold and* $r_C = 0$. *Let* p_o *be the retail price and* y_o *the stock level chosen by the retailer when* $r_R = 0$. *Let* $\tilde{p}$ *be the equilibrium retail price, and* $\tilde{r}_R$ *the equilibrium rebate that will arise under the retailer rebate game. The stock level that arises under this equilibrium is given by* $y^*(\tilde{p}, 0, \tilde{r}_R)$ *where* y^* *is as defined by* (13.2) . *Then:*

(a) $p_o - \tilde{r}_R \leq \tilde{p} \leq p_o$,

(b) $y_o \leq y^*(\tilde{p}, 0, \tilde{r}_R)$ *and*

(c) *If* $0 < \tilde{r}_R \leq w - c$, *then* $\Pi_{SC}(\tilde{p}, 0, \tilde{r}_R) > \Pi_{SC}(p_o, 0, 0)$.

The first two parts of the proposition state that, as expected, the retail price will decrease and the stock level will increase when retailer rebates are used. We should note that, in parts (a) and (b) of the proposition, the inequalities are not strict, since the equilibrium may turn out to be the no-rebate case; i.e., $\tilde{r}_R$ may be zero. It is interesting to note here how the role played by retailer rebates under endogenous retail pricing differs from that under an exogenously-fixed retail price. When the retail price is exogenous, the rebate helps the manufacturer by increasing the retailer's margin on every unit sold, thereby increasing the quantity ordered by the retailer. On the other hand, when the retail price is endogenous, the rebate serves a dual purpose for the manufacturer: As before, the rebate increases the order quantity of the retailer by increasing the retailer's margin on every unit sold, but, in addition, the rebate causes a decrease in the retail price (as stated in part (a) of the proposition), thereby increasing the customer demand, which causes a further increase in retailer's order quantity.

The last part of the proposition states that if the equilibrium rebate is (strictly) positive and below the manufacturer's unit profit margin (which would be expected to be the case in practice), then the supply chain will be strictly better off as a result of the use of the retailer rebate. This result is not surprising. Intuitively speaking,

the higher the retailer rebate, the closer the supply chain becomes to one that is owned by a single decision maker, since increasing the retailer rebate brings the retailer's underage cost closer to the integrated supply chain's underage cost. Therefore, the higher the retailer rebate, the closer the performance of the supply chain becomes to that of the integrated one. We should note that the constant-split property does not hold when only retailer rebates are used.

Next, we discuss the case in which only consumer rebates are used.

5 Consumer Rebate Only

The *consumer rebate game* is the game between the retailer and the manufacturer in Sect. 3 with the restriction that $r_R = 0$. The structural results on the retailer's and manufacturer's profit functions stated in Sect. 3 (adapted for $r_R = 0$) carry over. The following proposition states how the optimal price chosen by the retailer responds to a change in the consumer rebate.

Proposition 6 *Suppose (A1) through (A4) hold and* $r_R = 0$. *Let* $p^*(r_C)$ *be the optimal price chosen by the retailer as a response to a given* r_C. *Then:*

(a) $0 < \dfrac{dp^*(r_C)}{dr_C} < \alpha.$

(b) *There exists a Nash equilibrium for the consumer rebate game.*

The proposition states that when the manufacturer offers an additional \$1 rebate to the consumer, the retailer will take advantage of this offer, and will increase the retail price, but the increase will be less than α. This means that, as is commonly thought, a consumer rebate will bring about a price increase, however the effective retail price paid by the consumer will still be less than the price that would be paid if the rebate did not exist. Unfortunately, a result on how the optimal consumer rebate responds to price eludes us. In the absence of such a result, we are not able to claim that the Nash equilibrium under the consumer rebate game will be unique. The following proposition summarizes our results for this game.

Proposition 7 *Suppose (A1) through (A4) hold and* $r_R = 0$. *Let* p_o *denote the price and* y_o *the stock level chosen by the retailer when* $r_C = 0$. *Let* $\tilde{p}$ *be an equilibrium retail price under the consumer rebate game, and* $\tilde{r}_C$ *the corresponding equilibrium rebate. Suppose that* $\tilde{r}_C > 0$. *The stock level that arises under this equilibrium is given by* $y^*(\tilde{p}, \tilde{r}_C, 0)$ *where* y^* *is as defined by* (13.2) . *Then:*

(a) $p_o \le \tilde{p} \le p_o + \alpha \tilde{r}_C$,

(b) $y_o \le y^*(\tilde{p}, \tilde{r}_C, 0)$,

(c) *If* $\alpha \ge \beta$ *and* $\tilde{r}_C \le w - c$, *then* $\Pi_{SC}(\tilde{p}, \tilde{r}_C, 0) \ge \Pi_{SC}(p_o, 0, 0)$ *and*

(d) $\dfrac{\Pi_R(\tilde{p}, \tilde{r}_C, 0)}{\Pi_M(\tilde{p}, \tilde{r}_C, 0)} = \dfrac{\alpha}{\beta}$

The first two parts of the proposition state the intuitive results that the retail price and the stock level will increase when consumer rebates are used. However, the increase in retail price will not be larger than the effective fraction of the consumer rebate, so consumers are still better off as a result of the rebate. The third part of the proposition states that, if α is larger than β, the supply chain profit will improve as a result of the consumer rebate (provided that the rebate is less than the manufacturer's profit margin, which we would expect to be the case). This result is expected: Essentially, when the cost of a \$1 rebate, modeled by β, is less than the effective fraction of the \$1 rebate, modeled by α, the supply chain is able to achieve the demand impact of an α-dollar price discount at a cost of $\beta < \alpha$ dollars. Also, we see from the last part of the proposition that the constant-split property of supply chain profit continues to hold when consumer rebates are used in isolation. Due to this constant-split property, we conclude that, even when multiple Nash equilibria exist, the equilibrium under which the supply chain profit is at its highest (among all other equilibria) is the one preferred by both parties. Furthermore, the constant-split property shows that, in an equilibrium solution, neither party is able to extract the entire supply chain profits. (Unless α or β is zero, which are not likely to be the case. Here, we assume that both α and β are strictly positive, and we do not cover the cases that arise when one or the other is zero.)

An interesting consequence of the constant-split property is that if the retailer's share of the supply chain profit under consumer rebates is larger than the manufacturer's, then it must be that $\alpha > \beta$ for the product in question, and, hence, by Proposition 7(c), the use of consumer rebates must have improved total supply chain profits.

From part (d) of Proposition 7, we observe that the manufacturer's share of the supply chain profit under consumer rebate equilibrium is $\frac{\beta}{\alpha+\beta}$. However, this observation does not imply that the manufacturer would necessarily like to design rebates so that β is high or α is low. In fact, in many numerical examples, we observed the opposite to be true. One such example is depicted in Fig. 13.1. In this example, with β fixed at 0.9, the manufacturer prefers a large α to a small one, since the manufacturer prefers getting a smaller share of the large supply chain profit

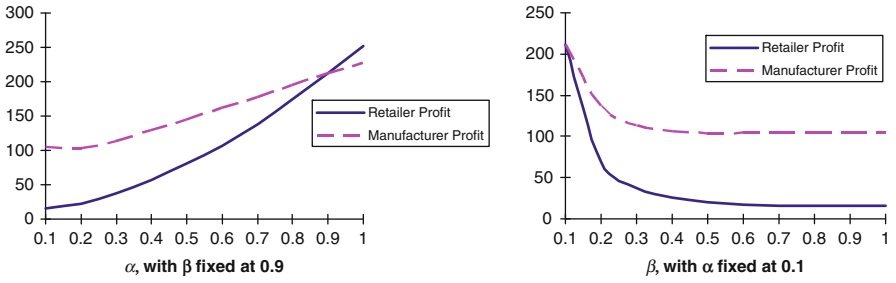


Fig. 13.1 Equilibrium retailer and manufacturer profits as a function of α (*left*) and β (*right*)

achieved under a large α value. Likewise, with α fixed at 0.1, the manufacturer prefers a small β to a large one. Note that the manufacturer's profit is not necessarily monotonic in α or β, which can be confirmed with careful scrutiny of the graphs. Another conclusion that applies to this example is that there is no conflict between the retailer and the manufacturer in terms of the attributes of a rebate: To the extent possible, both parties would like a rebate with a high customer valuation α and a small redemption rate β. We observed this to be the case in many other numerical examples. We return to this point in Sect. 7.

It is worthwhile to discuss the rationale behind the constant-split property. We will do so through a marginal analysis discussion. For the sake of the following discussion, define $\gamma(p) := -\frac{f'(p)}{f(p)}$; i.e., $\gamma(p)$ is a positive number representing the fractional decrease in expected demand in response to a marginal increase in price p. Under the consumer rebate equilibrium, the retail price must satisfy the FOC for the retailer. Hence, by part (a) of Lemma 6, the retail price p must satisfy

$$\int_0^{\frac{p+r_R-w}{p+r_R-v}} z(p, r_C, \xi)d\xi + \frac{w-v}{p+r_R-v} y^*(p, r_C, r_R) = \gamma(p+r_R-v) \int_0^{\frac{p+r_R-w}{p+r_R-v}} z(p, r_C, \xi)d\xi \tag{13.7}$$

The left-hand side of (13.7) is the expected sales of the product; a \$1 price increase means the retailer will make \$1 more on every unit sold, so the retailer's profit will increase by an amount equal to the expected sales. The right-hand side of (13.7) is γ times the (expected) profit of the retailer; a \$1 price increase will lead to a demand reduction, which will cause the retailer to lose some profit, and this loss turns out to be equal to γ times the profit of the retailer. (This is a consequence of the multiplicative demand model.) Therefore, as the FOC given by (13.7) implies, the optimal price chosen by the retailer must set the expected sales volume equal to γ times the retailer's profit.

Likewise, under the consumer rebate equilibrium, the consumer rebate must satisfy the FOC for the manufacturer. Hence, by part (b) of Lemma 6, the consumer rebate r_C must satisfy

$$\begin{aligned}
&\beta\left(\frac{w-v}{p+r_R-v} y^*(p, r_C, r_R) + \int_0^{\frac{p+r_R-w}{p+r_R-v}} z(p, r_C, \xi)d\xi\right) \\
&= \gamma\alpha\left[(w-c)y^*(p, r_C, r_R) - (\beta r_C + r_R)\left(\frac{w-v}{p+r_R-v} y^*(p, r_C, r_R) + \int_0^{\frac{p+r_R-w}{p+r_R-v}} z(p, r_C, \xi)d\xi\right)\right].
\end{aligned} \tag{13.8}$$

The left-hand side of (13.8) is β times the expected sales of the product; a \$1 rebate increase means the manufacturer will pay β dollars more per each unit sold, so the

manufacturer's profit will decrease by an amount equal to β times the expected sales. The right-hand side of (13.8) is $\gamma\ \alpha$ times the profit of the manufacturer; a \$1 rebate increase will lead to a demand increase, which will cause the manufacturer to gain some profit, and this gain turns out to be equal to $\gamma\ \alpha$ times the profit of the manufacturer. (Once again, this is a consequence of the multiplicative demand model.) Therefore, the optimal price chosen by the manufacturer must set β times the expected sales volume equal to $\gamma\ \alpha$ times the manufacturer's profit.

In summary, both parties are using the (expected) sales volume as a benchmark; one is trying to set its profit equal to the sales volume multiplied by $\frac{1}{\gamma}$, and the other is trying to set its profit equal to $\frac{\beta}{\gamma\alpha}$ times the sales volume. Since both parties will be seeing the same sales volume in equilibrium, the last part of the proposition follows.

In the next section, we provide some numerical examples to compare the effects of the retailer and consumer rebates on the profits of the supply chain partners.

6 Numerical Examples

One natural question to ask is which rebate type each player in the supply chain prefers. Unfortunately, there is no clear-cut answer to this question. In particular, as one would expect, the values of α and β have a significant impact on the equilibrium that arises under consumer rebates, and, therefore, whether a party prefers consumer rebates to retailer rebates depends very much on the values of α and β. Consider the equilibrium results depicted in Table 13.1. These equilibria are obtained under the assumption that $f(\cdot)$ is given by the logit demand function; i.e., $f(x) = \dfrac{exp(u_1 - x)}{exp(u_0) + exp(u_1 - x)}$, and ϵ is distributed uniformly between 50 and 250. The other parameter values were as follows: $w = 18.55$,

Table 13.1 Equilibria under different rebate scenarios

Rebate type	α	β	Retailer rebate	Consumer rebate	Price	Manufacturer profit	Retailer profit	Expected demand
No rebate	–	–	–	–	20.55	417.95	49.67	62.37
Retailer	–	–	3.44	–	19.32	689.79	204.72	106.33
Consumer	1	0.8	–	11.58	29.74	694.90	868.46	132.89
Consumer	1	1	–	8.94	27.37	621.60	621.17	128.35
Consumer	0.4	1	–	7.18	22.33	430.05	172.01	101.94
Consumer + retailer	1	0.8	0	11.58	29.74	694.90	868.46	132.89
Consumer + retailer	1	1	0	8.94	27.37	621.60	621.17	128.35
Consumer + retailer	0.4	1	0	7.18	22.33	430.05	172.01	101.94

$c = 4.08, v = 0, u_1 = 22.91, u_0 = 2.70$. (This is one of many randomly-generated numerical examples we tested.) For the three combinations of parameters considered for consumer rebates, there was only a single equilibrium to the "both rebate types" game and it specified zero retailer rebate.[5] Thus, the prices and profits are the same as those given in the table under consumer rebates only. When $\alpha = 1$ and $\beta = 0.8$, both the retailer and the manufacturer prefer consumer rebates to retailer rebates. However, if β increases to 1 while keeping α fixed at 1, the manufacturer will now suffer from the increased claim rate of rebates, and, therefore, will now prefer retailer rebates to consumer rebates, while the retailer's preference is not affected by the change in β. On the other hand, if α decreases to 0.4 while keeping β fixed at 1, consumer rebates will now have a smaller impact on consumer demand, and, hence, the retailer will now prefer retailer rebates to consumer rebates. Therefore, neither party always prefers one rebate type to another.

6.1 A Form of Prisoner's Dilemma in Choosing What Rebate(s) to Offer

Note from Table 13.1 that for $\alpha = 0.4$ and $\beta = 1$, a supply chain in which both rebate types are allowed will settle in the same equilibrium as a supply chain in which only consumer rebates are allowed. Notice that there is a form of prisoners' dilemma here: The retailer rebate game equilibrium, even though it is preferred by both parties, is not an equilibrium in this game with both types allowed. This leads to an interesting observation: When the supply chain plays the game where both types of rebates are allowed, the supply chain ends up using only consumer rebates in equilibrium, an outcome that hurts both parties when compared to what they could achieve if only retailer rebates are allowed. The policy implication of this observation is that there are environments in which both the retailer and the manufacturer will agree in advance, before prices and rebates are set, to not allow the use of consumer rebates.

6.2 Both Parties May Prefer Consumer Rebates Even When All Consumers Claim Them

There exist cases where both parties prefer to use the consumer rebates to stimulate customer demand. For example, when $w = 10, c = 4, v = 0, u_1 = 30, u_0 = 20$, and $\alpha = \beta = 1$, both parties prefer consumer rebates. (Under retailer rebates only, the equilibrium profits are 122.45 for the manufacturer and 32.92 for the retailer. Under

[5] Under consumer rebate equilibria, the manufacturer expected profit to retailer expected profit ratios are not precisely β: α, since our searches were over fine grids that were nevertheless discrete.

consumer rebates only, the equilibrium profits are 151.70 for both the manufacturer and the retailer.) Under the consumer rebate equilibrium, the retail price is 13.28 and the consumer rebate is 3.77, which yield an effective price of 9.51, less than the wholesale price of 10. Note that this is an environment where all consumers claim their rebates, i.e., β is one; nevertheless, both parties prefer consumer rebates to retailer rebates. Moreover, in this supply chain, even when both types of rebates are allowed, it turns out that retailer rebates are not offered in equilibrium. The policy implication is that, contrary to popular belief, there exist environments in which supply chains prefer consumer rebates even when all consumers claim them.

A variant of this result can be seen in Table 13.1, where the supply chain profits are higher under consumer rebates than retailer rebates when $\alpha = \beta = 1$. In this case, because the wholesale price is fixed so much higher than cost, the retail price and consumer rebate are both high, leading to an effective price lower than under retailer rebates, but with a much higher margin to the retailer on units sold with still a good margin to the manufacturer on an increased level of sales. The manufacturer gets slightly lower profits but the retailer gets dramatically more.

6.3 Retailer May Choose to Sell at a Loss to Make Money on Rebates

Rebates can play an interesting role in the supply chain when the exogenously-fixed wholesale price is high. For example, if $w = 20, c = 5, v = 0, u_1 = 40, u_0 = 20$, and $\alpha = \beta = 1$, then the retailer rebate game equilibrium has a retail price of 19.24, which is lower than the wholesale price, and the retailer rebate is 4.53. Thus, in this example, the wholesale price is so high that the retailer sells the product at a loss to stimulate customer demand, and makes money only on rebates collected from the manufacturer rather than directly from consumers.

6.4 The Effect of Wholesale Price

To further examine the effect of wholesale price on the equilibrium, consider the case where only consumer rebates are allowed. Figure 13.2 shows the effect of w on the rebate size in equilibrium as well as on the manufacturer's and retailer's profits. In this example, $w = 20, c = 5, v = 0, u_1 = 40, u_0 = 20$, and $\alpha = \beta = 1$.

Observe from the figure that there is a threshold for the wholesale price such that only if the wholesale price exceeds this threshold will the manufacturer offer a strictly positive consumer rebate. This is intuitive: As the wholesale price gets larger, the manufacturer's profit margin per unit gets larger as well, and the manufacturer becomes more willing to pay a rebate to drive the retailer's stock level up. In addition, the figure suggests that the manufacturer's profit is at its

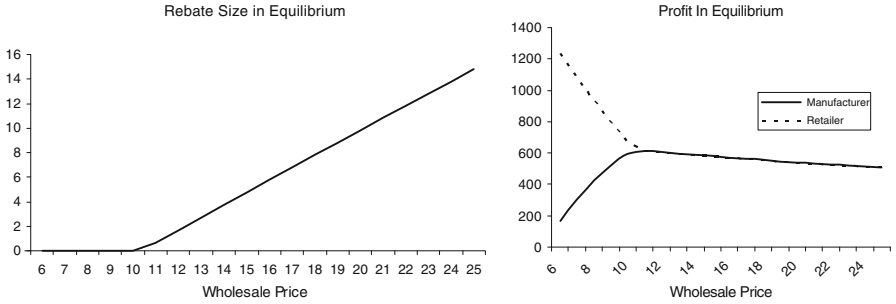


Fig. 13.2 Equilibrium rebate size (*left*) and profits (*right*) as a function of the wholesale price when only consumer rebates are allowed

highest at the threshold wholesale price. Therefore, if the manufacturer were to choose the wholesale price first, followed by a game where the consumer rebate and retail price are chosen simultaneously, then it would be optimal for the manufacturer to set the wholesale price equal to its threshold value, which would lead to a zero rebate in equilibrium. We have observed the same behavior in a number of numerical examples, but further analysis is needed to determine if this result is true in general.

7 Conclusion

We considered a supply chain where the retailer faces stochastic, effective-price-dependent demand and the manufacturer builds to order. We established some properties of the equilibrium that would arise when the manufacturer offers retailer and/or consumer rebates. We showed that supply chain profits are improved by the use of retailer rebates. On the other hand, consumer rebates may reduce the supply chain profit, but they will lead to an improvement whenever the effective fraction, α, is larger than the fraction of customers who claim their rebate, β. Furthermore, we showed that these two parameters have further significance: Under the equilibrium of the consumer rebate game, the ratio of (expected) retailer profits to (expected) manufacturer profits equals the ratio α/β. We discussed some interesting consequences of this property. We provided numerical examples to demonstrate that neither the retailer nor the manufacturer always prefers one particular kind of rebate to the other. In addition, our numerical examples suggest that, contrary to popular belief, it is possible for both firms to prefer consumer rebates even when all such rebates are redeemed.

In our model, we examined how the two rebate types differ from each other through their effects on the pricing and inventory decisions for a product. When the product's price is fixed, but the retailer is able to exert some type of hidden effort to sell the product; e.g., putting up in-store displays or advertising in local media, the

effects of retailer and consumer rebates are likely to differ again and are worthy of study. Another extension worthy of study is to address the moral hazard problem of misreporting retailer sales. One approach is to add buy-backs to the model (the manufacturer buys back unsold inventory at the end of the season at a set price), which could reduce the retailer's incentive to misreport sales. It would also be interesting to add a verification cost (of sold units) to the model.

We give a partial answer to the question of why consumer rebates are offered. Our numerical examples illustrate the existence of cases where the manufacturer will prefer offering consumer rebates to offering a retailer rebate. Consumer rebates help the manufacturer by increasing the stock level at the retailer, and our results suggest that they may be useful even when all customers claim them. Therefore, perhaps it is not too surprising that some firms choose to offer instantly redeemable rebates to online shoppers even though such rebates have high redemption rates. Bulkeley (1998) cites some alternative explanations for the use of consumer rebates. For example, consumer rebates may be seen as temporary price reductions, used in order to learn more about the customer population's price elasticity. Alternatively, in high-tech products, consumer rebates can be used to offer price discounts to consumers on older-generation products, which would eliminate the need for offering price protection to the retailer. Analysis of such uses for consumer rebates is left for future research. Hopefully, some of the structural results in this paper could prove useful for researchers who would like to further analyze the question of why consumer rebates are used. Another line of extension for this research is using more elaborate models for the redemption of consumer rebates, such as having heterogeneous consumer types, with differing values of α and β. A utility-based model that describes the customer's attitude towards redeeming a rebate would contribute to our understanding of the use of consumer rebates.

It is possible that some retailers will force manufacturers to move away from mail-in consumer rebates in the future. For example, in 2005 BestBuy announced that it would no longer stock products tied to mail-in rebates and it intended to implement this policy in the span of a few years (Menzies 2005). Indeed, according to an article published in the *US News & World Report* in 2008, BestBuy phased out mail-in rebates between 2005 and 2007 (Palmer 2008). BestBuy's stated reason was that mail-in rebates were cumbersome for the consumers. To the extent that our model captures the BestBuy environment (the major violation is likely to be that the wholesale price is not exogenous), it may be that BestBuy preferred the retailer rebate regime, although in our numerical examples where that happens, the manufacturer also prefers the retailer rebate regime, so would not resist dropping consumer rebates and instituting retailer rebates. Another explanation is that BestBuy was lobbying for having the consumer rebates instantaneously redeemable at the time of consumer purchase, as is done in the automotive industry. This might have the effect of increasing both the customer valuation α and redemption rate β to 1, which would make rebates the equivalent to price discounts offered directly by the manufacturer to consumers. Depending on what the values of α and β were prior to the cancelation of the mail-in rebates, such a change might have improved the total supply chain profit as well as been appreciated by consumers. There are other

explanations for BestBuy's position that are not covered by our model, such as that it helped BestBuy in its competition with other retailers. In any event, BestBuy could have been acting in its self interest, while claiming that its motivation was as a consumer advocate.

Acknowledgements The authors would like to thank the editors and an anonymous referee for their comments that helped improve the paper.

Appendix

For the purposes of the appendix, let $h(\cdot) = \frac{\phi_\epsilon(\cdot)}{1-\Phi_\epsilon(\cdot)}$ denote the failure rate of Φ_ϵ. Throughout the appendix, we will use the following short-hand notation by dropping the functional arguments: $f = f(p - \alpha r_C)$, $z = z(p, r_C, \xi)$, $y^* = y^*(p, r_C, r_R)$ and $h = h\left(\frac{y^*(p,r_C,r_R)}{f(p-\alpha r_C)}\right)$. In addition, define $\gamma := -\frac{f'}{f}$ and $\theta := \frac{f''}{f}$. Hence, by $(A3)$, $\gamma > 0$, and, by $(A4)$, $\gamma' = \gamma^2 - \theta \geq 0$. We first state and prove some lemmas that will be useful in the proofs of the propositions.

Lemma 1 *Suppose (A1) holds. For z(p,r*$_C$*,ξ) implicitly defined by* (13.4) *, we have:*

(a) $\nabla_1 z = -\gamma z$,
(b) $\nabla_2 z = \alpha\gamma z$,
(c) $\nabla^2_{11} z = \theta z$
(d) $\nabla^2_{22} z = \alpha^2\theta z$,
(e) $\nabla^2_{12} z = -\alpha\theta z$.

Proof of Lemma 1 By virtue of $(A1)$, we can rewrite (13.4) as $\Phi_\epsilon\left(\frac{z}{f}\right) = \xi$. Now, implicit differentiation of this identity with respect to p yields the following:

$$\nabla_1 z f - z f' = 0.$$

The first part of the lemma follows from the above equality recalling the definition of $\gamma := -\frac{f'}{f}$. The proof of the second part follows the same logic. The third part can be obtained directly by partial differentiation of the expression for $\nabla_1 z$. Likewise, the fourth and fifth parts are obtained by partial differentiation of the expression for $\nabla_2 z$. □

Lemma 2 *Suppose (A1) through (A4) hold. For* $y^*(p, r_C, r_R)$ *implicitly defined by* (13.2)*, we have:*

(a) $\nabla_1 y^* = -\gamma y^* + \dfrac{f}{(p + r_R - v)h}$,

(b) $\nabla_2 y^* = \alpha\gamma y^* > 0$,

(c) $\nabla_3 y^* = \dfrac{f}{(p + r_R - v)h} > 0$,

(d) $\nabla_1 y^* = -\dfrac{1}{\alpha}\nabla_2 y^* + \nabla_3 y^*$,

(e) $\nabla^2_{22} y^* = \alpha^2\theta y^*$,

(f) $\nabla^2_{33} y^* = -\dfrac{f}{(p + r_R - v)^2 h} - \dfrac{fh'}{(p + r_R - v)^2 h^3} < 0$,

(g) $\nabla^2_{23} y^* = \alpha\gamma \dfrac{f}{(p + r_R - v)h} > 0$,

(h) $\nabla^2_{13} y^* = -\dfrac{1}{\alpha}\nabla^2_{23} y^* + \nabla^2_{33} y^* < 0$.

Proof of Lemma 2 *Proofs of (a) through (d)* Due to (*A*1), we can rewrite (13.2) as

$$\Phi_\epsilon\left(\frac{y^*}{f}\right) = \frac{p + r_R - w}{p + r_R - v} \tag{13.9}$$

Now, implicit differentiation of (13.9) with respect to p yields

$$\frac{\nabla_1 y^* f - f' y^*}{f^2}\phi_\epsilon\left(\frac{y^*}{f}\right) = \frac{w - v}{(p + r_R - v)^2}.$$

Recalling the definition of $h(\cdot) = \dfrac{\phi_\epsilon(\cdot)}{1 - \Phi_\epsilon(\cdot)}$ and noting that $1 - \Phi_\epsilon\left(\dfrac{y^*}{f}\right) = \dfrac{w - v}{p + r_R - v}$ [this follows from (13.9)], we can leave $\nabla_1 y^*$ alone in the above expression to obtain part (a) of the lemma. The proofs of parts (b) and (c) follow the same line of argument. Part (d) of the lemma follows directly from parts (a) through (c).

Proofs of (e) through (g) These follow from partial differentiation of the expressions obtained in parts (a) through (c). To see why $\nabla^2_{33} y^* < 0$, recall that $h(\cdot)$ is the failure rate and it is an increasing function by (*A*2). To see why $\nabla^2_{23} y^* > 0$, recall that $\gamma > 0$ by (*A*3).

Proof of (h) This follows from part (d) of the lemma. □

Lemma 3 *Given r_C and r_R, if $\tilde{p}$ satisfies $\nabla_1 \Pi_R(\tilde{p}, r_C, r_R) = 0$, then $\nabla_1 y^*(\tilde{p}, r_C, r_R) < 0$.*

Proof of Lemma 3 Omitted. See Aydin and Porteus (2008) for the proof of the same result under more general conditions. □

Lemma 4 *Let f(x) be a twice-continuously-differentiable function of a single real variable defined on [a,∞). Suppose that $f''(x) < 0$ at any $x \geq a$ that satisfies $f'(x) = 0$. Then:*

(a) *(Petruzzi and Dada1999) If $f'(a) > 0$ and f(x) is strictly decreasing in x as x tends to infinity, then there exists a unique $x^* > a$ that satisfies $f'(x) = 0$, and x^* maximizes f(x).*
(b) *If $f'(a) \leq 0$ then f(x) is non-increasing for all $x \geq a$, and $x^* = a$ maximizes f(x).*

Proof of Lemma 4 Omitted. Lemma 4(a) is due to Petruzzi and Dada (1999). See Aydin and Porteus (2008) for a detailed proof. The proof of part (b) is very similar. □

Lemma 5 *In a two-player game, let $g_i(x_1, x_2)$ be the payoff function of player $i = 1,2$ when the strategies chosen by players 1 and 2 are x_1 and x_2, respectively. The strategy space for player i is $X_i := \{x : \underline{x}_i \leq x \leq \overline{x}_i\}$. Suppose that g_i is continuous and quasi-concave with respect to x_i, $i = 1,2$. Let $x_i^*(x_j)$ be the best response of player i when player j chooses strategy x_j; i.e., $x_i^*(x_j) = \text{argmax}_{x_i}(g_i(x_1, x_2))$. Then:*

(a) *There exists at least one pure strategy Nash equilibrium.*
(b) *If $\frac{dx_1^*(x_2)}{dx_2} \frac{dx_2^*(x_1)}{dx_1} < 1$, then there exists a unique pure strategy Nash equilibrium.*

Proof of Lemma 5 Omitted. See Cachon and Netessine (2004) for a summary of standard results in game theory. □

Lemma 6 *Suppose (A1) through (A4) hold. Let $\Pi_R(p, r_C, r_R)$ and $\Pi_M(p, r_C, r_R)$ be as defined by (13.5) and (13.6) , respectively. Then:*

(a)
$$\nabla_1 \Pi_R(p, r_C, r_R) = \int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi + \frac{w-v}{p+r_R-v} y^* - \gamma(p+r_R-v) \int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi$$

(b)
$$\nabla_2 \Pi_M(p, r_C, r_R) = (w-c)\alpha\gamma y^* - [\beta + \alpha\gamma(r_R + \beta r_C)] \left(\int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi + \frac{w-v}{p+r_R-v} y^* \right),$$

(c)

$$\nabla_3 \Pi_M(p, r_C, r_R) = \left[(w-c) - (r_R + \beta r_C)\frac{w-v}{p+r_R-v}\right]\frac{f}{(p+r_R-v)h} - \int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi - \frac{w-v}{p+r_R-v}y^*$$

(d) *For* $p + r_R > v$*:*

$$\nabla^2_{11}\Pi_R(p, r_C, r_R)\big|_{\nabla_1 \Pi_R = 0} = -\gamma \int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi + \frac{w-v}{p+r_R-v}\nabla_1 y^* - (p+r_R-v)\gamma' \int_0^{\frac{p+r_R-w}{p+r_R-v}} z < 0$$

(e) *For* $p + r_R > v$*:*

$$\nabla^2_{22}\Pi_M(p, r_C, r_R)\big|_{\nabla_2 \Pi_M = 0} = -\alpha\beta\gamma\left(\int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi + \frac{w-v}{p+r_R-v}y^*\right) + \left[(w-c)y^* - (r_R+\beta r_C)\left(\int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi + \frac{w-v}{p+r_R-v}y^*\right)\right](\alpha^2\theta - \alpha^2\gamma^2)$$

(f) *For* $p + r_R > v$*:*

$$\nabla^2_{33}\Pi_M(p, r_C, r_R)\big|_{\nabla_3 \Pi_M = 0} = -\frac{1}{p+r_R-v}\left[\int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi + \frac{w-v}{p+r_R-v}y^*\right] - \left[(w-c) - (r_R+\beta r_C)\frac{w-v}{p+r_R-v}\right]\frac{fh'}{(p+r_R-v)^2 h^3} + \left[-2\frac{w-v}{p+r_R-v} + (r_R+\beta r_C)\frac{w-v}{(p+r_R-v)^2}\right]\frac{f}{(p+r_R-v)h} < 0$$

(g) *For* $p + r_R > v$*:*

$$\nabla^2_{23}\Pi_M(p, r_C, r_R)\big|_{\nabla_2 \Pi_M = 0} = -\frac{w-v}{p+r_R-v}\frac{\beta f}{(p+r_R-v)h} < 0$$

(h) *For* $p + r_R > v$*:*

$$\nabla_{13}^2 \Pi_R(p, r_C, r_R)\big|_{\nabla_1 \Pi_R = 0} = \frac{w - v}{p + r_R - v} \nabla_1 y^* - \gamma \int_0^{\frac{p + r_R - w}{p + r_R - v}} z d\xi < 0$$

(i) *For* $p + r_R > v$*:*

$$\nabla_{13}^2 \Pi_M(p, r_C, r_R)\big|_{\nabla_3 \Pi_M = 0} = -\left[(w - c) - (r_R + \beta r_C)\frac{w}{p + r_R - v}\right] \frac{f h'}{(p + r_R - v)^2 h^3}$$
$$- \left[(2w - c) - 2(r_R + \beta r_C)\frac{w - v}{p + r_R - v}\right] \frac{f}{(p + r_R - v)^2 h} < 0$$

(j) *For* $p + r_R > v$*:*

$$\nabla_{12}^2 \Pi_R(p, r_C, r_R)\big|_{\nabla_1 \Pi_R = 0} = \alpha(p + r_R - v)\gamma' \int_0^{\frac{p + r_R - w}{p + r_R - v}} z d\xi > 0$$

Proof of Lemma 6 *Proof of (a)* The result follows from partial differentiation of $\Pi_R(p, r_C, r_R)$ [defined by (13.5)] with respect to p and substituting for $\nabla_1 z$ using Lemma 1(a).

Proof of (b) The result follows by partial differentiation of $\Pi_M(p, r_C, r_R)$ [defined by (13.6)] with respect to r_C and substituting for $\nabla_2 z$ and $\nabla_2 y^*$ from Lemma 1 (b) and from Lemma 2(b).

Proof of (c) The result follows by partial differentiation of $\Pi_M(p, r_C, r_R)$ [defined by (13.6)] with respect to r_R and substituting for $\nabla_3 y^*$ from Lemma 2(c).

Proof of (d) The second partial of $\Pi_R(p, r_C, r_R)$ with respect to p is given, after substituting for $\nabla_1 z$ and $\nabla_{11}^2 z$ using Lemma 1(a) and (c), by

$$\nabla_{11}^2 \Pi_R(p, r_C, r_R) = -2\gamma \int_0^{\frac{p + r_R - w}{p + r_R - v}} z d\xi + \frac{w - v}{p + r_R - v} \nabla_1 y^* - \frac{w - v}{p + r_R - v} \gamma y^*$$
$$+ (p + r_R - v)\theta \int_0^{\frac{p + r_R - w}{p + r_R - v}} z d\xi$$

Thus, when $\nabla_1 \Pi_R = 0$, using part (a) of the lemma, we have

$$\nabla_{11}^2 \Pi_R(p, r_C, r_R) = -\gamma \int_0^{\frac{p + r_R - w}{p + r_R - v}} z d\xi + \frac{w - v}{p + r_R - v} \nabla_1 y^* + (p + r_R - v)\left(\theta - \gamma^2\right)$$
$$\int_0^{\frac{p + r_R - w}{p + r_R - v}} z d\xi,$$

which is strictly negative, by Lemma 3 and since $\gamma > 0$ [by (*A*3)] and $\gamma' = \gamma^2 - \theta \geq 0$ [by (*A*4)].

Proof of (e) The second partial of $\Pi_M(p, r_C, r_R)$ with respect to r_C is given, after substituting for $\nabla_2 z$, $\nabla_{22}^2 z$, $\nabla_2 y^*$ and $\nabla_{22}^2 y^*$ from Lemma 1(b) and (d), and from Lemma 2(b) and (e), by

$$\nabla_{22}^2 \Pi_M(p, r_C, r_R) = (w - c)\alpha^2\theta y^* - [2\alpha\beta\gamma + (r_R + \beta r_C)\alpha^2\theta] \left(\int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi + \frac{w - v}{p + r_R - v} y^* \right)$$

Thus, when $\nabla_2 \Pi_M = 0$, using part (b) of the lemma, we have

$$\nabla_{22}^2 \Pi_M(p, r_C, r_R) = -\alpha\beta\gamma \left(\int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi + \frac{w - v}{p + r_R - v} y^* \right) + \left[(w - c)y^* - (r_R + \beta r_C) \left(\int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi + \frac{w - v}{p + r_R - v} y^* \right) \right] (\alpha^2\theta - \alpha^2\gamma^2),$$

which is strictly negative since $\gamma > 0$ [by (*A*3)], $\gamma' = \gamma^2 - \theta \geq 0$ [by (*A*4)] and the term in brackets is Π_M which should be positive when $\nabla_2 \Pi_M = 0$.

Proof of (f) The second partial of $\Pi_M(p, r_C, r_R)$ with respect to r_R is given, after substituting for $\nabla_3 y^*$ and $\nabla_{33}^2 y^*$ from Lemma 2(c) and (f), by

$$\nabla_{33}^2 \Pi_M(p, r_C, r_R) = \left[(w - c) - (r_R + \beta r_C) \frac{w - v}{p + r_R - v} \right] \left[-\frac{f}{(p + r_R - v)^2 h} - \frac{f h'}{(p + r_R - v)^2 h^3} \right] + \left[-2\frac{w - v}{p + r_R - v} + (r_R + \beta r_C) \frac{w - v}{(p + r_R - v)^2} \right] \frac{f}{(p + r_R - v)h}.$$

Thus, when $\nabla_3 \Pi_M = 0$, using part (c) of the lemma, we have

$$\nabla^2_{33}\Pi_M(p, r_C, r_R) = -\frac{1}{p + r_R - v}\left[\int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi + \frac{w-v}{p+r_R-v} y^*\right] - \left[(w-c) - (r_R + \beta r_C)\frac{w-v}{p+r_R-v}\right]\frac{f h'}{(p+r_R-v)^2 h^3} + \left[-2\frac{w-v}{p+r_R-v} + (r_R+\beta r_C)\frac{w-v}{(p+r_R-v)^2}\right]\frac{f}{(p+r_R-v)h}.$$

In order to show $\nabla^2_{33}\Pi_M(p, r_C, r_R)\big|_{\nabla_3 \Pi_M = 0} < 0$, first note that, by Lemma 6(c), if $\nabla_3 \Pi_M(p, r_C, r_R) = 0$, then we must have $(w-c) - (r_R + \beta r_C)\frac{w-v}{p+r_R-v} > 0$, in which case we will also have $-2\frac{w-v}{p+r_R-v} + (r_R + \beta r_C)\frac{w-v}{(p+r_R-v)^2} < 0$. (This can be verified through some algebra.) After making these observations, the desired result now follows since $h' > 0$ by assumption (*A*2).

Proof of (g) It can be verified that the cross-partial $\nabla_{23}\Pi_M(p, r_C, r_R)$ is given, after substituting for $\nabla_2 z$ from Lemma 1(b) and for $\nabla_2 y^*$, $\nabla_3 y^*$ and $\nabla_{23} y^*$ from Lemma 2(b), (c) and (g), by

$$\nabla^2_{23}\Pi_M(p, r_C, r_R) = -\left[(w-c) - (r_R + \beta r_C)\frac{w-v}{p+r_R-v}\right]\alpha\gamma\frac{f}{(p+r_R-v)h} - \alpha\gamma\left(\int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi + \frac{w-v}{p+r_R-v} y^*\right) - \beta\frac{w-v}{p+r_R-v}\frac{f}{(p+r_R-v)h}$$

Thus, when $\nabla_3 \Pi_M = 0$, using part (c) of the lemma, we have

$$\nabla^2_{23}\Pi_M(p, r_C, r_R) = -\beta\frac{w-v}{p+r_R-v}\frac{f}{(p+r_R-v)h}$$

Proof of (h) It can be verified that $\nabla^2_{13}\Pi_R(p, r_C, r_R)$ is given, after substituting for $\nabla_1 z$ from Lemma 1(a) and $\nabla_3 y^*$ from Lemma 2(c), by

$$\nabla^2_{13}\Pi_R(p, r_C, r_R) = \frac{w-v}{p+r_R-v}\left(-\gamma y^* + \frac{f}{(p+r_R-v)h}\right) - \gamma\int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi$$

Now, from part (a) of Lemma 2, we note that $-\gamma y^* + \frac{f}{h(p+r_R-v)} = \nabla_1 y^*$. The desired conclusion on the sign follows from $\gamma > 0$ [by (*A*3)] and Lemma 3.

Proof of (i) It can be verified that $\nabla_{13}\Pi_M(p, r_C, r_R)$ is given, after substituting for $\nabla_1 z$ from Lemma 1(a) and for $\nabla_1 y^*$, $\nabla_3 y^*$, and $\nabla_{13} y^*$ from Lemma 2(a), (c) and (h), by

$$\begin{aligned}\nabla^2_{13}\Pi_M(p, r_C, r_R) = &\left[(w-c) - (r_R + \beta r_C)\frac{w-v}{p+r_R-v}\right] \\ &\left[-\gamma\frac{f}{(p+r_R-v)h} - \frac{f}{(p+r_R-v)^2 h} - \frac{fh'}{(p+r_R-v)^2 h^3}\right] \\ &+ \gamma\int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi - \frac{w-v}{p+r_R-v}\left[-\gamma y^* + \frac{f}{(p+r_R-v)h}\right] \\ &+ (r_R + \beta r_C)\frac{w-v}{(p+r_R-v)^2}\frac{f}{(p+r_R-v)h}\end{aligned}$$

Now, using part (c) of the lemma and the above expression, one can verify through some algebra that the following is true when $\nabla_3\Pi_M = 0$:

$$\begin{aligned}\nabla^2_{13}\Pi_M(p, r_C, r_R) = &-\left[(w-c) - (r_R + \beta r_C)\frac{w-v}{p+r_R-v}\right]\frac{fh'}{(p+r_R-v)^2 h^3} \\ &- \left[(2w-c) - 2(r_R + \beta r_C)\frac{w-v}{p+r_R-v}\right]\frac{f}{(p+r_R-v)^2 h}\end{aligned}$$

In order to show that $\nabla^2_{13}\Pi_M(p, r_C, r_R)\big|_{\nabla_3\Pi_M=0} < 0$, note that, by part (c) of the lemma, if $\nabla_3\Pi_M(p, r_C, r_R) = 0$, then we must have $(w-c) - (r_R + \beta r_C)\frac{w}{p+r_R-v} > 0$, in which case we will also have $(2w-c) - (r_R + \beta r_C)\frac{w}{p+r_R-v} > 0$. The desired result now follows since $h' > 0$ by assumption (*A*2).

Proof of (j) It can be verified that $\nabla^2_{12}\Pi_R(p, r_C, r_R)$ is given, after substituting for $\nabla_2 z$ and $\nabla^2_{12} z$ from Lemma 1(b) and (e) and for $\nabla_2 y^*$ from Lemma 2(b) by

$$\nabla^2_{12}\Pi_R(p, r_C, r_R) = \alpha\gamma\int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi + \frac{w-v}{p+r_R-v}\alpha\gamma y^* - \alpha\theta(p+r_R-v)\int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi$$

Now, when $\nabla_1\Pi_R = 0$, the following relationship can be verified through algebra, using part (a) of the lemma and the above expression:

$$\nabla^2_{12}\Pi_R(p, r_C, r_R) = \alpha(\gamma^2 - \theta)(p + r_R - v)\int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi,$$

which is strictly positive since $\gamma' = \gamma^2 - \theta > 0$ by virtue of (*A*4). □

Proof of Proposition 1 Using Lemma 6(a), one can verify that $\nabla_1\Pi_R(w - r_R, r_C, r_R) > 0$. Again using Lemma 6(a), one can also verify that $\nabla_1\Pi_R(p, r_C, r_R) < 0$ as $p \to \infty$. Given these observations, the result now follows from Lemmas 4(a) and 6(d). □

Proof of Proposition 2 *Proof of (a)* Given p and r_R, using Lemma 6(b), one can verify that $\nabla_2\Pi_M(p, r_C, r_R) < 0$ as $r_C \to \infty$. From Lemma 6(e), we know that $\nabla^2_{22}\Pi_M(p, r_C, r_R)|_{\nabla_2\Pi_M=0} < 0$. The result now follows by applying parts (a) and (b) of Lemma 4.

Proof of (b) We can focus on r_R such that $p + r_R \geq w$ (and, hence, $p + r_R \geq v$), since the retailer would stock zero units otherwise, and the manufacturer would make zero profits. Given p and r_C, using Lemma 6(c), one can verify that $\nabla_3\Pi_M(p, r_C, r_R) < 0$ as $r_R \to \infty$. From Lemma 6(f), we know that $\nabla^2_{33}\Pi_M(p, r_C, r_R)|_{\nabla_3\Pi_M=0} < 0$. The result now follows by applying parts (a) and (b) of Lemma 4.

Proof of (c) When $\nabla_2\Pi_M(p, r_C, r_R) = 0$, we can use Lemma 6(b) to write

$$\beta\int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi + \beta\frac{w-v}{p+r_R-v}y^* = \left[(w-c) - (r_R + \beta r_C)\frac{w-v}{p+r_R-v}\right]\alpha\gamma y^* - (r_R + \beta r_C)\alpha\gamma\int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi.$$

Note that for the above equality to hold, we need to have $(w-c) - (r_R + r_C)\frac{w-v}{p+r_R-v} > 0$ [since $\gamma > 0$ by assumption (*A*3)]. Similarly, when $\nabla_3\Pi_M(p, r_C, r_R) = 0$, we can use Lemma 6(c) to write

$$\int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi + \frac{w-v}{p+r_R-v}y^* = \left[(w-c) - (r_R + \beta r_C)\frac{w-v}{p+r_R-v}\right]\frac{f}{(p+r_R-v)h}.$$

Again, note that for the above equality to hold, we need to have $(w-c) - (r_R + r_C)\frac{w-v}{p+r_R-v} > 0$. By using the last two equalities, we obtain:

$$-(r_R + \beta r_C)\alpha\gamma\int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi = \left[(w-c) - (r_R + \beta r_C)\frac{w-v}{p+r_R-v}\right]\left[\frac{\beta f}{(p+r_R-v)h} - \alpha\gamma y^*\right]$$

Now, note that the second term in brackets on the right-hand side of the equality above is $-\nabla_2 y^* + \beta\nabla_3 y^*$ (from parts (b) and (c) of Lemma 2). Also, as noted above, we must have $(w-c) - (r_R + r_C)\frac{w-v}{p+r_R-v} > 0$. The term on the left-hand side is negative [since $\gamma > 0$ by assumption (*A*3)]. The desired result now follows. □

Proof of Proposition 3 Under an equilibrium solution $(\tilde{p}, \tilde{r}_C, \tilde{r}_R)$ with $\tilde{r}_C > 0$, we need to have $\nabla_2 \Pi_M(\tilde{p}, \tilde{r}_C, \tilde{r}_R) = \nabla_1 \Pi_R(\tilde{p}, \tilde{r}_C, \tilde{r}_R) = 0$ (by Proposition 1 and part (a) of Proposition 2). Since $\nabla_2 \Pi_M(\tilde{p}, \tilde{r}_C, \tilde{r}_R) = 0$, we know from Lemma 6(b) that

$$\begin{aligned}\beta\frac{w-v}{\tilde{p}+\tilde{r}_R-v}y^* + \beta\int_0^{\frac{\tilde{p}+\tilde{r}_R-w}{\tilde{p}+\tilde{r}_R-v}} z d\xi &= (w-c)\alpha\gamma y^* \\ &\quad -(\tilde{r}_R+\beta\tilde{r}_C)\alpha\gamma\left(\int_0^{\frac{\tilde{p}+\tilde{r}_R-w}{\tilde{p}+\tilde{r}_R}-v} z d\xi + \frac{w-v}{\tilde{p}+\tilde{r}_R-v}y^*\right), \\ &= \alpha\gamma\Pi_M(\tilde{p}, \tilde{r}_C, \tilde{r}_R) \text{ by (6)}\end{aligned} \tag{13.10}$$

Also, since $\nabla_1 \Pi_R(\tilde{p}, \tilde{r}_C, \tilde{r}_R) = 0$, we know from Lemma 6(a) that

$$\begin{aligned}\frac{w-v}{\tilde{p}+\tilde{r}_R-v}y^* + \int_0^{\frac{\tilde{p}+\tilde{r}_R-w}{\tilde{p}+\tilde{r}_R-v}} z d\xi &= (\tilde{p}+\tilde{r}_R-v)\gamma\int_0^{\frac{\tilde{p}+\tilde{r}_R-w}{\tilde{p}+\tilde{r}_R-v}} z d\xi, \\ &= \gamma\Pi_R(\tilde{p}, \tilde{r}_C, \tilde{r}_R) \text{ by (5)}\end{aligned} \tag{13.11}$$

Now, (13.10) and (13.11) together allow us conclude $\frac{\Pi_M(\tilde{p}, \tilde{r}_C, \tilde{r}_R)}{\Pi_R(\tilde{p}, \tilde{r}_C, \tilde{r}_R)} = \frac{\beta}{\alpha}$. □

Proof of Proposition 4 *Proof of (a)* Throughout the proof, recall that $p^*(r_R)$ will satisfy $\nabla_1 \Pi_R(p^*(r_R), 0, r_R) = 0$ at any given r_R (by Proposition 1). By implicit differentiation of this identity with respect to r_R, we obtain $\frac{dp^*(r_R)}{dr_R} = -\frac{\nabla^2_{13}\Pi_R(p^*(r_R),0,r_R)}{\nabla^2_{11}\Pi_R(p^*(r_R),0,r_R)}$. Hence, we will conclude the proof of part (a) if we can show that $\nabla^2_{11}\Pi_R(p^*(r_R), 0, r_R) \le \nabla^2_{13}\Pi_R(p^*(r_R), 0, r_R) < 0$. From Lemma 6 (d) and (h), we know that $\nabla^2_{11}\Pi_R(p^*(r_R), 0, r_R) < 0$ and $\nabla^2_{13}\Pi_R(p^*(r_R), 0, r_R) < 0$. Again, from Lemma 6(d) and (h), note that:

$$\nabla^2_{11}\Pi_R(p^*(r_R), 0, r_R) - \nabla^2_{13}\Pi_R(p^*(r_R), 0, r_R) = -(p + r_R - v)\gamma'\int_0^{\frac{p+r_R-w}{p+r_R-v}} z \le 0, \tag{13.12}$$

where the inequality follows from $\gamma' \ge 0$ [by (*A*4)]. Thus, we are able to conclude that

$$\nabla^2_{11}\Pi_R(p^*(r_R), 0, r_R) \le \nabla^2_{13}\Pi_R(p^*(r_R), 0, r_R) < 0,$$

which concludes the proof of part (a).

Proof of (b) Given p, w and $r_C = 0$, it follows from Proposition 2(b) that either ${r_R}^*(p) = 0$ or $r_R^*(p) > 0$ in which case $r_R^*(p)$ satisfies $\nabla_3 \Pi_M(p, 0, r_R^*(p)) = 0$. If ${r_R}^*(p) = 0$ for all $p > 0$, then part (b) holds trivially. Suppose now there exists a p at which ${r_R}^*(p) > 0$ and satisfies $\nabla_3 \Pi_M(p, 0, r_R^*(p)) = 0$. By implicit differentiation of this identity with respect to p, we obtain $\dfrac{dr_R^*(p)}{dp} = -\dfrac{\nabla_{13}^2 \Pi_M(p, 0, r_R^*(p))}{\nabla_{33}^2 \Pi_M(p, 0, r_R^*(p))}$. We already know from Lemma 6(f) and (i) that $\nabla_{33}^2 \Pi_M(p, 0, r_R^*(p)) < 0$ and $\nabla_{13}^2 \Pi_M(p, 0, r_R^*(p)) < 0$. Furthermore, again from Lemma 6(f) and (i), one can verify that

$$\begin{aligned}\nabla_{13}^2 \Pi_M(p, r_C, r_R)\big|_{\nabla_3 \Pi_M = 0} = \nabla_{33}^2 \Pi_M(p, r_C, r_R)\big|_{\nabla_3 \Pi_M = 0} + \frac{w - v}{p + r_R - v} \frac{f}{(p + r_R - v)h} \\ - \frac{1}{p + r_R - v} \left\{ - \int_0^{\frac{p + r_R - w}{p + r_R - v}} z d\xi - \frac{w - v}{p + r_R - v} y^* \right. \\ \left. + \left[(w - c) - (r_R + \beta r_C) \frac{w - v}{p + r_R - v} \right] \frac{f}{(p + r_R - v)h} \right\}\end{aligned}$$

From Lemma 6(c), we observe that the term in curly brackets above is in fact $\nabla_3 \Pi_M(p, r_C, r_R)$. Therefore, from the above expression, we obtain:

$$\nabla_{13}^2 \Pi_M(p, r_C, r_R)\big|_{\nabla_3 \Pi_M = 0} = \nabla_{33}^2 \Pi_M(p, r_C, r_R)\big|_{\nabla_3 \Pi_M = 0} + \frac{w - v}{p + r_R - v} \frac{f}{(p + r_R - v)h}$$

Hence, from the last equality, we conclude that $\nabla_{33}^2 \Pi_M(p, 0, r_R^*(p)) < \nabla_{13}^2 \Pi_M(p, 0, r_R^*(p))$, which, along with $\nabla_{33}^2 \Pi_M(p, 0, r_R^*(p)) < 0$ and $\nabla_{13}^2 \Pi_M(p, 0, r_R^*(p)) < 0$, allows us to conclude that $-1 < \frac{dr_R^*(p)}{dp} < 0$. Recall that we assumed p is such that $r_R^*(p) > 0$. For some p', we will have $r_R^*(p') = 0$, and ${r_R}^*(p)$ will remain zero for all $p > p'$, and hence $\frac{dr_R^*(p)}{dp}$ will be zero for all $p > p'$. (If ${r_R}^*(p)$ were to become positive for some $p'' > p'$, this would be a contradiction to the result that $\frac{dr_R^*(p)}{dp} < 0$ when $r_R^*(p) > 0$.)

Proof of (c) The existence of the Nash equilibrium follows from Lemma 5(a), Propositions 1 and 2(b). The uniqueness of the Nash equilibrium follows from Lemma 5(b) and parts (a) and (b) of this proposition. (Note that, in order to apply Lemma 5, we need upper bounds on the decision variables of the retailer and the manufacturer, p and r_R, respectively. We could satisfy this requirement by picking arbitrarily large numbers to bound the feasible choices for p and r_R.) □

Proof of Proposition 5 Throughout the proof, let $p^*(r_R)$ denote the optimal retail price chosen by the retailer at a given r_R when $r_C = 0$. *Proof of (a)* Note that $p_o = p^*(0)$ whereas $\tilde{p} = p^*(\tilde{r}_R)$. Therefore, $\tilde{p} - p_o = \int_0^{\tilde{r}_R} \frac{dp^*(r_R)}{dr_R} dr_R$. By Proposition 4(a), $-1 < \frac{dp^*(r_R)}{dr_R} < 0$. The desired result follows.

Proof of (b) Note that $y_o = y^*(p^*(0), 0, 0)$ whereas $\tilde{y} = y^*(p^*(\tilde{r}_R), 0, \tilde{r}_R)$. Now, $\tilde{y} - y_o = \int_0^{\tilde{r}_R} \frac{dy^*(p^*(r_R),0,r_R)}{dr_R} dr_R$. Therefore, we will conclude the proof if we can show that $\frac{dy^*(p^*(r_R),0,r_R)}{dr_R} > 0$. Note that $\frac{dy^*(p^*(r_R),0,r_R)}{dr_R} = \nabla_3 y^*(p^*(r_R), 0, r_R) + \frac{dp^*(r_R)}{dr_R} \nabla_1 y^*(p^*(r_R), 0, r_R)$. Now, $\nabla_3 y^*(p^*(r_R), 0, r_R) > 0$ from Lemma 2(c), $\frac{dp^*(r_R)}{dr_R} < 0$ from of Proposition 4(a) and $\nabla_1 y^*(p^*(r_R), 0, r_R) < 0$ from Lemma 3. (To see why Lemma 3 can be applied here, recall that $\nabla_1 \Pi_R(p^*(r_R), 0, r_R) = 0$ by Proposition 1 since $p^*(r_R)$ optimizes Π_R.) These observations imply that $\frac{dy^*(p^*(r_R),0,r_R)}{dr_R} > 0$, which yields the desired result.

Proof of (c) Note that $\Pi_{SC}(\tilde{p}, 0, \tilde{r}_R) = \Pi_{SC}(p^*(\tilde{r}_R), 0, \tilde{r}_R)$ and $\Pi_{SC}(p_o, 0, 0) = \Pi_{SC}(p^*(0), 0, 0)$. Therefore, $\Pi_{SC}(\tilde{p}, 0, \tilde{r}_R) - \Pi_{SC}(p_o, 0, 0) = \int_0^{\tilde{r}_R} \frac{d\Pi_{SC}(p^*(r_R),0,r_R))}{dr_R} dr_R$. Hence, if we can show that $\Pi_{SC}(p^*(r_R), 0, r_R)$ is increasing in r_R for $r_R \le w - c$, then the desired result will follow. Hence, we want to show that

$$\frac{d\Pi_{SC}(p^*(r_R), 0, r_R)}{dr_R} = \frac{dp^*(r_R)}{dr_R} \nabla_1 \Pi_{SC}(p^*(r_R), 0, r_R) + \nabla_3 \Pi_{SC}(p^*(r_R), 0, r_R)$$

is positive. The following equalities can be verified using (13.5) and (13.6):

$$\begin{aligned}
\nabla_1 \Pi_{SC}(p, r_C, r_R) &= \nabla_1 \Pi_R(p, r_C, r_R) + \nabla_1 \Pi_M(p, r_C, r_R) \\
&= \nabla_1 \Pi_R(p, r_C, r_R) + \left[w - c - (r_R + \beta r_C) \frac{w - v}{p + r_R - v} \right] \nabla_1 y^* \\
&\quad - (r_R + \beta r_C) \int_0^{\frac{p + r_R - w}{p + r_R - v}} \nabla_1 z \, d\xi \\
\nabla_3 \Pi_{SC}(p, r_C, r_R) &= \nabla_3 \Pi_R(p, r_C, r_R) + \nabla_3 \Pi_M(p, r_C, r_R) \\
&= \left[w - c - (r_R + \beta r_C) \frac{w - v}{p + r_R - v} \right] \nabla_3 y^*
\end{aligned}$$

Note that $\nabla_1 \Pi_R(p^*(r_R), 0, r_R) = 0$ by definition of $p^*(r_R)$ and Proposition 1. Thus, after substitution and rearranging terms, we get

$$\frac{d\Pi_{SC}(p^*(r_R),0,r_R)}{dr_R} = \left(1+\frac{dp^*(r_R)}{dr_R}\right)\left[w-c-r_R\frac{w-v}{p^*(r_R)+r_R-v}\right]\nabla_3 y^* + \frac{dp^*(r_R)}{dr_R}\left\{\left[w-c-r_R\frac{w-v}{p^*(r_R)+r_R-v}\right](\nabla_1 y^*-\nabla_3 y^*) - r_R\int_0^{\frac{p^*(r_R)+r_R-w}{p^*(r_R)+r_R-v}}\nabla_1 z\,d\xi\right\}$$

By Lemma 2(c) and Proposition 4(a), the first term above is positive. We show that the second term is also positive, to conclude the proof. Since $\frac{dp^*(r_R)}{dr_R} < 0$, all we need to show is

$$\left[w-c-r_R\frac{w-v}{p^*(r_R)+r_R-v}\right](\nabla_1 y^*-\nabla_3 y^*) - r_R\int_0^{\frac{p^*(r_R)+r_R-w}{p^*(r_R)+r_R-v}}\nabla_1 z\,d\xi < 0. \tag{13.13}$$

Now, for $\xi \le \dfrac{p^*(r_R)+r_R-w}{p^*(r_R)+r_R-v}$,

$$\begin{aligned}\nabla_1 y^*-\nabla_3 y^* &= -\gamma y^*\\ &= \nabla_1 z\left(p^*(r_R),0,\frac{p^*(r_R)+r_R-w}{p+r_R-v}\right)\\ &< \nabla_1 z(p^*(r_R),0,\xi)\end{aligned}$$

The first equality follows from Lemma 2(a) and (c), the second from Lemma 1(a), and the inequality holds because $\xi \le \dfrac{p^*(r_R)+r_R-w}{p^*(r_R)+r_R-v}$. Thus, using $r_R \le w-c$, (13.13) holds. □

Proof of Proposition 6 *Proof of (a)* Note that $p^*(r_C)$ will satisfy $\nabla_1\Pi_R(p^*(r_C), r_C, 0) = 0$ at any given r_C (by Proposition 1). By implicit differentiation of this identity with respect to r_C, we obtain $\frac{dp^*(r_C)}{dr_C} = -\frac{\nabla^2_{12}\Pi_R(p^*(r_C),r_C,0)}{\nabla^2_{11}\Pi_R(p^*(r_C),r_C,0)}$. We know from Lemma 6(d) and (j) that

$$\nabla^2_{11}\Pi_R(p^*(r_C),r_C,0) < 0 \text{ and } \nabla^2_{12}\Pi_R(p^*(r_C),r_C,0) > 0.$$

Therefore, it follows that $\frac{dp^*(r_C)}{dr_C} > 0$. Furthermore, from Lemma 6(d) and (j), we can write:

$$\nabla^2_{12}\Pi_R(p,r_C,r_R)\big|_{\nabla_1\Pi_R=0} = -\alpha\nabla^2_{11}\Pi_R(p,r_C,r_R)\big|_{\nabla_1\Pi_R=0} + \alpha\frac{w-v}{p+r_R-v}\nabla_1 y^* - \alpha\gamma\int_0^{\frac{p+r_R-w}{p+r_R-v}} z d\xi$$

From the equality above, since $\nabla_1 y^* < 0$ when $\nabla_1\Pi_R = 0$ (from Lemma 3) and $\gamma > 0$ [by (A3)], we have $\nabla^2_{12}\Pi_R(p,r_C,r_R)\big|_{\nabla_1\Pi_R=0} < -\alpha\nabla^2_{11}\Pi_R(p,r_C,r_R)\big|_{\nabla_1\Pi_R=0}$. Therefore, we have

$$\nabla_{12}\Pi_R(p^*(r_C),r_C,0) < -\alpha\nabla_{11}\Pi_R(p^*(r_C),r_C,0).$$

This observation yields $\frac{dp^*(r_C)}{dr_C} < \alpha$.

Proof of (b) The existence of the Nash equilibrium follows from Lemma 5(a), Propositions 1 and 2(a). (Note that, in order to apply Lemma 5, we need upper bounds on the decision variables of the retailer and the manufacturer, p and r_C, respectively. We could satisfy this requirement by picking arbitrarily large numbers to bound the feasible choices for p and r_C.) □

Proof of Proposition 7 Throughout the proof, let $p^*(r_C)$ denote the optimal retail price chosen by the retailer at a given r_C when $r_R = 0$. *Proof of (a)* Note that $p_o = p^*(0)$ whereas $\tilde{p} = p^*(\tilde{r}_C)$. Therefore, $\tilde{p} - p_o = \int_0^{\tilde{r}_C} \frac{dp^*(r_C)}{dr_C} dr_C$. By Proposition 6, $0 < \frac{dp^*(r_C)}{dr_C} < \alpha$. The desired result follows.

Proof of (b) Note that $y_o = y^*(p^*(0),0,0)$ whereas $\tilde{y} = y^*(p^*(\tilde{r}_C),\tilde{r}_C,0)$. Now, $\tilde{y} - y_o = \int_0^{\tilde{r}_C} \frac{dy^*(p^*(r_C),r_C,0)}{dr_C} dr_C$. We will conclude the proof if we can show that $\frac{dy^*(p^*(r_C),r_C,0)}{dr_C} > 0$. Note that $\frac{dy^*(p^*(r_C),r_C,0)}{dr_C} = \nabla_2 y^*(p^*(r_C),r_C,0) + \frac{dp^*(r_C)}{dr_C}\nabla_1 y^*(p^*(r_C),r_C,0)$. Since $0 < \frac{dp(r_C)}{dr_C} < \alpha$ by Proposition 6 and $\nabla_1 y^*(p^*(r_C),r_C,0) < 0$ by Lemma 3, we obtain $\frac{dy^*(p^*(r_C),r_C,0)}{dr_C} > \nabla_2 y^*(p^*(r_C),r_C,0) + \alpha\nabla_1 y^*(p^*(r_C),r_C,0)$. Using this last inequality and substituting for $\nabla_1 y^*(p^*(r_C),r_C,0)$ from Lemma 2(a) and for $\nabla_2 y^*(p^*(r_C),r_C,0)$ from Lemma 2(b), we can deduce that $\frac{dy^*(p^*(r_C),r_C,0)}{dr_C} > 0$, which concludes the proof of this part.

Proof of (c) As in the proof of part (c) of Proposition 5, we will show that $\Pi_{SC}(p^*(r_C),r_C,0)$ is increasing in r_C for $r_C \le w - c$ when $\alpha \ge \beta$. The desired result would then follow. Now, the following equalities can be verified by partial differentiation of (13.5) and (13.6):

$$\begin{aligned}\frac{d\Pi_{SC}(p^*(r_C),r_C,0)}{dr_C} &= \frac{dp^*(r_C)}{dr_C}\nabla_1\Pi_{SC}(p^*(r_C),r_C,0)+\nabla_2\Pi_{SC}(p^*(r_C),r_C,0)\\ &= \frac{dp^*(r_C)}{dr_C}\nabla_1\Pi_R(p^*(r_C),r_C,0)\\ &\quad+\left(w-c-\beta r_C\frac{w-v}{p^*(r_C)-v}\right)\left(\nabla_2 y^*+\frac{dp^*(r_C)}{dr_C}\nabla_1 y^*\right)\\ &\quad+\beta r_C\left(\int_0^{\frac{p^*(r_C)-w}{p^*(r_C)-v}}\nabla_2 z d\xi-\frac{dp^*(r_C)}{dr_C}\int_0^{\frac{p^*(r_C)-w}{p^*(r_C)-v}}\nabla_1 z d\xi\right)\\ &\quad+p^*(r_C)\int_0^{\frac{p^*(r_C)-w}{p^*(r_C)-v}}\nabla_2 z d\xi\\ &\quad-\beta\left(\int_0^{\frac{p^*(r_C)-w}{p^*(r_C)-v}} z d\xi+\frac{w-v}{p^*(r_C)-v}y^*\right)\end{aligned}$$

Now, the first term is zero, by definition of $p^*(r_C)$. The second term is positive, because, as in the proof of part (b), $\nabla_2 y^*(p^*(r_C),r_C,0)+\frac{dp^*(r_C)}{dr_C}\nabla_1 y^*(p^*(r_C),r_C,0)>0$, and $r_C\le w-c$. The third term is positive by virtue of Lemma 1(a)–(b) and Proposition 6(a). Using Lemma 6 (a), Lemma 1(a) and (b) and the fact that $\nabla_1\Pi_R(p^*(r_C),r_C,0)=0$ we get that

$$p^*(r_C)\int_0^{\frac{p^*(r_C)-w-v}{p^*(r_C)-v}}\nabla_2 z d\xi=\alpha\left(\int_0^{\frac{p^*(r_C)-w}{p^*(r_C)-v}} z d\xi+\frac{w-v}{p^*(r_C)-v}y^*\right).$$

Thus, the sum of the last two terms can be written as

$$(\alpha-\beta)\left(\int_0^{\frac{p^*(r_C)-w}{p^*(r_C)-v}} z d\xi+\frac{w-v}{p^*(r_C)-v}y^*\right),$$

which is positive because $\alpha\ge\beta$.

Proof of (d) The proof of this part is almost identical to the analogous result in Proposition 3. Set $\tilde{r}_R=0$ and the proof follows the same line of argument. □

References

Arcelus, F. J., Kumar, S., & Srinivasan, G. (2012). The effectiveness of manufacturer vs. retailer rebates within a newsvendor framework. *European Journal of Operational Research, 219*, 252–263.

Aydin, G., & Porteus, E. L. (2008). Joint inventory and pricing decisions for an assortment. *Operations Research, 56*, 1247–1255.

Barlow, R. E., & Proschan, F. (1965). *Mathematical theory of reliability*. New York: Wiley.

Bulkeley, W. M. (1998, February 10). 'Rebates' secret appeal to manufacturers: Few consumers actually redeem them. *The Wall Street Journal*.

Cachon, G., & Netessine, S. (2004). Game theory in supply chain analysis. In D. Simchi-Levi, D. Wu, & Z.-J. Shen (Eds.), *Handbook of quantitative supply chain analysis: Modeling in the ebusiness era*. New York: Springer.

Chen, X., Li, C.-L., Rhee, B.-D., & Simchi-Levi, D. (2007). The impact of manufacturer rebates on supply chain profits. *Naval Research Logistics, 54*, 667–680.

Cho, S.-H., McCardle, K. F., & Tang, C. S. (2009). Optimal pricing and rebate strategies in a two level supply chain. *Production and Operations Management, 18*, 426–446.

Demirag, O. C., Baysar, O., Keskinocak, P., & Swann, J. L. (2010). The effects of customer rebates and retailer incentives on a manufacturer's profits and sales. *Naval Research Logistics (NRL), 57*, 88–108.

Demirag, O. C., Chen, Y., & Li, J. (2011a). Customer and retailer rebates under risk aversion. *International Journal of Production Economics, 133*, 736–750.

Demirag, O. C., Keskinocak, P., & Swann, J. L. (2011b). Customer rebates and retailer incentives in the presence of competition and price discrimination. *European Journal of Operational Research, 215*, 268–280.

Dreze, X., & Bell, D. R. (2003). Creating win-win trade promotions: Theory and empirical analysis of scan-back trade deals. *Marketing Science, 22*, 16–39.

Geng, Q., & Mallik, S. (2011). Joint mail-in rebate decisions in supply chains under demand uncertainty. *Production and Operations Management, 20*, 587–602.

Gerstner, E., & Hess, J. D. (1991). A theory of channel price promotions. *The American Economic Review, 81*, 872–886.

Gerstner, E., & Hess, J. D. (1995). Pull promotions and channel coordination. *Marketing Science, 14*, 43–60.

Kalyanam, K. (1996). Pricing decisions under demand uncertainty: A bayesian mixture model approach. *Marketing Science, 15*, 207–221.

Khouja, M., & Zhou, J. (2010). The effect of delayed incentives on supply chain profits and consumer surplus. *Production and Operations Management, 19*, 172–197.

Krishnan, H., Kapuscinski, R., & Butz, D. A. (2004). Coordinating contracts for decentralized channels with retailer promotional effort. *Management Science, 50*, 48–63.

Lal, R. (1990). Price promotions: Limiting competitive encroachment. *Marketing Science, 9*, 247–262.

Lal, R., Little, J. D. C., & Villas-Boas, J. M. (1996). A theory of forward buying, merchandising and trade deals. *Marketing Science, 15*, 21–37.

Menzies, D. (2005, September 12). Mail-in rebates rip. *Marketing*.

Millman, H. (2003, April 17). Customers tire of excuses for rebates that never arrive. *The New York Times*.

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

Olenick, D. (2002, October 14). emachines drops rebate program. *TWICE*.

Palmer, K. (2008, January 18). Why shoppers love to hate rebates. *US News and World Report*.

Petruzzi, N. C., & Dada, M. (1999). Pricing and the newsvendor problem: A review with extensions. *Operations Research, 47*, 183–194.

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

Raju, J. S., Srinivasan, V., & Lal, R. (1990). The effects of brand loyalty on competitive price promotional strategies. *Management Science, 36*, 276–304.

Rao, R. C. (1991). Pricing and promotions in asymmetric duopolies. *Marketing Science, 10*, 131–144.

Ricadela, A., & Koenig, S. (1998, September). Rebates' pull is divided by hard and soft lines. *Computer Retail Week*.

Taylor, T. A. (2002). Supply chain coordination under channel rebates with sales effort effects. *Management Science, 48*, 992–1007.

Yang, S., Munson, C. L., & Chen, B. (2010). Using msrp to enhance the ability of rebates to control distribution channels. *European Journal of Operational Research, 205*, 127–135.

Chapter 14
Clearance Pricing in Retail Chains

Stephen A. Smith

1 Introduction

1.1 Background

As an application of management science, retail clearance pricing has been an outstanding success, Pilot studies conducted in the 1990s (Smith and Achabal 1998), found that installing a computer based clearance pricing algorithm at a major retail chain resulted in 10–15 % increases in the revenue capture rate during the clearance period. Increases in sell-through and shorter markdown cycle times also freed up capital and floor space for the retailer's follow-on products. Similar revenue gains during the clearance period have been achieved by commercially offered clearance markdown systems (Merrick 2001). Spotlight Systems, Inc.,[1] a seller of clearance markdown software systems, reported in 2002 that the average gain in gross margin dollars for the department and specialty stores that had implemented their system amounted to about 4 % of revenue, or $40 million for every $1 billion of sales. Since U.S. department store sales now exceed $500 Billion per year, there is a very large potential dollar impact, if similar results can be obtained across the industry. More recently, Caro and Gallien (2012) reported that a system that was implemented at a major Spanish retailer (Zara) resulted in a 6 % increase in clearance sales revenue relative to the previous manual system based on managerial judgment. Major vendors of ERP systems are now making price

[1] Spotlight Systems was acquired by Profit Logic, Inc. in 2003, which was in turn acquired by Oracle Corporation in 2005.

S.A. Smith (✉)
Department of Operations Management and Information Systems, Leavey School of Business, Santa Clara University, 500 El Camino Real, Santa Clara, CA 95053, USA
e-mail: ssmith@scu.edu

N. Agrawal, S.A. Smith (eds.), *Retail Supply Chain Management*,
International Series in Operations Research & Management Science 223,
DOI 10.1007/978-1-4899-7562-1_14

optimization a cornerstone of their retail applications suites (Sullivan 2005). This background section discusses why clearance pricing is such an attractive application for retailers and what has allowed it to be successfully implemented through computer based models.

1.2 Trends in Retail Pricing

Retail department and specialty stores are selling an ever increasing fraction of their merchandise on markdowns, which now account for over one third of all sales.[2] This is a result of four general trends in these retailers' merchandising strategies:

1. More products in the assortment
2. A greater proportion of "fashion" merchandise
3. Shorter seasons and
4. More private label (store brand) merchandise.

While these trends give customers a wider selection of product choices and are essential for retailers to remain competitive, they also increase the difficulty of managing the retail supply chain. Fashion and private label items tend to have long lead times for orders from the manufacturer and the total order quantity for the season is usually fixed in advance. This decision is based on the initial sales forecasts, which tend to be inaccurate for fashion and seasonal merchandise. Also, well over half of the retailer's total order for seasonal and fashion items is usually sent to the stores at the start of the season to create an attractive presentation of the merchandise. Since inter-store transfers are often not economical, it is difficult to rebalance this inventory if the initial allocation is incorrect. When sales in a given category or group of items are lower than expected, retailers must find a way to clear the excess merchandise to make way for the new product arrivals of the coming season. The cycle time for this process becomes shorter still for "fast fashion" retailers who use very short seasons. [See, e.g., Caro and Gallien 2012.]

Clearance pricing involves two decisions: when to start clearance markdowns and how "deep" the markdowns should be, both of which depend on the remaining inventory. Traditionally, these decisions have been made by the buyer who originally chose the merchandise and ordered it from the manufacturer. This may create a disincentive for taking markdowns early enough, since an early decision to mark down really amounts to admitting that the product has underperformed. For seasonal items such as swimsuits and winter coats, demand decreases rapidly near the end of the season; thus delaying a markdown can be very costly. For simplicity, buyers have traditionally taken the same markdown at all stores, or for all stores within a region. This is suboptimal when there are significant inventory imbalances across stores. These factors tend to make clearance markdowns a very complex

[2] National Retail Federation data for Department and Specialty Stores.

decision that buyers would be happy to delegate to a computer based pricing algorithm. At the same time, retail managers require a clear demonstration of the "payback," i.e., the return on investment, for any newly implemented system. Thus, any clearance markdown pricing system needs to be able to pay for itself through improvements in gross margin dollars during the clearance period.

The computing resources necessary for clearance pricing have only recently been available to retailers. As late as the 1990s many retailers did not retain store level item sales figures for more than 90 days and sales results were often reported only in dollars of revenue. Often, there were no detailed records of how many units of each item were sold at a given price. The economics of data storage tended to be the deciding factor in these decisions, because a department store retailer with 100,000 SKUs and 1,000 stores simply could not afford to store all this transaction data for all time periods. Computing resources were also limited among retail staff members, because of the high costs of training and support. Since retail staff members tend to change job assignments frequently, it is important to standardize and document all decision making procedures, and to make the results easily understandable by retail personnel who are not technically trained in using computers. The exponential decline in the cost of data storage and the growth in popularity of personal computers that occurred during the 1990's have removed these barriers to implementing computer based clearance pricing algorithms.

1.3 Mathematical Models for Clearance Pricing

An analytical approach to clearance markdown management requires the successful implementation of three system components:

1. A sales forecasting model
2. A clearance price optimization algorithm that works at the store and item level
3. Financial performance measurement of the effectiveness of the system

This section discusses a number of the models in the literature that relate to these components of the clearance pricing system.

The modeling assumptions in this paper were motivated by discussions with buyers who manage clearance markdowns at several retail department and specialty store chains. The author also assisted three major retailers in designing computer-based systems that incorporated these models. One unique aspect of this chapter's pricing model is that sales depend explicitly on the retailer's on-hand inventory. The pricing analysis implies that when the rate of sale is sensitive to the inventory level, it is optimal to have higher prices early in the season, followed by deeper markdowns later in the clearance period. Furthermore, inventory sensitivity in the demand makes it optimal to have some amount of leftover merchandise at the end of the clearance period. This leftover inventory, which is typically found in department store chains, may be sold to a discounter, transferred to other channels operated by the retailer or possibly donated to charity. Many retailers recognize the

advantage of setting clearance prices at the store level to account for the variation in inventory levels and sales rates across stores. Due to the complexity and time consuming nature of localized pricing, computer-based clearance pricing algorithms are required to implement these store level markdown decisions.

2 Related Research

In general, optimal clearance pricing for retailers involves some type of dynamic pricing. Surveys on dynamic pricing policies appear in papers by Elmaghraby and Keskinocak (2003) and by Bitran and Caldenty (2003), and are also included in the monograph by Talluri and van Ryzin (2004). The surveyed papers include a variety of factors such as seasonally varying or declining demand, varying customer response to price changes, demand uncertainty, inventory dependent demand and simultaneous pricing and inventory decisions. Since no tractable model can incorporate all of these factors simultaneously, the choice of modeling assumptions requires tradeoffs. The literature summary below focuses on specific subsets of the pricing literature in marketing, economics and inventory management that are relevant to the retail clearance pricing application.

Intertemporal pricing issues similar to those found in clearance markdowns are studied in a deterministic setting by Stokey (1979), Kalish (1983), Dhebar and Oren (1985), Rajan et al. (1992), Braden and Oren (1994). Stokey's analysis considered a family of customer utility functions that decline with time and identified conditions under which the optimal price trajectory is constant or decreasing. Kalish (1983) considered sales rates that vary with both price and cumulative sales-to-date and obtained conditions on sales rate and production cost that determine whether the optimal price trajectories are increasing or decreasing. Dhebar and Oren (1985) determined the optimal price trajectory when there is a positive network externality and decreasing supply cost. Khmelnitsky and Gerchak (2002) applied an optimal control model to a production system in which demand is positively influenced by inventory level, but with a predetermined constant price. The other two papers are discussed below.

Demand uncertainty has been included in dynamic pricing models in a variety of ways. Lazear (1986) and Pashigian (1988) considered clearance markdowns for a single item sold to heterogeneous customers who have a time invariant probability distribution of reservation prices. Gallego and van Ryzin (1994) developed a continuous time optimal pricing model in which demand is generated by Poisson arrivals. Feng and Gallego (1995) develop a continuous time Markov process formulation with stochastic demand that determines the optimal timing and duration of a single price reduction. Bitran et al. (1998), Bitran and Mondschein (1997) and Zhao and Zheng (2000) generalize this by modeling customer demand as Poisson arrivals whose reservation prices change over time. The net result is a nonhomogeneous Poisson process multiplied by a price sensitivity function. While these models capture demand uncertainty, they do not include the influence of

inventory level on demand, which we found was often significant in retail sales. Significant effects of inventory levels on retail sales have been found by Wolfe (1968), Bhat (1985), Smith and Achabal (1998) and Caro and Gallien (2012).

Learning can play a role in dynamic pricing for either the buyer or the seller. Lazear (1986) allowed the seller to infer customers' reservation prices through their responses to a decreasing sequence of discrete prices. Braden and Oren (1994) derive an optimal nonlinear price structure that improves the seller's information about the distribution of heterogeneous customers' price sensitivities. Lariviere and Porteus (1999) considered a multi-period pricing and inventory model with learning, in which the seller uses varying inventory levels as opposed to price changes to obtain information.

The impact of strategic customers on retail pricing decisions has also been analyzed in a variety of contexts. Besanko and Winston (1990) investigated the role of customers' knowledge of future prices in intertemporal pricing. Cachon and Swinney (2009) consider the impact of strategic customers on the retailer's purchasing and pricing decisions.

The marketing literature on price promotions provides a number of empirically tested functional forms for price response. (See e.g., Gaur and Fisher 2005.) This paper adopts a multiplicative form with exponential price sensitivity, which has been analyzed and empirically tested by Narasimhan (1984), Russell and Bolton (1988), Bolton (1989), Achabal et al. (1990), Smith et al. (1994) and Kalyanam (1996). Exponential sensitivity is also applicable for modeling how price influences purchases of consumer durables; Kalish (1985) compared several variations.

There are a number of related papers that develop combined strategies for pricing and inventory management. Eliashberg and Steinberg (1987) considered pricing, inventory and production management policies for a marketing channel subject to seasonal variations. Rajan et al. (1992) considered dynamic pricing and inventory decisions with a variable time horizon and shrinkage costs. Bitran et al. (1998) consider the coordination of prices and inventories across multiple retail outlets in which there are initial allocations of inventories and a further reallocation to rebalance inventories in response to sales. This formulation includes many of the aspects of retail markdown pricing, but the result is a dynamic programming problem with such a large state space that it is likely to be intractable. The authors propose and test some myopic heuristics for approximate solutions. Mantrala and Rao (2001) discuss a decision support system called MARK, which determines discrete prices and inventory levels based on a time varying elasticity demand model. Monahan et al. (2004) analyze a newsvendor model with combined pricing and inventory decisions at discrete time points. Cheng and Sethi (1999) develop a Markov decision model to determine promotion and inventory decisions in a discrete periodic review system. Ray et al. (2005) develop a combined pricing and inventory management model for a two echelon serial supply chain using a demand function with an additive uncertainty term and random delivery times. Netessine (2004) models price and inventory changes at discrete time points, considering the optimization of both prices and the discrete timing of the price changes. Caro and Gallien (2012) consider a clearance markdown model that

incorporates inventory effects, discrete price choices and groupings of similar items to facilitate clearance management. They also give a detailed description of a successful estimation and implementation of the model at Zara.

In summary, the model in this chapter differs from those discussed above in that it combines seasonal variations and demand dependence on inventory level with a price trajectory optimization based on optimal control theory. At the same time, this paper's model requires the time horizon to be fixed, and ignores time dependent inventory costs and discounting. It allows a single inventory level adjustment, while a number of the previous papers on combined dynamic policies consider more general inventory strategies. Also, this chapter's pricing model does not explicitly include demand uncertainty. However, the updating of the clearance price at discrete time points, as discussed in the last section, provides an approximate myopic solution to the dynamic pricing problem with demand changes. Also, the deterministic optimization formulation allows a closed form pricing solution to be obtained from optimal control theory. For the retail clearance markdown application, it appears that these modeling assumptions are a good compromise that results in a workable clearance pricing model.

This chapter extends the specific results in Smith and Achabal (1998) in several ways. First, it discusses the highly successful application results that have been achieved by commercially available clearance markdown systems since the publication of the original paper. Second, it extends the earlier model to obtain FONC and approximate solutions for the case in which prices change only at pre-assigned discrete time points. An approximate discrete pricing solution is developed, and the continuous solution is used to obtain bounds on the maximum error associated with the approximation. Finally, it obtains closed form expressions for the maximum profit function and presents illustrative numerical analyses for the discrete pricing case.

3 Model Specifications and Optimality Conditions

In developing a decision making framework for clearance markdowns, it is important to note three ways in which clearance prices differ from other types of retail pricing decisions: (1) clearance markdowns are permanent, i.e., prices are not permitted to increase later, (2) demand tends to decrease at the end of the clearance period due to items becoming "out of season," as well as incomplete assortments and reduced merchandise selection, (3) optimal clearance prices typically differ by location due to inventory imbalances.

Motivated by these observations, the modeling assumptions are as follows:

- Sales rate depends explicitly on price, seasonal variations and inventory level.
- Competition, demand uncertainty, time discounting and time dependent holding costs are not explicitly included in the model.

These modeling choices can be explained as follows. Price dependence specifies the change in sales rate as a function of the percentage markdown. Seasonal variations capture the increase in sales rate that tends to occur during certain prime shopping periods such as Christmas and back-to-school, and the decrease that occurs at the end of the product's season. When the on-hand inventory is too low at a given store, the sales rate may also drop. This is especially true for apparel when there is an incomplete selection of sizes and colors. Additionally, for some items, it is important to have sufficient inventory to create an attractive in-store display to draw customers' attention to the product.

Retailers tend to intentionally schedule larger deliveries during periods with high sales forecasts, e.g., during promotions. In analyzing the corresponding sales data after the fact, this may sometimes seem to imply a false "causality," in that the higher sales during promotions should not be attributed to higher inventories, even though a positive correlation exists. On the other hand, most buyers seem to feel that low inventories do reduce sales, which was supported by our regression results. Retailers often define a minimum on-hand inventory for each product, sometimes called "fixture fill," which is the quantity required for adequate presentation. This is used as a reference level in defining the inventory effect in the model.

Competition and demand uncertainty are not explicitly captured in the sales rate model. However, sales lost to competitors are implicitly reflected in the retailer's seasonally adjusted rate of sale. This is appropriate as long as the competitors do not react directly to the retailer's price changes. For clearance markdowns taken at the store level, competitive reactions seem unlikely, given that most retail chains have hundreds of stores, each with different local competitive environments.

Demand uncertainty clearly exists, but modeling it complicates the analysis to a great extent. Optimal clearance pricing in the presence of gradually decreasing demand uncertainty would require multistage pricing decisions, which would need to be jointly optimized by stochastic dynamic programming. The state space for this problem is extremely large, because it must capture all the possible changes in the states of information that influence each update of the pricing policy. Because the clearance period is relatively short and sales rates are declining, the early clearance markdowns tend to be the dominant decisions economically, thus reducing the importance of multi-stage optimization. The short clearance period also justifies the lack of time discounting and time dependent inventory costs in this model. We therefore develop a deterministic pricing formulation without discounting.

3.1 Model Formulation

The model is specified as a continuous function of time with the following parameters

t_0 = current time of the season
t_e = end of the season, sometimes known as the "outdate"

t = an arbitrary time $t_0 \le t \le t_e$
I_0 = on hand inventory at time t_0
$p(t)$ = price trajectory at time
$s(t)$ = cumulative sales from time t_0 to time t
$I(t) = I_0 - s(t)$ = the on-hand inventory at time t
s_e = total units sold by the outdate t_e
$x(p,I,t)$ = the sales rate at time t, with price p and on-hand inventory I.
c_e = salvage value per unit at the end of the season
$c(I_0)$ = cost of adjusting I_0, if changes are permitted
$R(I_0)$ = total revenue obtained from the I_0 units

The total sales $s(t)$ up to time t clearly satisfies

$$s(t) = I_0 - I(t) = \int_{t_0}^{t} x(p(\tau), I(\tau), \tau) d\tau, \tag{14.1}$$

which implies the differential equation

$$I'(t) = -x(p(t), I(t), t) \quad \text{for each } t. \tag{14.2}$$

It is also required that $s_e \le I_0$, where the unsold units $I_0 - s_e = I(t_e)$ are salvaged.

In general, the retailer's objective is to maximize total revenue during the clearance period, since the cost of ordering I_0 is a sunk cost. However, changes in I_0 with costs captured by the function $c(I_0)$ may be permitted in some cases. The net profit can then be expressed as:

$$R(I_0) - c(I_0) = \int_{t_0}^{t_e} p(t)x(p(t), I(t), t)dt + c_e(I_0 - s_e) - c(I_0),$$

$$\text{subject to } I_0 \ge s_e = \int_{t_0}^{t_e} x(p(t), I(t), t)dt. \tag{14.3}$$

This objective function can be optimized using optimal control methods, as discussed in detail in Smith and Achabal (1998). These results will be summarized below and then extended to develop exact and approximate solutions for the discrete pricing case.

First order necessary conditions (FONC) for maximizing (14.3) with respect to $p(t)$, subject to the stated constraints can be obtained by forming the Hamiltonian $H = (p - \lambda)x$ and treating $I(t)$ as the state variable and $p(t)$ as the control (see, e.g., Kamien and Schwartz 1981, pp. 143–8). The Lagrange multipliers are

θ = the Lagrange multiplier for the constraint $I_0 - s_e \ge 0$
$\lambda(t)$ = the Lagrange multiplier for $I'(t) = -x(p(t),I(t),t)$ at time t.

The FONC for the optimal control $p(t)$ and the corresponding state variable $I(t)$ are[3]:

$$\partial H/\partial I = [p - \lambda]x_I = -\lambda' \quad \partial H/\partial p = [p - \lambda]x_p + x = 0, \tag{14.4}$$

with the boundary condition

$$\lambda(t_e) = c_e + \theta. \tag{14.5}$$

Eliminating $p - \lambda$ from the two partial derivative equations gives

$$\lambda' = xx_I/x_p \qquad \text{and} \qquad p + x/x_p = \lambda. \tag{14.6}$$

Evaluating (14.6) at $t = t_e$ and combining with (14.5) yields the boundary condition for θ

$$\left(p + x/x_p\right)_{t=t_e} = c_e + \theta. \tag{14.7}$$

3.1.1 The Separable Sales Rate Case

Specific assumptions concerning the functional form of the sales rate allow (14.6) and (14.7) to be solved explicitly for the optimal price trajectory. For this paper, a multiplicative, separable function with exponential price sensitivity is assumed,

$$x(p, I, t) = k(t)y(I)e^{-\gamma p}, \tag{14.8}$$

where $k(t)$ = the seasonal demand at time t

$y(I)$ = the inventory effect when on-hand inventory is I
γ = the price sensitivity parameter for demand.

Although much of this paper's development can be carried through for a more general demand function, a closed form solution can be obtained only for a separable demand function like (14.8). A slightly different closed form solution can also be obtained for constant elasticity price dependence of the form $p^{-\gamma}$. Both exponential price sensitivity and constant elasticity demand functions have been widely studied in marketing. These have generally been found to be superior to linear price sensitivity in empirical studies. [See, e.g., Kalyanam (1996) and Smith et al. (1994) for references.]

For the separable form (14.8), we have that $x/x_p = -1/\gamma$ is a constant. From (14.6), it therefore follows that $p'(t) = \lambda'(t)$. Thus, (14.6) yields an ordinary differential equation that can be solved for $p(t)$

[3] Subscripts p and I denote partial derivatives and the independent variable t has been suppressed for notational compactness.

$$p'(t) = xx_I/x_p = -\frac{1}{\gamma}k(t)y'(I(t))e^{-\gamma p(t)}. \tag{14.9}$$

Mathematically similar formulations have been studied in other contexts. Kalish (1983), Dhebar and Oren (1985) and Mahajan et al. (1990) developed formulations that are sensitive to experience effects rather than inventory, which lead to similar necessary conditions for the optimal price trajectories. Rajan et al. (1992) obtained optimal price solutions for a separable demand form that is analogous to (14.8), but with a time varying γ. Gallego and van Ryzin (1994) obtained an optimal price trajectory for the case of exponential price sensitivity and Poisson demand arrivals. These formulations do not consider the dependence of sales on the current inventory level or seasonal variations, however.

Rajan et al. allow a variable cycle length and they explicitly consider shrinkage and other inventory costs. They obtain closed form optimal price trajectories for the cases of linear and exponential price sensitivities. Variable cycle length is used for clearance pricing of some discontinued non-seasonal items, but seasonal items, which constitute the bulk of retail clearance items, have a fixed clearance calendar to coincide with the planned arrival of new merchandise.

3.1.2 Compensating Prices

Equation (14.9) can be solved by proving that the optimal $p(t)$ adjusts the sales rate so as to exactly compensate for any reduction in sales due to $y(I(t))$. This result is stated as the following lemma.

Lemma 1 *For the multiplicatively separable sales rate function given by (14.8), (14.9) implies that the optimal policy is to adjust* p(t) *so that sales remain proportional to* k(t).

Proof We wish to show that for the optimal $p(t)$

$$\frac{x(p(t), I(t), t)}{k(t)} = y(I(t))e^{-\gamma p(t)} \text{ is constant in } t. \tag{14.10}$$

Suppressing the dependence on t and I and differentiating, we have

$$\frac{d}{dt}(ye^{-\gamma p}) = \left[I'y' - \gamma y p'\right]e^{-\gamma p} = \left[-ky'e^{-\gamma p} - \gamma p'\right]ye^{-\gamma p} = 0,$$

from (14.9), after substituting $I' = -kye^{-\gamma p}$ from (14.2). ■

Lemma 1 implies that the price $p(t)$ at any time t can be expressed in terms of the final price $p(t_e)$ and the ending inventory $I(t_e))$ as follows

$$y(I(t))e^{-\gamma p(t)} = y(I(t_e))e^{-\gamma p(t_e)} \quad \text{for all} \quad t, \tag{14.11}$$

Equation (14.11) also shows that the optimal price depends upon *I(t)* but not upon *t*. Therefore, by defining a new function $P(I(t)) = p(t)$, (14.9) can be solved for the price trajectory as a function of the inventory level

$$P(I) = p(t_e) + \frac{1}{\gamma} \ln\left(\frac{y(I)}{y(I(t_e))}\right). \tag{14.12}$$

The total sales s_e must satisfy from (14.1)

$$\begin{aligned} s_e &= \int_{t_0}^{t_e} k(t)y(I(t))e^{-\gamma p(t)}dt = y(t_e)e^{-\gamma p(t_e)}K, \\ \text{where} \quad K &= K(t_e) = \int_{t_0}^{t_e} k(t)dt. \end{aligned} \tag{14.13}$$

One of two possible cases must hold at time t_e. Either $\theta \geq 0$ and $s_e = I_0$, or $\theta = 0$ and thus $p(t_e) = c_e + 1/\gamma$ from (14.7). If $\theta = 0$, we determine s_e from the relationship

$$s_e = y(I_0 - s_e)Ke^{-\gamma c_e - 1}. \tag{14.14}$$

This has a unique solution since $y(I_o - s_e)$ is decreasing in s_e.

3.1.3 Determining Optimal Inventory and Maximum Profit

We can use the change of variable $I = I(t)$ and the price function *P(I)* to rewrite the integral in the total revenue as

$$\int_{t_0}^{t_e} p(t)x(p(t), I(t), t)dt = \int_{t_0}^{t_e} p(t)\left(-I'(t)\right)dt = \int_{I_0 - s_e}^{I_0} P(I)dI. \tag{14.15}$$

Substituting for *P(I)* from (14.12), we have

$$R(I_0) = s_e p(t_e) + \frac{1}{\gamma}\int_{I_0 - s_e}^{I_0} \ln\left(\frac{y(I)}{y(I_0 - s_e)}\right)dI + c_e(I_0 - s_e). \tag{14.16}$$

This allows us to compute the revenue that will be obtained by using the optimal pricing policy.

Equation (14.16) can also be used to solve for the optimal I_0, if it is a decision variable, subject to the relationships between I_0 and s_e specified above. For the case in which $y(0) = 0$, FONC can be obtained by maximizing $R(I_0) - C(I_0)$ with respect to I_0 and s_e, subject to (14.14). Letting η be the Lagrange multiplier for (14.14), it can be shown that the FONC imply that $\eta = -1/\gamma$ and that

$$p(I_0) = c_e + \frac{1}{\gamma}\left\{1 + \ln\left(\frac{y(I_0)}{y(I_0 - s_e)}\right)\right\} = c'(I_0) + 1/\gamma. \tag{14.17}$$

This can be solved simultaneously with (14.14) to obtain the optimal I_0 and s_e.

Conceptually, it is also possible to use the solution of (14.17) and (14.14) to optimize the initial inventory purchase at the beginning of the season. However, there are practical reasons why this is generally not advisable. Expanding the size of the time interval $[t_0, t_e]$ to include the whole season implies that the same exponential price sensitivity must hold for the demand during the entire time interval. Intuitively, it seems unlikely that this will be true, since price sensitivity may increase or decrease or even require different functional forms during different parts of the season. Thus, it does not seem appropriate to include the original inventory purchase as a decision variable in the context of the clearance pricing model. Smith and Achabal (1998) discuss some adjustments in on-hand inventory that may be possible during the clearance period.

3.1.4 Adding Demand Uncertainty to the Model

Let us consider the case in which demand at time t has a multiplicative uncertainty factor given by the random variable $\xi(t)$. Let us also that there is a common unknown parameter w such that the conditional random variables $\xi(t|w)$ are independent of each other across time. so that assume that the $\xi(t)$ values are independent of each other. Let $\Omega(t)$ be the expected value of $\xi(t)$

4 Discrete Price Changes

In practice, retailers change prices at discrete points in time, rather than continuously. In this section, optimal discrete pricing will be derived and compared to the results for continuous pricing. The discrete pricing case is considerably more complex to solve than the continuous case. However, an approximate discrete solution and error bound can be derived.

An approximate solution for the discrete case can be obtained by choosing prices in each time interval that yield the same unit sales as the continuous case for that time interval. It is shown that the typical revenue losses from this approximation are no more than 1–2 % for two or more price points. The continuous solution is used to

bound the maximum error for the approximate discrete solution, since the exact discrete solution can never be better than the continuous solution.

Suppose the retailer may change prices at n previously set times, e.g., once per week. Let

t_i = time of the ith price change
p_i = price for time period i
$s_i(t)$ = cumulative sales up to time t for $t_{i-1} \le t \le t_i$
$s_i = s_i(t)$ = cumulative sales to the end of period i.

The continuous functions $s_i(t)$, $i = 1, \ldots, n$ satisfy the differential equation

$$s_i'(t) = k(t)y(I_0 - s_i(t))e^{-\gamma p_i} \text{ for } t_{i-1} \le t \le t_i, \tag{14.18}$$

with boundary conditions $s_i(t) = s_i$ for $i = 1, \ldots, n$. The discrete optimization problem is

$$\max_{p_1, \cdots p_n} \sum_{i=1}^{n} p_i(s_i - s_{i-1}), \tag{14.19}$$

subject to (14.18) and its boundary conditions. The variables separate in (14.18), to yield

$$\frac{ds_i}{y(I_0 - s_i)} = e^{-\gamma p_i}k(t)dt. \tag{14.20}$$

The differential equation in (14.20) can be solved for a specific function *y(I)*, if the left hand side can be integrated. The optimization problem can then be solved by a discrete search over the vector of prices $p_1, \ldots, p_n$, subject to the functions $s_i(t)$ obtained from (14.20).

4.1 Solution for the Power Function Form

In this section, we will solve the special case in which the inventory sensitivity follows a power function[4] of the form

$$y(I) = (I/I_r)^{\alpha}, \text{ for a fixed reference value } I_r. \tag{14.21}$$

This form gives considerable flexibility since for various choices of α, it can be either convex, concave or a linear function of the on-hand inventory. This form has $y(0) = 0$,

[4] In Smith and Achabal (1998), additional solution details are given for the general function *y(I)* and numerical analyses are performed for a linear function *y(I)*.

which implies that $\theta = 0$ in (14.7) and s_e is determined from (14.14). Thus, there will be left over inventory to be salvaged at the end of the season in this case. In practice, this occurs for virtually all clearance items. Also, $p(t_e) = p_e = c_e + 1/\gamma$ from (14.7) for $\theta = 0$.

Sometimes in (14.21) the effect of inventory dependence can be truncated at $I = I_r$. This assumes that inventory larger than I_r, does not affect sales. This may often be an appropriate assumption because, as noted previously, higher inventories may sometimes falsely appear to cause higher sales. Thus, whether or not to truncate the inventory effect is really a judgment call, based on the nature of the sales environment that is being analyzed.

For the power function (14.21), the fraction of units sold

$$f_e = s_e / I_0 \tag{14.22}$$

is related to I_0 from (14.14) as follows

$$f_e = \left(\frac{I_0}{I_r}\right)^{-\alpha} \frac{K}{I_0} e^{-\gamma(c_e + 1/\gamma)} (1 - f_e)^{\alpha}. \tag{14.23}$$

The price and total revenue equations then can be written as

$$P(I) = p_e + \frac{\alpha}{\gamma} \ln\left(\frac{I/I_0}{1 - f_e}\right) \tag{14.24}$$

$$R(I_0) = I_0 \left[c_e + f_e/\gamma - \frac{\alpha}{\gamma}\{f_e + \ln(1 - f_e)\}\right]. \tag{14.25}$$

Note that in (14.24) and (14.25) I_r and K do not appear, but f_e depends on I_0/I_r and K/I_0 through (14.23).

Some of the characteristics of these functions can be summarized as follows:

Lemma 2 *The fraction f_e of the inventory sold is decreasing in I_0 for $\alpha < 1$, increasing in I_0 for $\alpha > 1$ and constant for $\alpha = 1$. For $\alpha = 1$, we have*

$$f_e = \frac{a}{1 + a}, \quad \text{where } a = \frac{K e^{-\gamma p_e}}{I_r}. \tag{14.26}$$

Thus, the revenue $R(I_0)$ is linear in I_0 for $\alpha = 1$.

Proof: Taking the total derivative of (14.23) and rearranging terms, we obtain

$$\frac{df_e}{dI_0} = \frac{(1 - f_e)(\alpha - 1)}{I_0^{2-\alpha}(1 - f_e)^{1-\alpha} + \alpha I_0 a}. \tag{14.27}$$

This shows the behavior of f_e, with respect to changes in I_0, based on the term $\alpha - 1$ in the numerator. ■

4.1.1 Optimal Discrete Pricing

The differential equation (14.18) can be solved by integration for the special case (14.21) to obtain

$$-\frac{I_r^{\alpha}\{I_0 - s_i(t)\}^{1-\alpha}}{1-\alpha} = e^{-\gamma p_i}[K(t) - K(t_{i-1})] + Z_i, \tag{14.28}$$

where Z_i is the constant for the function $s_i(t)$ and $K(t)$ is the cumulative seasonal coefficient function from (14.13). At the initial condition $t = t_{i-1}$ in (14.28) $s_i(t_{i-1}) = s_{i-1}$ and we obtain the constant term

$$Z_i = -\frac{I_r^{\alpha}\{I_0 - s_{i-1}\}^{1-\alpha}}{1-\alpha}.$$

Equation (14.28) then acts as a constraint in solving the optimization problem (14.19). No closed form solution can be obtained, but the optimal $p_1, \ldots p_n$ can be determined by numerical methods.

Discrete Pricing to Match the Optimal Continuous Sales

An approximate pricing solution can be obtained by choosing p_i so that the sales in period i match those obtained for the continuous pricing case. That is, we calculate the cumulative sales obtained up to time t_i in the continuous case

$$s_i = y(t_e)e^{-\gamma p_e}K(t_i), \quad \text{for } i = 1, \ldots, n. \tag{14.29}$$

Using this s_i, we determine the corresponding prices by solving the relationships

$$e^{-\gamma p_i}[K(t_i) - K(t_{i-1})] = \frac{\{I_0 - s_{i-1}\}^{1-\alpha} - \{I_0 - s_i\}^{1-\alpha}}{(1-\alpha)I_r^{-\alpha}} \tag{14.30}$$

from (14.28) for $p_1, \ldots p_n$. Here it is convenient to express the p_i in terms of

$f_i =$ the fraction of units sold up to time t_i.

Because of the compensating price property, it follows that

$$f_i = \frac{s_i}{I_0} = f_e \frac{K(t_i)}{K}, \tag{14.31}$$

when the optimal price trajectory $P(I)$ is used. Thus, once f_e is determined from (14.23), the f_i follow immediately from (14.31). Therefore

$$p_i = -\frac{1}{\gamma}\ln\left(\frac{I_0}{1-\alpha}\left(\frac{I_r}{I_0}\right)^{\alpha}\frac{\{1-f_{i-1}\}^{1-\alpha}-\{1-f_i\}^{1-\alpha}}{K(t_i)-K(t_{i-1})}\right). \tag{14.32}$$

The total revenue obtained using this discrete pricing is then given by

$$\overline{R}(I_0) = I_0\left[\sum_{i=1}^{n} p_i[f_i - f_{i-1}] + (1-f_e)c_e\right]. \tag{14.33}$$

Since the maximum revenue $R(I_0)$ obtained with the optimal continuous pricing solution is greater than or equal to the revenue that can be obtained with any discrete solution, it bounds the maximum discrete revenue obtained from (14.30) as well as the revenue obtained with the approximate solution in (14.33). Thus we have proved the following lemma.

Lemma 3 *The percentage profit loss from using the approximate discrete prices obtained from (14.30) in place of the exact discrete price solution obtained from (14.28) is bounded as follows*

$$\text{Profit Loss } \% \;\le\; \frac{R(I_0)-\overline{R}(I_0)}{R(I_0)}. \tag{14.34}$$

Furthermore, the profit loss from using optimal discrete pricing obtained from (14.19) instead of optimal continuous pricing from (14.12) has this same upper bound. It is illustrated in the next section that this percentage loss is less than 1–2 % for typical parameter values.

5 Numerical Examples

In this section, we compute the price trajectories, total sales and total revenue for some parameter values to gain insights about the sensitivity of the results to the various input parameters. We will also compare the continuous and discrete pricing solutions.

To reduce the number of variables, all cases use the values

$I_0 = I_r = 1{,}000$ U, $t_0 = 0$, $t_e = 1$ and $K(t) = tK$.

That is, we assume that there are no seasonal variations and the on hand inventory exactly equals I_r. The solutions can be extended to other I_0 values from (14.24) and (14.25). The time unit scale can be chosen arbitrarily, since all time variations can be expressed as functions of the inventory level I. Solutions are obtained by solving (14.23) for s_e by a one dimensional search, e.g., the Excel Goal Seek function, and then computing the prices and total revenues from (14.24) and (14.25).

Different demand rates can be tested by changing K or by changing the ratio K/I_0. Since K is difficult to interpret intuitively, we define the Base demand parameter

$$\text{Base (demand)} = Ke^{-\gamma p_e}, \tag{14.35}$$

which corresponds to the total unit demand at the minimum price p_e with no inventory effect ($\alpha = 0$). Also note that $p_e = c_e + 1/\gamma$ is the optimal price when $\alpha = 0$ and inventory can be obtained at a unit cost c_e. We will use a retail price of $p_0 = \$10.00$ as a reference value and write all other costs and revenues as multiples of p_0. For these graphs, I_0 is not a decision variable, so $c(I_0)$ is a sunk cost that can be omitted from this numerical analysis.

For the first set of graphs, we use the following parameter values, which represent typical numbers for an apparel item

$c_e = 20\ \%$, $\gamma = 3.33$, $\alpha = 0.5$, 1.0 or 1.5 and Base = 500–1,500.

Let us first consider the total sales $s_e = f_e I_0$ in Fig. 14.1 as a function of the Base values and $\alpha = 0.5$, 1.0 and 1.5. These curves are concave increasing, as one might expect, and the smaller values of α give the largest total sales in every case. This is because the negative effects of inventory on sales are less for smaller values of α.

Now let us consider Fig. 14.2, the optimal price trajectory for the single fixed Base Demand = 1,000. From Fig. 14.1, the total sales for $\alpha = 0.5$, 1.0 and 1.5 are 838, 677 and 578, respectively. Each curve in Fig. 14.2 shows the compensating behavior of the optimal price trajectory, as more inventory is sold. Also, we know from Fig. 14.1 that $\alpha = 1.5$ corresponds to the least total inventory sold. In all cases, it is best to price higher initially and then gradually decrease the price to compensate for the increasing inventory effect, as described by (14.24). The crossing patterns of the price curves in Fig. 14.2 can be explained as follows. We know that $\alpha = 1.5$ must have the steepest drop, because it compensates for the largest inventory effect, while $\alpha = 0.5$ must yield the flattest curve. All curves must have

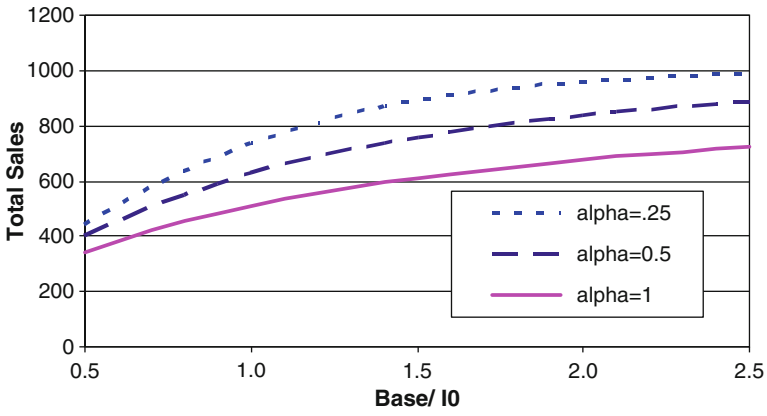


Fig. 14.1 Total sales $f_e I_0$ versus Base/I0 $I_0 = I_r = 1{,}000$

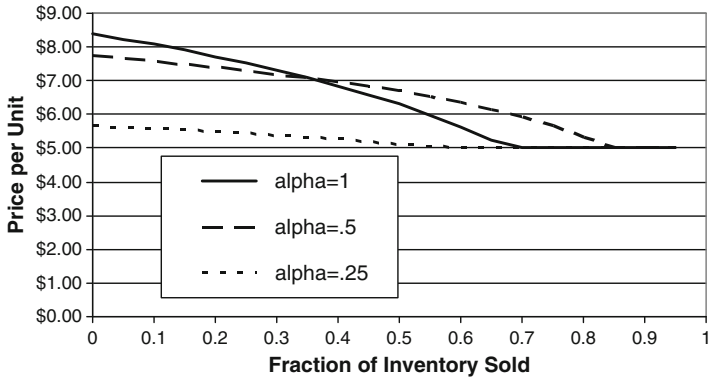


Fig. 14.2 Optimal price trajectories Base/$I_0 = 1.0$

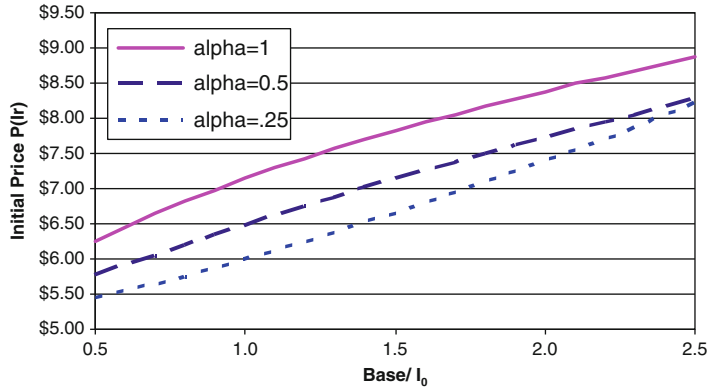


Fig. 14.3 Optimal initial price

the same terminal price p_e. The highest initial price therefore occurs for $\alpha = 1.5$. Figure 14.3 shows the behavior of the optimal initial price $p(I_0)$ for other values of Base Demand.

Figure 14.4 shows the total revenue obtained by using the optimal price trajectory (14.24) in each case. It is interesting to note in Fig. 14.4 that the revenues generated for the three values of α are fairly close to each other. This implies that if inventory effects are modeled correctly, then the almost the same revenue can be obtained through appropriate pricing. For larger α values, higher prices maximize the profit by selling fewer units.

Figure 14.5 shows the bound on the profit loss as a result of approximating the optimal continuous price trajectory with the discrete prices (14.32) that match the continuous sales at the discrete points. That is, the percentage losses in Fig. 14.5 are obtained from (14.34). The other assumptions behind Fig. 14.5 are as follows. For $\alpha \leq 1$, it is intuitively clear that $\alpha = 1$ yields the worst percentage loss, since the

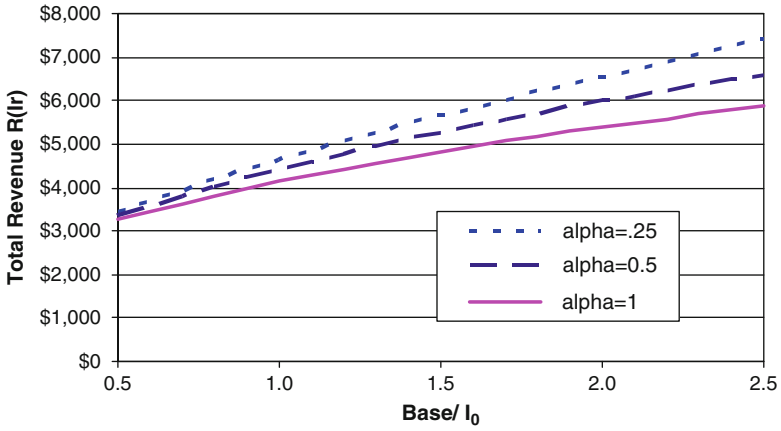


Fig. 14.4 Total revenue versus Base/I_0

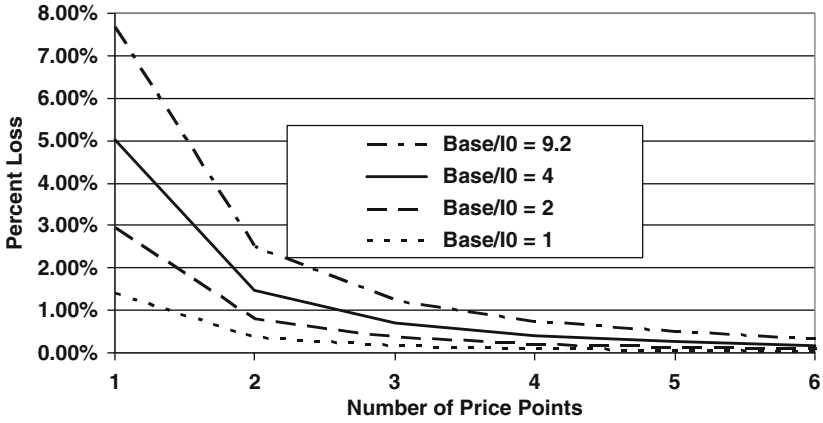


Fig. 14.5 Profit loss bounds for approximate discrete prices for $c_e \geq 0$, $\alpha \leq 1$ & $\gamma > 0$

price drops more rapidly for higher α. This was also verified by extensive calculations. Second, $c_e = 0$ is also the worst case for percentage loss, because with no salvage value the price trajectory drops must achieve all the profits. But with $c_e = 0$, we see that the factor I_0/γ appears in both (14.25) and (14.33), and so I_0/γ cancels out in (14.34). Thus, the curve in Fig. 14.5 holds for all I_0 and γ as well. It is clear from Fig. 14.5 that errors are generally less than 1 or 2 % if at least two price points are used. The worst case occurs for Base/$I_0 = 9.2$, which corresponds to the lowest demand level that requires an optimal price higher than the base price of $p_0 = \$10.00$.

6 Conclusions

Both practical and theoretical insights can be drawn from the experiences with the clearance markdown methodology described in this paper. From a practical standpoint, improvements in clearance markdown policies have had major financial impacts on a number of firms because clearance sales volumes are substantial and any increased revenues from improved clearance policies go directly to the bottom line. Clearance markdown algorithms are now a key component of merchandise pricing for many retail chains, which are part of a sector with sales exceeding $500 billion per year.

The markdown response model in this chapter differs from other dynamic pricing models in that it includes a dependence on inventory level. Retail buyers in the initial studies, particularly for apparel products, felt that having adequate inventory for presentation strongly affects sales. Regression analyses have also found that low inventories are highly correlated with reduced sales. Adopting a multiplicative, exponential price response function, which has previously been successful in modeling the response to promotional markdowns, leads to an optimal clearance price trajectory that exactly compensates for the effects of reduced inventory, independent of the form of the inventory sensitivity.

General properties of the optimal pricing policy for merchandise that is sensitive to inventory level can provide guidelines for developing corporate strategies for these products. Inventory sensitivity implies that prices should be set higher before the clearance period begins, and then reduced gradually during the clearance period. For many products, it is optimal to leave some quantity of merchandise unsold at the end of the season, especially if it has a salvage value. At the same time, our pricing studies indicated that the initial clearance markdowns should be deeper than buyers were accustomed to taking, while excessive markdowns at the end of the season should be avoided in favor of salvaging, or even discarding, unsold merchandise.

One of the implementation requirements is parameter estimation. Smith and Achabal (1998) discuss some regression based approaches for estimating the parameters for sales forecasting and markdown response models. These methods have often been combined with subjective estimation of certain response parameters, or use of seasonal variations that were computed at a higher level of aggregation. While these estimation methods based partially on subjective choices, they have been sufficiently accurate to achieve significant improvements in operating results at a number of retailers.

This model can also provide a basis for further research in pricing policies that include dependence on inventory effects. Possible enhancements, which have been considered in other related research, include time discounted cash flows and time dependent inventory holding costs. When the clearance markdown period is longer, these time dependent aspects become more important. Another interesting generalization is the use of initial clearance prices to elicit information about the customer markdown response parameters. When combined with the sensitivity of

sales to inventory, this remains an unsolved problem to the author's knowledge. Finally, these successful practical applications should encourage others to apply management science models in situations that require a combination of regression analysis and subjective parameters choices.

Acknowledgement The author is especially indebted to Professor Dale Achabal, Director of the Retail Management Institute at Santa Clara University, for initiating the clearance markdown research, and for leading the projects that resulted in the successful implementations of clearance pricing.

References

Achabal, D., McIntyre, S., & Smith, S. (1990, Winter). Maximizing profits from department store promotions. *Journal of Retailing, 66*(4), 383–407.

Besanko, D., & Winston, W. L. (1990, May). Optimal price skimming by a monopolist facing rational consumers. *Management Science, 36*(5), 555–567.

Bhat, R. R. (1985). *Managing the demand for fashion items*. Ann Arbor, MI: UMI Research Press.

Bitran, G. R., & Caldenty, R. (2003). An overview of pricing models for revenue management. *Manufacturing and Service Operations Management, 5*(3), 203–229.

Bitran, G., Caldenty, R., & Mondschein, S. (1998). Coordinating clearance markdown sales of seasonal products in retail chains. *Operations Research, 46*, 609–624.

Bitran, G. R., & Mondschein, S. V. (1997, January). Periodic pricing of seasonal products in retailing. *Management Science, 43*(1), 64–79.

Bolton, R. N. (1989, Spring). The relationship between market characteristics and promotion price elasticities. *Marketing Science, 8*(2), 153–169.

Braden, D. J., & Oren, S. S. (1994, Summer). Nonlinear pricing to produce information. *Marketing Science, 13*(3), 310–326.

Cachon, G., & Swinney, R. (2009). Purchasing, pricing, and quick response in the presence of strategic consumers. *Management Science, 55*(3), 497–511.

Caro, F., & Gallien, J. (2012, Nov–Dec). Clearance price optimization for a fast fashion retailer. *Operations Research, 60*(6), 1404–1422.

Cheng, F., & Sethi, S. P. (1999). A periodic review inventory model with demand influenced by promotions decisions. *Management Science, 45*(11), 1510–1523.

Dhebar, A., & Oren, S. (1985). Optimal dynamic pricing for expanding networks. *Marketing Science, 4*(Fall), 336–351.

Eliashberg, J., & Steinberg, R. (1987, August). Marketing-production decisions in an industrial channel of distribution. *Management Science, 33*, 981–1000.

Elmaghraby, W., & Keskinocak, P. (2003, October). Dynamic pricing in the presence of inventory considerations. *Management Science, 49*(10), 1287–1309.

Feng, Y., & Gallego, G. (1995, August). Optimal starting times for end-of-season sales and optimal stopping times for promotional fares. *Management Science,* 41(8), 1371–1391.

Gallego, G., & van Ryzin, G. (1994, August). Optimal dynamic pricing of inventories with stochastic demand. *Management Science, 40*(8), 999–1020.

Gaur, V., & Fisher, M. (2005). In store experiments to determine the impact of price on sales. *Production and Operations Management, 14*(4), 377–387.

Kalish, S. (1983, Spring). Monopolistic pricing with dynamic demand and production cost. *Marketing Science, 2*(2), 135–159.

Kalish, S. (1985, December). A new product adoption model with price, advertising and uncertainty. *Management Science, 31*(12), 1569–1585.

Kalyanam, K. (1996). Pricing decisions under demand uncertainty: A Bayesian mixture model approach. *Marketing Science, 15*(3), 207–221.

Kamien, M. I., & Schwartz, N. (1981). *Dynamic optimization*. New York: North Holland.

Khmelnitsky, E., & Gerchak, Y. (2002). Optimal control approach to production systems with inventory dependent demand. *IEEE Transactions on Automatic Control, 47*(2), 289–292.

Lariviere, M., & Porteus, E. (1999, March). Stalking information: Bayesian inventory management with unobserved lost sales. *Management Science, 45*(3), 346–363.

Lazear, E. P. (1986, March). Retail pricing and clearance sales. *The American Economic Review, 76*, 14–32.

Mahajan, V., Muller, E., & Bass, F. (1990, January). New product diffusion models in marketing: A review and directions for research. *Journal of Marketing, 54*, 1–26.

Mantrala, M. K., & Rao, S. (2001). A decision support systems that helps retailers decide order quantities and markdowns for fashion goods. *Interfaces, 31*(3), S146–S165.

Merrick, A. (2001, August 7). Priced to move: Retails attempt to get a leg up on markdowns with new software. *Wall Street Journal,* A1, A6.

Monahan, G., Petruzzi, N., & Zhao, W. (2004). The dynamic pricing problem from a newsvendor's perspective. *Manufacturing and Service Operations Management, 6*(1), 73–91.

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

Netessine, S. (2004). Dynamic pricing of inventory/capacity with infrequent price changes. *European Journal of Operations Research, 174*(1), 553–580.

Pashigian, B. P. (1988, December). Demand uncertainty and sales: a study of fashion and markdown pricing. *The American Economic Review, 78*(5), 936–953.

Rajan, A., Rakesh, A., & Steinberg, R. (1992, February). Dynamic pricing and ordering decisions by a monopolist. *Management Science, 38*(2), 240–262.

Ray, S., Li, S., & Song, Y. (2005). Tailored supply chain decision making under price sensitive demand and delivery uncertainty. *Management Science, 51*(12), 1873–1891.

Russell, G. J., & Bolton, R. N. (1988, August). Implications of market structure for elasticity structure. *Journal of Marketing Research, 25*(3), 229–241.

Smith, S. A., Achabal, D., & McIntyre, S. (1994). A two stage sales forecasting procedure using discounted least squares. *Journal of Marketing Research, 31*(February), 44–56.

Smith, S. A., & Achabal, D. D. (1998). Clearance pricing and inventory policies for retail chains. *Management Science, 44*(3), 285–300.

Stokey, N. (1979, August). Inter-temporal price discrimination. *Quarterly Journal of Economics, 93*, 355–371.

Sullivan, L. (2005, November 28). Getting the price right: SAP goes shopping. *Information Week.*

Talluri, K. G., & van Ryzin, G. (2004). *The theory and practice of revenue management*. New York: Springer.

Wolfe, H. B. (1968, Winter). A model for control of style merchandise. *Industrial Management Review* [now Sloan Mgt. Rev.]*, 9*, 69–82.

Zhao, W., & Zheng, Y.-S. (2000). Optimal dynamic pricing for perishable assets with nonhomogeneous demand. *Management Science, 46*(3), 375–388.

Chapter 15
Markdown Competition

Seungjin Whang

1 Introduction

Dynamic price optimization, as a branch of revenue management, investigates the price as a key decision variable in a dynamic business environment. In particular, it studies how to "operationalize" pricing decisions by considering additional dimensions like time and inventories. Perhaps the most canonical example is the groundbreaking work by Gallego and van Ryzin (1993) who study the optimal price trajectory based on the actual realization of sales and the length of remaining sales period. Since then, a wide variety of dynamic pricing models came into existence. In those models, demands may be deterministic or *stochastic* (Gallego and van Ryzin 1993), the set of prices *predetermined* or arbitrary (Feng and Xiao 2000), the number of price changes *limited* or unlimited (Feng and Gallego 1995), time continuous or *discrete* (Dudey 1992), customers *strategic* or myopic (Aviv and Pazgal 2003), the setting of the game *completely known* or revealing over time (Lazear 1986), and sellers *monopolistic* or competing (Belobaba 1987). See Talluri and van Ryzin (2004) or Bitran and Caldency (2003) for an extensive review of the literature.

Competition, although present in almost every real setting, has not received enough attention in the dynamic pricing literature, compared to other aspects. This paper attempts to fill the gap by presenting a stylized model of dynamic markdown competition. We consider two retailers who compete in a market with a fixed level of initial inventory. The initial inventory level is only known to the corresponding retailer, and not to the other. To maximize the profit, each retailer would permanently mark down once at a time of his individual choice. The model assumes deterministic demands, a single chance of price change, and a predetermined set of prices. We consider a two-parameter strategy set where a retailer chooses the timing

S. Whang (✉)
Graduate School of Business, Stanford University, Stanford, CA, USA

N. Agrawal, S.A. Smith (eds.), *Retail Supply Chain Management*,
International Series in Operations Research & Management Science 223,
DOI 10.1007/978-1-4899-7562-1_15

of markdown as a function of the current time, his inventory level, and the other's move so far. We characterize the equilibrium of the game and derive managerial insights.

Dynamic markdown competition—where a retailer marks down as a counter to the competitor's move—is a familiar facet of business practice. Consider, for example, the cut-throat competition in the game device market:

> Microsoft cut the price of its Xbox game console by about a third in the U.S. and Canada and announced a similar price cut for Japan Wednesday. The move had been expected by market watchers and comes on the heels of Sony Computer Entertainment America's price reduction for the PlayStation 2 on Tuesday. Effective immediately, Xbox consoles will cost $199.99 in the U.S., down from $299.99, Microsoft says in a statement. Xbox, Sony's PlayStation 2, and Nintendo's GameCube now all cost about $200 in the U.S. In Japan, where Xbox sales have been sluggish since its launch late February, the Xbox will be cut to $193 from $270 effective May 22, Microsoft says. (Evers 2002).

Our model extracts two elements of the business practice captured in the article—the timing of markdown in response to the competitor's move and based on its own inventory position.

This is not the first research work on dynamic price competition. For example, Dudey (1992) studies a model where two duopolistic firms face multiple customers, one at a time in sequence. For each customer, the two firms simultaneously submit their price quotes, and the customer would take the lower offer so far as the price is lower than her reservation price. Each firm starts with a fixed quantity of inventory, so that the price quote is a function of the time, her own inventory level and the other firm's inventory level, as well as the customer's reservation price. Assuming that both firms have complete information of the game (including the evolution of inventory positions), the paper characterizes the equilibrium strategy of each firm.

Varian (1980) and Lal (1990) interpret price promotions as a mixed equilibrium strategy among competing retailers. Lal (1990), for example, considers three retailers, two national brands and one local brand, in a market consisting of switchers and loyals. Loyals are loyal to their preferred national brand, while switchers always buy the cheapest available. The dilemma facing a national brand is that he cannot extract all the surplus from his loyals *and* win switchers' market segment, too, due to the threat coming from the local brand. Thus, implicit collusion is supported as a non-cooperative equilibrium, where the two national brands take turns lowering the price in the form of promotion. Hence, the regular price extracts loyals' surplus, and the promotional price attracts switchers. In a similar market setting, Rao (1991) also studies two retailers—a national brand and a local brand—competing in promotion. Each firm makes a three-stage sequential decision of regular price, promotion depth and promotion frequency. Two firms simultaneously take actions at each stage, and the outcome of the previous stages is jointly observed before moving on to the next stage. They characterize the equilibrium of the multi-stage, multi-decision game with complete information. In the above line of work the players in this game are allowed to change prices, but not as an *ex-post* counter to the other's decisions.

Netessine and Shumsky (2004) study horizontal competition in which two airlines compete over "overflow passengers." Each airline has a fixed capacity and offers two classes, high-fare and low-fare, of seats at two different prices. Each airline faces a random demand to each class, which is exogenously given. Each airline sets a "booking limit" to the number of low-fare seats, so the overflow customers denied tickets at one airline attempt to purchase tickets at the other airline. The paper investigates the strategy of each airline in choosing the booking limit in this non-cooperative game with complete information.

Our model differs from the above work in that it is set up as a non-cooperative game with incomplete information, and players' strategy is the timing of markdown. The rest of the paper is organized as follows. In Sect. 2 we provide the details of the model. Section 3 analyzes the problem of a monopolistic retailer who would choose the time of markdown in the base model. Section 4 forms the core of the paper where we demonstrate the equilibrium strategies of two duopolistic retailers in choosing the markdown time. The last section concludes with a summary and managerial implications.

2 The Model

Consider a pair of retailers (denoted by $i = 1, 2$) competing in a seasonal or fashion product market. At time 0, each retailer, facing uncertain demand, orders a fixed quantity of the product, based on his individual forecast. The order arrives before the selling season starts. The two retailers are symmetric in terms of market power and cost structure, but may differ in their forecasts and order quantities. The forecast as well as the order quantity is privately known to the respective retailer. The order quantity by one retailer is viewed to the other as a random variable drawn from a common distribution F over $[0, \infty)$. At time 1 the selling season starts, and the demand rate at each possible pair of retail prices is revealed to both retailers. Retailers have no chance to replenish the stock even if they realize the demand is larger than initially forecasted.

In standard microeconomics, the demand function defines the 'total' demand level at each price. It does not capture how the demand materializes across time. To fix this, we introduce a 'demand trajectory' that shows the distribution of demand over time. In the present paper we assume a specific demand trajectory in the form of $e^{-\tau/\beta}$ over time $\tau \in [0, \infty)$, where $\beta(>0)$ is the 'demand rate' defining the demand intensity. Thus, the demand arriving in the time interval $[0, t]$ is here given by $\int_0^t e^{-\tau/\beta} d\tau$ or $\beta[1 - e^{-t/\beta}]$, and the total demand over the entire season is β. This particular demand trajectory assumes that the demand of the product peaks upon its introduction and exponentially declines over time. Even if the selling season is infinitely long in this setup, the exponential decay (with the right choice of β) will ensure that the demand fades away fast in time, thereby approximating the demand pattern of a seasonal or fashion product. Further, note that the demand realization

process has no uncertainties once the demand parameter is revealed. Obviously, it is a strong assumption, but it keeps the analysis tractable. In addition, the deterministic model will serve as an anchor case to stochastic models in developing a heuristic or an upper bound (see Gallego and van Ryzin 1993).

Note that the higher the demand rate β, the slower the demand decays over time and the larger the total demand. β is determined by the prices set by the retailers. Each retailer starts the season with the price set at p_0, but may choose to mark down to $p_1 (< p_0)$ at a time of his individual choice. p_0 and p_1 are prefixed prior to the season. This price change would change the demand rates for both retailers. To simplify the notation, let β_{ij} $(i,j \in \{0,1\})$ denote the demand rates β facing the retailer whose own price is p_i and the other's is p_j. For example, if his price is p_0 and hers is p_1, he faces β_{01} and she faces β_{10} as the demand rate. We assume that $\beta_{10} > \beta_{11} > \beta_{00} > \beta_{01}$. In case he marks down and she does not, for example, his demand rate β_{10} will be the highest of the four cases (due to the combination of a larger market and bigger market share), and hers β_{01} will be the lowest. If both mark down, the demand rate β_{11} facing each retailer falls somewhere between the two extremes, but will be higher than β_{00} the initial demand rate, due to a larger market.

We assume that sales are permanently lost from the market if the retailer visited stocks out. One scenario that supports this assumption is the following: If a potential customer visits a retailer who is out of stock, she will not learn about the existence of the product, so she will not search for it at the other retailer's. More generally, we assume that stockouts at one retailer's do not affect the sales at the other retailer's. This adds another strong assumption that if one stocks out, the current demand intensity continues to hold at the other retailer.

Compared to the existing literature, the present model imposes a series of simplifying assumptions of deterministic demands, a single chance of price change, and a prefixed set of prices. Further, we do not discount cash flow for simplicity, and assume that any unsold items at the end of the season are thrown away at zero salvage value and zero cost. In return, the model highlights the timing of competitive markdowns under asymmetric information (about the initial stock level).

3 The Case of a Monopolistic Retailer

Before we study the case of competition, we first consider a monopolistic retailer who starts the season at price p_0 with the stock level S. Assume that the demand parameter at price p_i is β_i for $i = 0, 1$, where $p_0 > p_1$ and $\beta_0 < \beta_1$. Suppose now that the retailer would choose the time to mark down. The demand trajectory enables us to evaluate the impact of a price change on the season's overall profit to each retailer and to formulate the markdown-timing problem as follows.

$$\max_{t \geq 0} \int_0^t p_0 e^{-\tau/\beta_0} d\tau + \int_t^T p_1 e^{-\tau/\beta_1} d\tau = p_0 \beta_0 \left(1 - e^{-t/\beta_0}\right) + p_1 \beta_1 \left(e^{-t/\beta_1} - e^{-T/\beta_1}\right), \tag{P1}$$

where

$$\beta_0\left(1-e^{-t/\beta_0}\right)+\beta_1\left(e^{-t/\beta_1}-e^{-T/\beta_1}\right)\le S. \tag{15.1}$$

Inequality (15.1) is the capacity constraint that ensures that total sales do not exceed the initial inventory, where T denotes the time of running out of stock. We assume that T can take the value of infinity, which happens when S is large enough.

We form the Lagrangian function:

$$\begin{aligned}\mathcal{L}(t,T,\lambda)=&\,p_0\beta_0\left(1-e^{-t/\beta_0}\right)+p_1\beta_1\left(e^{-t/\beta_1}-e^{-T/\beta_1}\right)\\&-\lambda\left[\beta_0\left(1-e^{-t/\beta_0}\right)+\beta_1\left(e^{-t/\beta_1}-e^{-T/\beta_1}\right)-S\right],\end{aligned} \tag{P2}$$

where λ is the Lagrangian multiplier associated with the capacity constraint. After straightforward manipulation, the Kuhn–Tucker theorem yields the following result.

Theorem 1 *To the monopolistic retailer with a starting inventory S, the optimal time $t^*(S)$ to mark down is given by*

$$t^*(S)=\begin{cases}\infty, & \text{if } S<\beta_0;\\ \dfrac{\beta_0\beta_1}{\beta_1-\beta_0}\ \ln\dfrac{p_0-\lambda(S)}{p_1-\lambda(S)} & \text{if } \beta_0\le S\le S^*;\\ \dfrac{\beta_0\beta_1}{\beta_1-\beta_0}\ \ln\dfrac{p_0}{p_1} & \text{if } S>S^*,\end{cases}$$

where $\lambda(S)$, the (non-negative) Lagrangian multiplier to the capacity constraint, satisfies

$$S=\beta_0\left[1-\left(\frac{p_1-\lambda(S)}{p_0-\lambda(S)}\right)^{\frac{\beta_1}{\beta_1-\beta_0}}\right]+\beta_1\left(\frac{p_1-\lambda(S)}{p_0-\lambda(S)}\right)^{\frac{\beta_0}{\beta_1-\beta_0}}, \tag{15.2}$$

and S^* is the smallest value of S with $\lambda(S)=0$; that is,

$$S^*=\beta_0\left[1-\left(\frac{p_1}{p_0}\right)^{\frac{\beta_1}{\beta_1-\beta_0}}\right]+\beta_1\left(\frac{p_1}{p_0}\right)^{\frac{\beta_0}{\beta_1-\beta_0}}. \tag{15.3}$$

Also, $\beta_0<S^*<\beta_1$.

If the retailer has tight supply, he will never mark down, or equivalently, his optimal markdown time will be infinity. This is because in the absence of cash flow discounting, he has no incentive to mark down if he can sell everything he has even if it takes a long time. The cutoff inventory level is β_0, which is the quantity he can

sell without a markdown. Here the choice of the value ∞ is somewhat arbitrary. To be exact, the solution to (*P*2) in this range of S is $t^*(S) = T^*$, where $T = T^*$ satisfies (15.1) in equality. This means that the retailer marks down at the time he runs out of stock. This is equivalent to the event of no markdown ever (especially as observed by the other retailer if she exists as in later sections), hence comes our choice of infinity. In the other extreme case (i.e., an ample inventory), he cannot sell all he has, so he will maximize his profit by lowering the price at time $\frac{\beta_0\beta_1}{\beta_1 - \beta_0} \ln \frac{p_0}{p_1}$, which remains constant to any retailer whose inventory level is larger than S^*. In the middle range of the inventory, the timing of his markdown will depend on the inventory level. The higher the inventory level, the quicker comes the markdown. In this case, the retailer will time the markdown to sell all his inventory. Loosely speaking, $t^*(S)$ is decreasing in $S \in [0, \infty)$.[1] The monopolist with a high inventory will be more anxious, so he will rush to cut the price to move the volume.

4 Markdown Competition

We now turn to the case of two retailers competing in the choice of markdown timing. The strategy for each retailer is the choice of its markdown time, taking the other retailer's strategy as given. More specifically, retailer $i (= 1, 2)$ (he) will choose the time $\sigma_i(S_i, \mathcal{H}_t)$ to mark down, where σ_i is not only a function of his private inventory level S_i, but also of the history $\mathcal{H}_t$ of the game up until his decision time t. In our model that has assumed away demand uncertainties, the relevant information contained in $\mathcal{H}_t$ is the actions taken by the other retailer j (she) and the current time. The strategy determines in advance what to do in each contingency, as the game evolves and uncertainties are resolved. The strategy will maximize the expected profit at each time point for the rest of the game based on the realized path.

Retailer i's expected profit depends on his own inventory level S_i, as well as retailer j's strategy σ_j that depends on her inventory level S_j. To derive his optimal strategy, retailer i must take into account the uncertainties about S_j to predict her strategy and develop his own strategy. Our equilibrium concept is similar to Bayesian subgame-perfect equilibrium (Kreps 1990). Further, we restrict our attention to 'symmetric' equilibrium in which the two retailers use the same strategy function and play with different arguments.

Now consider the set $\mathcal{S} = \{\tilde{\sigma}(t_a, t_b, \mathcal{H}_t) | 0 \le t_a \le t_b\}$ (or $\{\tilde{\sigma}(t_a, t_b)\}$ for short) of two-parameter strategies for each retailer that operate as follows: "Wait and see if the other retailer marks down; if the latter does before t_b, then mark down either immediately or at t_a, whichever comes later. If the other does not mark down until t_b, then don't wait any longer and mark down before the other." When both retailers

[1] This statement is not mathematically accurate since the function $t^*(S)$ is not well defined in the interval $[0, \beta_0]$, but the meaning is clear in the present context.

play strategies in $\mathcal{S}$, retailer i faces three alternative scenarios depending on retailer j's markdown time τ. τ may fall in one of the three time intervals $I_a := [0, t_a)$, $I_b := [t_a, t_b)$, and $I_c := [t_b, \infty]$. If it falls in I_a, retailer i is not "ready" yet, so he will wait and mark down later at t_a. If in I_b, he will immediately match retailer j's markdown. In I_c, retailer i will move first without further waiting for retailer j's move.

While this strategy set appears to contain a wide set of plausible actions, it is not exhaustive by any means. For example, one can consider a three-parameter strategy like "Wait and see if the other retailer marks down; if the latter does before t_a, then mark down at $t'_a (> t_a)$. If the latter does after t_a but before t_b, then mark down at t_b. If the other does not mark down until t_b, then don't wait any longer and mark down before the other." Clearly, this example, although not so convincing on its own, alludes to an infinite number of possible strategy sets, underscoring the fact that $\mathcal{S}$ is just one of them.

Now retailer i's decision is to find a pair $(t_a^*(S_1), t_b^*(S_1))$, or simply (t_a^*, t_b^*), that determine his optimal strategy in $\mathcal{S}$. To derive t_a^* first, suppose that the game started at time 0, and soon retailer j marked down at time t in I_a. The current demand rate for retailer i is β_{01}, but his markdown decision would change it to β_{11}. We now solve

$$\max_{t_a \geq t} \int_t^{t_a} p_0 e^{-\tau/\beta_{01}} d\tau + \int_{t_a}^{T} p_1 e^{-\tau/\beta_{11}} d\tau = p_0 \beta_{01} \left(1 - e^{-t_a/\beta_{01}}\right) + p_1 \beta_{11} \left(e^{-t_a/\beta_{11}} - e^{-T/\beta_{11}}\right) \tag{P3}$$

subject to

$$\beta_{01} \left(e^{-t/\beta_{01}} - e^{-t_a/\beta_{01}}\right) + \beta_{11} \left(e^{-t_a/\beta_{11}} - e^{-T/\beta_{11}}\right) \leq S_i - \beta_{00} \left(1 - e^{-t/\beta_{00}}\right).$$

After adding a constant $\int_0^t p_0 e^{-\tau/\beta_{01}} d\tau$ to the objective and slight modification of the constraint, we have:

$$\max_{t_a \geq t} \int_0^{t_a} p_0 e^{-\tau/\beta_{01}} d\tau + \int_{t_a}^{T} p_1 e^{-\tau/\beta_{11}} d\tau = p_0 \beta_{01} \left(1 - e^{-t_a/\beta_{01}}\right) + p_1 \beta_{11} \left(e^{-t_a/\beta_{11}} - e^{-T/\beta_{11}}\right) \tag{P3'}$$

subject to

$$\beta_{01} \left(1 - e^{-t_a/\beta_{01}}\right) + \beta_{11} \left(e^{-t_a/\beta_{11}} - e^{-T/\beta_{11}}\right) \leq S_{it},$$

where $S_{it} := S_i - \left[\beta_{00}\left(1 - e^{-t/\beta_{00}}\right) - \beta_{01}\left(1 - e^{-t/\beta_{01}}\right)\right] := S_i - \Delta_t$. It is easy to verify that Δ_t is positive and monotone increasing in t.

This problem has the same structure as (*P*1), with β_0, β_1 and S_i replaced by β_{01}, β_{11} and S_{it}. Hence, we have the following solution from Theorem 1.

$$t_a^*(S_{it}) = \begin{cases} \infty, & \text{if } S_{it} \le \beta_{01}; \\ \dfrac{\beta_{01}\beta_{11}}{\beta_{11}-\beta_{01}} \ln \dfrac{p_0 - \lambda(S_{it})}{p_1 - \lambda(S_{it})}, & \text{if } \beta_{01} < S_{it} < S^\circ; \\ \dfrac{\beta_{01}\beta_{11}}{\beta_{11}-\beta_{01}} \ln \dfrac{p_0}{p_1}, & \text{if } S_{it} \ge S^\circ, \end{cases} \tag{15.4}$$

where $\lambda(S_{it})$, the (non-negative) Lagrangian multiplier to the capacity constraint, satisfies

$$S_{it} = \beta_{01}\left[1 - \left(\frac{p_1 - \lambda(S_{it})}{p_0 - \lambda(S_{it})}\right)^{\frac{\beta_{11}}{\beta_{11}-\beta_{01}}}\right] + \beta_{11}\left(\frac{p_1 - \lambda(S_{it})}{p_0 - \lambda(S_{it})}\right)^{\frac{\beta_{01}}{\beta_{11}-\beta_{01}}}, \tag{15.5}$$

and

$$S^\circ = \beta_{01}\left[1 - \left(\frac{p_1}{p_0}\right)^{\frac{\beta_{11}}{\beta_{11}-\beta_{01}}}\right] + \beta_{11}\left(\frac{p_1}{p_0}\right)^{\frac{\beta_{01}}{\beta_{11}-\beta_{01}}}. \tag{15.6}$$

Also, note that $\beta_{01} < S^\circ < \beta_{11}$.

Suppose now that the time point t_a^* has passed without retailer j's move. The new time interval I_b starts, so retailer i will immediately adopt if the other marks down. But if she does not, retailer i cannot wait forever for her move, so he faces the problem of choosing "the preemptive markdown time" t_b, i.e., the time to stop waiting and mark down first.

To find the optimal t_b^*, we first introduce some notation. For the moment, assume that $t_b^*(\cdot)$ is monotone decreasing. Let $G(\tau)$ denote the probability of the other retailer marking down by time τ, with $\bar{G}(\tau) := 1 - G(\tau)$ and $g(\tau) := G'(\tau)$. Also let $\bar{G}^o(\tau|t)$ denote the probability that retailer j will mark down later than time τ on the condition that she has not marked down until time t; i.e., $\bar{G}^o(\tau|t) := 1 - G^o(\tau|t) = \bar{G}(\tau)/\bar{G}(t)$, for $\tau \ge t$. Let g^o and g respectively denote the probability density (or frequency) function of G^o and G.

At time $t(>t_a^*)$, retailer i will choose t_b^* by solving the following (*P4*):

$$\begin{aligned}
\max_{t_b \geq t} & \int_t^{t_b^-} \left[p_0\beta_{00}\left(e^{-t/\beta_{00}} - e^{-\tau/\beta_{00}}\right) + p_1\beta_{11}\left(e^{-\tau/\beta_{11}} - e^{-T_1(\tau)/\beta_{11}}\right)\right] dG^\circ(\tau|t) \\
& + \left[p_0\beta_{00}\left(e^{-t/\beta_{00}} - e^{-t_b/\beta_{00}}\right) + p_1\beta_{11}\left(e^{-t_b/\beta_{11}} - e^{-T_2/\beta_{11}}\right)\right] g^\circ(t_b|t) \\
& + \int_{t_b^+}^{\infty} \left[p_0\beta_{00}\left(e^{-t/\beta_{00}} - e^{-t_b/\beta_{00}}\right) + p_1\beta_{10}\left(e^{-t_b/\beta_{10}} - e^{-\tau/\beta_{10}}\right)\right. \\
& \left. + p_1\beta_{11}\left(e^{-\tau/\beta_{11}} - e^{-T_3(\tau)/\beta_{11}}\right)\right] dG^\circ(\tau|t) \\
& + \left[p_0\beta_{00}\left(e^{-t/\beta_{00}} - e^{-t_b/\beta_{00}}\right) + p_1\beta_{10}\left(e^{-t_b/\beta_{10}} - e^{-T_4/\beta_{10}}\right)\right] g^\circ(\infty|t), \qquad \text{(P4)}
\end{aligned}$$

subject to the following capacity constraints

$$\begin{gathered}
\beta_{00}(1 - e^{-\tau/\beta_{00}}) + \beta_{11}(e^{-\tau/\beta_{11}} - e^{-T_1(\tau)/\beta_{11}}) \leq S_i, \quad \forall \tau \in [t, t_b) \\
\beta_{00}(1 - e^{-t_b/\beta_{00}}) + \beta_{11}(e^{-t_b/\beta_{11}} - e^{-T_2/\beta_{11}}) \leq S_i \\
\beta_{00}(1 - e^{-t_b/\beta_{00}}) + \beta_{10}(e^{-t_b/\beta_{10}} - e^{-\tau/\beta_{10}}) + \beta_{11}(e^{-\tau/\beta_{11}} - e^{-T_3(\tau)/\beta_{11}}) \leq S_i, \forall \tau \in (t_b, \infty) \\
\beta_{00}(1 - e^{-t_b/\beta_{00}}) + \beta_{10}(e^{-t_b/\beta_{10}} - e^{-T_4/\beta_{10}}) \leq S_i.
\end{gathered}$$

In the above, T_i $(i = 1, 2, 3, 4)$ represents the time to run out of inventory under four different scenarios; $T_1(\tau)$ is the time to run out of stock when both retailers mark down at time $\tau \in [0, \tau_b)$, T_2 when both mark down at t_b, $T_3(\tau)$ when i first marks down at t_b and j follows at $\tau \in (t_b, \infty)$, and T_4 when i first marks down at t_b and j does not follow. The objective function in (*P4*) represents the expected profit to retailer i when he plays $\tilde{\sigma}_i(t_a^*, t_b)$ while retailer j plays $\tilde{\sigma}_j(t_a^*, t_b^*)$.

Note that G can be derived from the distribution of random variables S_j via $t_a^*(\cdot)$ and $t_b^*(\cdot)$, and is a mixed (i.e., continuous and discrete) distribution. Regrettably, (*P4*) is very difficult to solve. One way to tackle the problem is to form a Lagrangian and obtain its saddle point (Luenberger 1969). To derive the equilibrium strategy, we obtain the FOC of the Lagrangian for (*P4*), and then invoke the symmetric equilibrium assumption, so retailer i's choice of t_b should be equal to retailer j's optimal ${t_b}^*$, hence $t_b^{*^{-1}}(t_b) = t_b^{*^{-1}}(t_b^*(S_i)) = S_i$. Then, we have (see the details in the Appendix):

$$
\begin{aligned}
&\left(p_0 e^{-t_b^*/\beta_{00}} - p_1 e^{-t_b^*/\beta_{11}}\right) F(S_i) + p_1\left(e^{-t_b^*/\beta_{11}} - e^{-t_b^*/\beta_{10}}\right) F\left(t_a^{*-1}\left(t_b^*\right) + \Delta_{t_b}\right) \\
&\quad - \lambda_1\left(t_b^*\right)\left[\beta_{00}\left(1 - e^{-t_b^*/\beta_{10}}\right) - \beta_{11}\left(e^{-t_b^*/\beta_{11}} - e^{-T_1\left(t_b^*\right)/\beta_{11}}\right) - S_i\right] \\
&\quad - \lambda_2'\left(t_b^*\right)\left[\beta_{10}\left(1 - e^{-t_b^*/\beta_{10}}\right) - \beta_{11}\left(e^{-t_b^*/\beta_{11}} - e^{-T_2/\beta_{11}}\right) - S_i\right] \\
&\quad - \lambda_2\left(t_b^*\right)\left(e^{-t_b^*/\beta_{10}} - e^{-T_2/\beta_{11}}\right) \\
&\quad + \lambda_3\left(t_b^*\right)\left[\beta_{00}\left(1 - e^{-t_b^*/\beta_{10}}\right) + \beta_{11}\left(e^{-t_b^*/\beta_{11}} - e^{-T_3\left(t_b^*\right)/\beta_{11}}\right)\right] \\
&\quad - \bar{\Lambda}_3\left(t_b^*\right)\left(e^{-t_b^*/\beta_{00}} - e^{-t_b^*/\beta_{10}}\right) + \left(p_0 e^{-t_b^*/\beta_{00}} - p_1 e^{-t_b^*/\beta_{10}}\right) F(\beta_{01}) \\
&\quad + \lambda_4\left(t_b^*\right)\left[\beta_{00}\left(1 - e^{-t_b^*/\beta_{00}}\right) + \beta_{10}\left(e^{-t_b^*/\beta_{10}} - e^{-T_4/\beta_{10}}\right) - S_i\right] = 0.
\end{aligned}
\tag{15.7}
$$

A corner solution to (*P4*) occurs when retailer i has an initial inventory less than β_{01}. He would ultimately sell out even at the regular price, so he would never mark down, or his markdown time will be infinity.

Hence, the following theorem summarizes the equilibrium.

Theorem 2 *Consider the set* $\mathcal{S} = \{\tilde{\sigma}(t_a, t_b, \mathcal{H}_t) \mid 0 \le t_a \le t_b\}$ *of two-parameter strategies for each retailer that operate as follows: "Wait and see if the other retailer marks down; if the latter does before* t_b*, then mark down either immediately or at* t_a*, whichever comes later. If the other does not mark down until* t_b*, then don't wait any longer and mark down before the other." Let*

$$
t_a^*(S_{it}) = \begin{cases} \infty, & \text{if } S_{it} \le \beta_{01}; \\ \dfrac{\beta_{01}\beta_{11}}{\beta_{11} - \beta_{01}} \ln \dfrac{p_0 - \lambda(S_{it})}{p_1 - \lambda(S_{it})}, & \text{if } \beta_{01} < S_{it} < S^\circ; \\ \dfrac{\beta_{01}\beta_{11}}{\beta_{11} - \beta_{01}} \ln \dfrac{p_0}{p_1}, & \text{if } S_{it} \ge S^\circ, \end{cases}
\tag{15.4}
$$

where $\lambda(S_{it})$*, the (non-negative) Lagrangian multiplier to the capacity constraint, satisfies*

$$
S_{it} = \beta_{01}\left[1 - \left(\frac{p_1 - \lambda(S_{it})}{p_0 - \lambda(S_{it})}\right)^{\frac{\beta_{11}}{\beta_{11} - \beta_{01}}}\right] + \beta_{11}\left(\frac{p_1 - \lambda(S_{it})}{p_0 - \lambda(S_{it})}\right)^{\frac{\beta_{01}}{\beta_{11} - \beta_{01}}},
$$

and

$$
S^\circ = \beta_{01}\left[1 - \left(\frac{p_1}{p_0}\right)^{\frac{\beta_{11}}{\beta_{11} - \beta_{01}}}\right] + \beta_{11}\left(\frac{p_1}{p_0}\right)^{\frac{\beta_{01}}{\beta_{11} - \beta_{01}}}.
$$

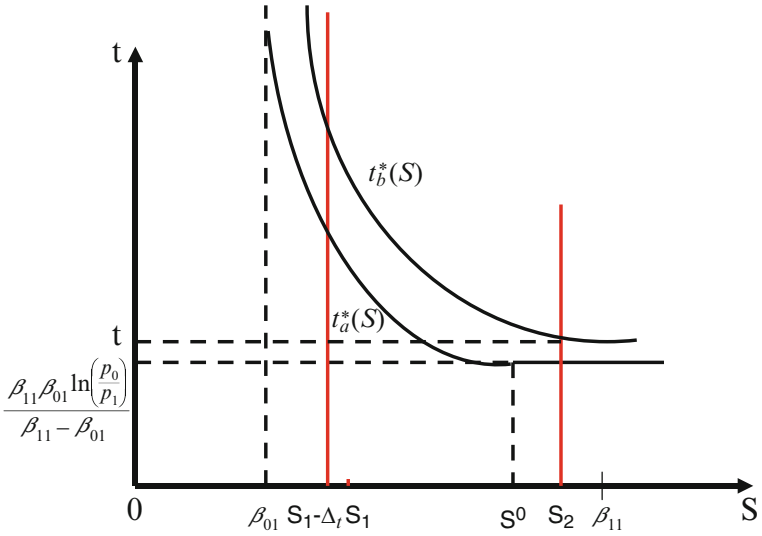


Fig. 15.1 Equilibrium for markdown competition—case 1

And, let

$$t_b^*(S_i) = \begin{cases} \infty, & \text{if } S_i \le \beta_{01}; \\ B^*(S_i), & \text{otherwise}, \end{cases}$$

where $B = B^*(S_i)$ *satisfies (* 15.7 *). If* $B^*(\cdot)$ *is monotone decreasing and* $t_b^*(S) \ge t_a^*(S - \Delta_t)$ *for each t,* $S \in [0,\infty)$*, then* $\tilde{\sigma}_i(t_a^*, t_b^*)$ *forms the equilibrium in* $\mathcal{S}$ *of the markdown game.*

The equilibrium is depicted in Figs. 15.1 and 15.2. Each instance of initial inventory S determines the two parameters $(t_a{}^*, t_b{}^*)$, which in turn define his markdown strategy. As an example, suppose two retailers 1 and 2 start with inventory positions S_1 and S_2, respectively, as shown in Fig. 15.1. At the beginning both retailers sell at the regular price p_0. As time passes (moving up in the Y axis on the figure), retailer 2 with a higher inventory S_2 reaches the time point $t_b^*(S_2)$ and marks down to price p_1. Let $t := t_b^*(S_2)$. Since $t_a^*(S_1 - \Delta_t) > t$ in the figure, retailer 1 does not immediately match the markdown, but instead waits until $t_a^*(S_1 - \Delta_t)$ and marks down. Thus, the two markdowns will be separated by some time. Now consider another pair of retailers that start with inventory levels S_1 and S'_2 as in Fig. 15.2. Again, retailer 2 moves first at time $t_b^*(S'_2) := t'$. But this time $t_a^*(S_1 - \Delta_{t'}) < t'$, so retailer 1 will immediately follow the markdown. This is the case where markdowns are "clustered" around the same time. The first mover disturbs the status quo to the other, who is then forced to take a mitigating action.

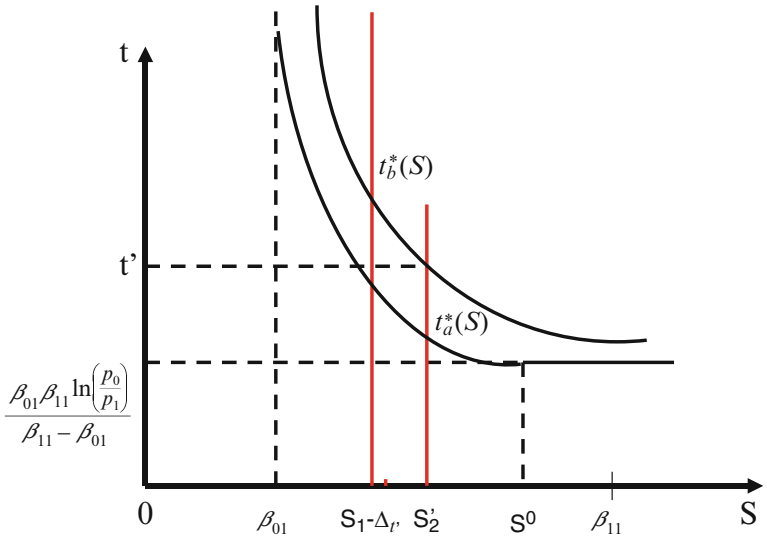


Fig. 15.2 Equilibrium for markdown competition—case 2

5 Managerial Implications and Conclusion

The paper studies how two retailers compete in choosing markdown times. We have restricted our search to a set $\mathcal{S}$ of two-parameter strategies that capture a lot of plausible behaviors. The equilibrium strategy is a function of three elements—the competitor's move so far, the current time (relative to $t_a{}^*$ and $t_b{}^*$), and his own inventory level (captured through $t_a^*(S_i - \Delta_t)$). In our deterministic model, the latter two are overlapping. In equilibrium one retailer's markdown may prompt the other to match instantaneously, especially if the latter has a large inventory and the selling season is almost over (e.g., S_{it} is large and λ is zero). Unfortunately, we could not obtain a closed-form solution to one parameter $t_b{}^*$, but the structure of the solution provides several managerial insights.

First, the markdown policy has a direct impact on its preceding inventory decisions. One can view the inventory and markdown decisions together as a bigger sequential game—first considering the subgame (*P*4) of markdown competition, and then rolling up the solution to the inventory decision. That is, one should solve the following inventory-markdown integrated problem:

$$\max_S \Pi(\tilde{\sigma}\left(t_a^*(S), t_b^*(S)\right) - C(S), \tag{P5}$$

where $\Pi(\tilde{\sigma}\left(t_a^*(S), t_b^*(S)\right)$ is the expected profit under the optimal strategy $\tilde{\sigma}\left(t_a^*(S), t_b^*(S)\right)$ [solving (*P*4)], and $C(S)$ is the cost of procuring S. This may lead to larger or smaller inventory levels than the traditional newsvendor solution, depending on

model parameters. On the one hand, the unit margin or the underage cost is not as high as the newsvendor operation without a markup, so the optimal inventory level will be smaller than the newsvendor solution. On the other hand, however, the demand will be higher at a marked-down price. Hence, the retailer who is willing to mark down if necessary may possibly choose a larger-than-newsvendor inventory level if the markdown still grants the retailer a positive margin.

Second, we anticipate that markdowns will be frequently clustered around a certain time. Note in Fig. 15.2 that clustering happens when the two retailers start with similar levels of inventory. Since their demand signals are likely to be positively correlated or if they have a uni-modal density function like the normal distribution, they will order similar quantities, so clustering of markdowns will be more likely. See Gul and Lundholm (1995) and its references for other instances of clustering.

Third, the present work proposes an alternative model of price dispersion. Economists have long studied various models of price dispersion as a deviation from the traditional "law of one price" (see Varian 1980 and its references). For example, Varian (1980) (plus its Errata, Varian 1981) analyzes the competition among n retailers facing two types of customers—informed and uninformed. Informed customers know the price distribution of a certain item and purchase the item at the store with the lowest price. Uninformed customers randomly choose a store and buy the item there if the price is lower than her reservation price. Each store's strategy is the assignment of probabilities to different prices to charge. Varian demonstrates, among others, that no symmetric equilibrium exists where all stores charge the same price, and even strongly, that there would be no point masses in the equilibrium pricing strategies. Thus, price-randomization is the only equilibrium, hence arises price dispersion. Our model presents another possibility of price dispersion. It differs from Varian in two major ways (besides other differences like permanent vs. temporary price changes, and information asymmetry vs. symmetry). First, the model allows a retailer to choose a dynamic strategy of taking, or not taking, an action upon observing the other's move, while each retailer in Varian sets a price randomly drawn from a pre-determined density function. The difference boils down to whether retailers can monitor each other's price. Obviously, it will vary across different markets and products, but given the Internet and the mass media, prices are getting easier to monitor these days.

The other key difference of our model it that it captures the inventory position as a driver of price dispersion. Note from the figures that the retailer's markdown time is a decreasing function of his initial inventory position. Markdown happens either on its own initiative (due to a high inventory level and a disappointing demand rate) or motivated by the competitor's markdown. In either case, competition redirects the market demand from one retailer with a low inventory to another with a high inventory. On the one hand, it is similar to the behavior of a monopolist who "shapes demands" across different products by dynamically adjusting the prices of two products to shift the demand away from a low-stock product to a high-stock product. But markdown competition would enhance economic efficiency by achieving inventory pooling. Note this happens in a decentralized manner and despite informational asymmetry—as envisioned by Hayek (1945) (who assumed

there is "only one price for any commodity" in one market). Unfortunately, for lack of a consumer choice model, our model would be insufficient to formally investigate the efficiency issue.

Fourth and last, note from Fig. 15.1 or Theorem 2 that a markdown will happen only after a certain time $A^*\left(:=\frac{\beta_{01}\beta_{11}}{\beta_{11}-\beta_{01}}\ln\frac{p_0}{p_1}\right)$. This comes from two observations: (1) the preemptive markdown time $t_b^*(\cdot)$ is a decreasing function of the initial inventory level, and (2) even if a retailer has a lot of inventory (larger than S°), his optimal markdown time remains at A^*. This seems consistent with our perception that markdowns are what we expect towards the end of lifecycle.

The paper deliberately took a minimalist approach, loaded with a series of simplifying assumptions. Relaxation of these assumptions (e.g., deterministic demand) would be desirable. But given that we could not obtain any crisp results from the present simple model, I would rather hope to see a model that is even simpler and yet insightful, or empirical study that would supplement our modeling approach.

Acknowledgements I would like to thank the editor and the referees for offering valuable input to earlier drafts.

Appendix: A Sketchy Derivation of (15.7)

Note first in (P4) that since $\bar{G}^o(\tau|t)=\bar{G}(\tau)/\bar{G}(t)$ and $g^o(\tau|t)=g(\tau)/\bar{G}(t)$, every $G^\circ(\cdot\,|\,t)$ and $g^\circ(\cdot\,|\,t)$ can be respectively replaced by $G(\cdot)$ and $g(\cdot)$. Note also that G can be derived from the distribution of random variables S_j via $t_a^*(\cdot)$ and $t_b^*(\cdot)$, and is a mixed (i.e., continuous and discrete) distribution. The first term is his expected profit when retailer j first marks down and he follows immediately. Thus, the probability of retailer j's markdown happening no later than τ is given by $G(\tau)=P\left(t_b^*\left(S_j\right)\le\tau\right)=P\left(S_j\ge t_b^{*^{-1}}(\tau)\right)=1-F\left(t_b^{*^{-1}}(\tau)\right)$. Thus, $g(\tau)=-dF\left(t_b^{*^{-1}}(\tau)\right)/d\tau$. The second term captures the case where retailer i first marks down at t_b and retailer j immediately follows. In this case $g(\tau)$ has a probability mass at $\tau=t_b$, since any retailer j whose S_{jt_b} (or $S_j+\Delta_{t_b}$) value satisfies $t_a^*(S_{jt_b})<t_b\le t_b^*(S_j)$ will immediately follow retailer i's move. Thus, $g(t_b)=F\left(t_b^{*^{-1}}(t_b)\right)-F\left(t_a^{*^{-1}}(t_b)+\Delta_{t_b}\right)$. The third captures the case where retailer i first marks down at t_b and retailer j follows later at τ. Retailer i's demand rate changes from β_{00}, to β_{10} (at t_b) and then to β_{11} (at τ). In this case $G(\tau)=P\left(t_a^*\left(S_j-\Delta_{t_b}\right)\le\tau\right)=P\left(S_j-\Delta_{t_b}\ge t_a^{*^{-1}}(\tau)\right)=1-F\left(t_a^{*^{-1}}(\tau)+\Delta_{t_b}\right)$, giving $g(\tau)=-dF\left(t_a^{*^{-1}}(\tau)+\Delta_{t_b}\right)/d\tau$. The last term covers the case where retailer i first marks down, but retailer j never follows, since her initial inventory is lower than β_{01}, so she can sell all at the regular price even in the worst scenario (i.e., at demand rate β_{01}). This happens with

probability $F(\beta_{01})$, which is here denoted by $g(\infty)$. The constraints ensure that sales do not exceed the inventory in each instance of τ.

Regrettably, (*P4*) is very difficult to solve. One way to tackle the problem is to form a Lagrangian and obtain its saddle point (Luenberger 1969). The FOC of the Lagrangian, after straightforward manipulation and letting $t=0$ without loss of generality, gives:

$$\begin{aligned}
&\left(p_0 e^{-t_b/\beta_{00}} - p_1 e^{-t_b/\beta_{11}}\right) F\left(t_b^{*^{-1}}(t_b)\right) + p_1\left(e^{-t_b/\beta_{11}} - e^{-t_b/\beta_{10}}\right) F\left(t_a^{*^{-1}}(t_b) + \Delta_{t_b}\right) \\
&\quad - \lambda_1(t_b)\left[\beta_{00}\left(1 - e^{-t_b/\beta_{10}}\right) - \beta_{11}\left(e^{-t_b/\beta_{11}} - e^{-T_1(t_b)/\beta_{11}}\right) - S_i\right] \\
&\quad - \lambda_2'(t_b)\left[\beta_{10}\left(1 - e^{-t_b/\beta_{10}}\right) - \beta_{11}\left(e^{-t_b/\beta_{11}} - e^{-T_2/\beta_{11}}\right) - S_i\right] - \lambda_2(t_b)\left(e^{-t_b/\beta_{10}} - e^{-T_2/\beta_{11}}\right) \\
&\quad + \lambda_3(t_b)\left[\beta_{00}\left(1 - e^{-t_b/\beta_{10}}\right) + \beta_{11}\left(e^{-t_b/\beta_{11}} - e^{-T_3(t_b)/\beta_{11}}\right)\right] - \bar{\Lambda}_3(t_b)\left(e^{-t_b/\beta_{00}} - e^{-t_b/\beta_{10}}\right) \\
&\quad + \left(p_0 e^{-t_b/\beta_{00}} - p_1 e^{-t_b/\beta_{10}}\right) F(\beta_{01}) \\
&\quad + \lambda_4(t_b)\left[\beta_{00}\left(1 - e^{-\tau_b/b_{00}}\right) + \beta_{10}\left(e^{-t_b/\beta_{10}} - e^{-T_4/\beta_{10}}\right) - S_i\right] = 0,
\end{aligned} \tag{15.7}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are the Lagrangian multipliers to the four constraints of (*P4*) in that order, and $\bar{\Lambda}_3(t_b) := \int_{t_b^+}^{\infty} \lambda_3(\tau) d\tau$. By definition of symmetric equilibrium, retailer i's choice of t_b should be equal to retailer j's optimal t_b^*, hence $t_b^{*^{-1}}(t_b) = t_b^{*^{-1}}\left(t_b^*(S_i)\right) = S_i$. Applying this to (15.7), we have:

$$\begin{aligned}
&\left(p_0 e^{-t_b^*/\beta_{00}} - p_1 e^{-t_b^*/\beta_{11}}\right) F(S_i) + p_1\left(e^{-t_b^*/\beta_{11}} - e^{-t_b^*/\beta_{10}}\right) F\left(t_a^{*^{-1}}(t_b^*) + \Delta_{t_b}\right) \\
&\quad - \lambda_1(t_b^*)\left[\beta_{00}\left(1 - e^{-t_b^*/\beta_{10}}\right) - \beta_{11}\left(e^{-t_b^*/\beta_{11}} - e^{-T_1\left(t_b^*\right)/\beta_{11}}\right) - S_i\right] \\
&\quad - \lambda_2'(t_b^*)\left[\beta_{10}\left(1 - e^{-t_b^*/\beta_{10}}\right) - \beta_{11}\left(e^{-t_b^*/\beta_{11}} - e^{-T_2/\beta_{11}}\right) - S_i\right] - \lambda_2(t_b^*)\left(e^{-t_b^*/\beta_{10}} - e^{-T_2/\beta_{11}}\right) \\
&\quad + \lambda_3(t_b^*)\left[\beta_{00}\left(1 - e^{-t_b^*/\beta_{10}}\right) + \beta_{11}\left(e^{-t_b^*/\beta_{11}} - e^{-T_3\left(t_b^*\right)/\beta_{11}}\right)\right] - \bar{\Lambda}_3(t_b^*)\left(e^{-t_b^*/\beta_{00}} - e^{-t_b^*/\beta_{10}}\right) \\
&\quad + \left(p_0 e^{-t_b^*/\beta_{00}} - p_1 e^{-t_b^*/\beta_{10}}\right) F(\beta_{01}) \\
&\quad + \lambda_4(t_b^*)\left[\beta_{00}\left(1 - e^{-t_b^*/\beta_{00}}\right) + \beta_{10}\left(e^{-t_b^*/\beta_{10}} - e^{-T_4/\beta_{10}}\right) - S_i\right] = 0.
\end{aligned} \tag{15.8}$$

References

Aviv, W., & Pazgal, A. (2003). *Optimal pricing of seasonal products in the presence of forward-looking consumers*. Working Paper, Olin School of Business, Washington University, St. Louis, MO.

Belobaba, P. P. (1987). Airline yield management: An overview of seat inventory control. *Transportation Science, 29*(3), 63–73.

Bitran, R., & Caldency, R. (2003). Pricing models for revenue management. *Manufacturing and Service Operations Management, 5*(3), 203–229.

Dudey, M. (1992). Dynamic Edgeworth-Bertrand competition. *The Quarterly Journal of Economics, 107*(4), 1461–1477.
Evers, J. (2002). Microsoft announces Xbox price cut. *PCWorld*. Accessed May 15, 2002, from http://www.pcworld.com/news/article/0,aid,99524,00.asp.
Feng, Y., & Gallego, G. (1995). Optimal starting times for end-of-season sales and optimal stopping times for promotional fares. *Management Science, 41*(98), 1371–1391.
Feng, Y., & Xiao, B. (2000). Optimal policies of yield management with multiple predetermined prices. *Operations Research, 48*(2), 332–343.
Gallego, G., & van Ryzin, G. (1993). Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science, 40*, 999–1020.
Gul, F., & Lundholm, R. (1995). Endogenous timing and the clustering of agents' decisions. *Journal of Political Economy, 103*(5), 1039–1066.
Hayek, F. (1945). The use of knowledge in society. *American Economic Review, 35*, 519–530.
Kreps, D. (1990). *A course in microeconomic theory*. Princeton: Princeton Book Company.
Lal, R. (1990). Price promotions: Limiting competitive encroachment. *Marketing Science, 9*(3), 247–262.
Lazear, E. (1986). Retail pricing and clearance sales. *The American Economic Review, 76*(1), 14–32.
Luenberger, D. G. (1969). *Optimization by vector space methods*. New York: Wiley.
Netessine, S., & Shumsky, R. (2004). *Revenue management games: Horizontal and vertical competition*. Working Paper, Wharton School, The University of Pennsylvania.
Rao, R. (1991). Pricing and promotions in asymmetric duopolies. *Marketing Science, 10*(2), 131–144.
Talluri, K., & van Ryzin, G. (2004). *Theory and practice of revenue management*. New York: Springer.
Varian, H. (1980). A model of sales. *American Economic Review, 70*(4), 651–659.
Varian, H. (1981). Errata: A model of sales. *American Economic Review 71*(3), 517.

Index

A
Achabal, D.D., 21, 23, 235, 387, 391, 392, 394, 398, 399, 406
Active decisions, 309, 310
Actual inventory, 54, 55, 58, 59, 65, 66
  and inventory record, discrepancy between, 65
Adams, P.W., 287
Adelman, D., 338
Adjusted inventory turnover (AIT), 26, 29, 30, 34–38
Advertising, influence on customer decisions, 12
Aggregate demand model, 267
Aggregate echelon inventory positions, 325
Aggregate level inventory management in retailing, directions for future research on, 49
Agrawal, N., 1–9, 11–23, 67, 82, 188, 196–197, 209, 230, 247, 270, 278, 294, 304, 319–343
Agrawal and Smith (2003), modeling and solution approach in, 209
Ahire, S.L., 327
Akella, R., 202, 269
Albert Heijn (Dutch supermarket chain)
  assortment planning in practice, 219–229
  price and promotion at, 214
  replenishment system at, 197
Allocation decision, 153, 176, 325, 328, 336, 337
Allocation policies
  types of, 328
  used at warehouse, 323–324
Alptekinoglu, A., 251
Anderson, E., 265
Anderson, P., 60
Anderson, S.P., 179, 185, 186, 190
Andrews, R.L., 298
Anily, S., 338
Antitrust
  concerns, and category captainship, 167–168
Anupindi, R., 6, 82, 217, 265–289, 331, 335, 336
Apparel specialty retailer, 239
Archibald, T.W., 335
Argote, L., 61
Arrow Electronics, high inventory record accuracy in, mechanism for, 59
Arrow, K., 271
Assets, of retail firm, 26
Assortment
  and inventory planning, 4–7, 195
    importance in product availability, 53
  key marketing objectives for, 12
  optimization heterogeneity in, 313
  ornamentation, 13
  problem, solution technique for, 283
  reductions, 184, 268, 269
  and stocking problem
    basic formulation, 274–227
    modeling no purchase, 277–278
    optimization model and some special cases, 279–280
    reformulation, 278–279
  strategy, based on percent sales *vs.* sales velocity matrix, 224

N. Agrawal, S.A. Smith (eds.), *Retail Supply Chain Management*,
International Series in Operations Research & Management Science 223,
DOI 10.1007/978-1-4899-7562-1

Assortment-based substitution, 176, 183, 184, 191, 192, 196, 198, 201, 218, 303
estimation of, 218
in exogenous demand model, 187–189
Assortment customization, at Borders Books, 219
Assortment decision, 7, 17, 22, 158, 159, 161, 164, 190, 194, 203, 204, 206, 208, 211, 223, 228, 229, 252, 266–268, 272, 274, 279, 280, 285, 287, 293, 296, 297, 300
seasonality of, 297
Assortment planning. *See also* Assortment selection, and inventory planning
at Albert Heijn (BV), 225–227
at Best Buy, 220–222
at Borders Group Ink, 222–223
constraints on, 13
consumer search in, 193
decentralized approach to, 211
in decentralized supply chains, 203–204
demand estimation
of MNL, 212–216
of substitution rates in exogenous demand models, 217–218
demand models
consumer driven substitution, 182–184
exogenous demand model, 187–189
locational choice model, 189–190
multinomial Logit (MNL) model, 184–186
used in, 6
directions for future research, 229–231
dynamic, 204–206
under exogenous demand models
Kök and Fisher model, 197–201
Smith and Agrawal model, 196–197
factors for consideration, 6
industry approach to
Albert Heijn (BV), 225–227
Best Buy, 220–222
Borders, 222–223
compared to academic, 227–229
Tanishq, 223–225
and inventory management, additional aspects of, 22
under locational choice, 201–203
with MNL: The van Ryzin and Mahajan model, 190–195
models
for retailers of multi-featured products (*see* Retail assortments optimization for diverse customer preferences)
for tractability, assumption made in, 188
models with multiple categories, 208–211
in multi-store, 231
in practice (*see* Assortment planning, industry approach to)
related literature
multi-item inventory models, 180
product variety and product line design, 178–179
shelf space allocation models, 181–182
variety, perception of, 182
and replenishment, attribute-focused, 228
at Tanishq, 223–225
Assortments
rationalization using household scanner panel data (*see* Product variety, on retail shelf, management of)
Assortment selection
and inventory planning
in decentralized supply chains, 203–204
dynamic, 204–206
under exogenous demand models, 196–201
under locational choice, 201–203
with MNL: van Ryzin and Mahajan model, 190–195
models with multiple categories, 208–211
optimal, 195
and presentation design decisions, relationship between, 13
Assortment selection decision, 5, 151, 158–162, 204
delegation of, 157
Assortments optimization, for diverse customer preferences
DVD player data base, illustrative application for, 309–314
comparing model's predictions to retailer's sales data, 310–311
impact of customer preference structure, 312–314
optimal assortment *vs.* expected revenue of retailer's assortment, 311–312
model description
modeling consumer's purchase, 296–302
optimal assortment, properties of, 306–308
optimization problem, solving of, 308–309
retailer's assortment optimization, 302–306
Atali, A., 66, 95, 97, 98, 102
Atkins, D.P., 338
Attraction model, 251

Automated replenishment, 54, 69, 81, 93, 101
Automated replenishment systems, 54, 93
  failure of, 63
Aviv, W., 409
Aviv, Y., 333
Avsar, Z.M., 180
Axsater, S., 7, 319, 329, 333, 341, 342
Aydin, G., 8, 158, 203, 229, 231, 349–384

B
Backlogging assumption, 65
Backordering, of unmet demand, 324, 328, 330, 332, 342
Backrooms, 68, 69
Baker, K.R., 339
Balakrishnan, R., 31
Balance assumption, 326, 329
Balintfy, J., 338
Banana Republic, 13
Bankruptcy, 34, 113
  effects on model estimation, 51
Barlow, R.E., 357
Base stock policy, 102, 104
  adjusted, 65
Basics products, 240, 241, 252
Basket profits, 207, 211
Basket shopping consumers, 177, 207, 209
Basket value, 124, 126, 129, 141
Bass, F., 396
Bassok, Y., 180, 202, 269, 331, 335, 336
Basuroy, S., 148, 185
Batch ordering, 329–330, 332, 342
Baumol, W.J., 209
Bawa, K., 273
Bayers, C., 54
Bayesian inventory record, 103
Bayesian learning mechanism, use in retail, 205
Bayesian models, 81, 84–86, 89, 91, 92, 105
Baykal-Gursoy, M., 180
Belgian supermarket chains, Corstjens and Doyle model for shelf space allocation in, 181
Bell, D.R., 209, 354
Belobaba, P.P., 409
Ben-Akiva, M., 185, 186, 210, 298, 299
Benchmarking, 4, 26, 28, 29, 37, 49
  of inventory productivity of retail firms, 26
Benefits, 22, 66, 68, 88, 94, 100, 101, 103, 115, 125, 127, 128, 135, 138, 142, 148–152, 157, 160–164, 166–168, 171, 203, 211, 221, 229, 231, 248, 272, 314, 315, 321, 330–333, 336, 352, 354
Bergman, R.P., 54
Bertrand, L.P., 335
Besanko, D., 391
Best Buy (electronics retailer)
  assortment planning in practice, 219–229
  rebate handliing at, 369
Bharadwaj, S., 266
Billesbach, T.J., 31
Bish, E.K., 193
Bitran, R., 409
Bluedorn, A., 61
Boatwright, P., 178, 265, 269
Bolton, R.N., 391
Bookbinder, J.H., 335
Borders Group Inc. (book and music retailer)
  assortment planning in practice, 219–229
  average employee turnover in, 60, 61
  drivers of misplaced products in, research analysis on, 71
  effect of product variety and inventory levels on store sales in, 73
  misplaced products in, 56
  reduced profits due to misplaced products in, 54
Borin, N., 181, 268
Borle, S., 269, 273
Bottom-up analysis, 15
Braden, D.J., 83, 390, 391
Bradford, J.W., 322
Bramel, J.D., 338
Brand(s). *See also* Category captainship
  choice, 186, 271, 272, 274
  of first and second preference, cross classification of, 281, 282
  management, by vendors, 6
  noncaptain, 204
  substitution between, asymmetries in patterns of, 283
Brand strength, 153, 157
Broniarczyk, S.M., 178, 182, 268, 269
Brooks, R.B., 54
Brown, K.A., 54
Bucklin, R.E., 212, 215, 277, 301
Bulkeley, W.M., 349, 351, 369
Bull whip effect, 28–30, 331
Bultez, A., 181, 268
Business analytics, 108, 255–256
Butz, D.A., 354

C
Cachon, G.P., 8, 28, 29, 158, 159, 186, 192, 193, 207, 209, 210, 217, 229, 230, 249, 254, 294, 295, 330, 332, 338, 372, 391

Caldency, R., 409
Caldenty, R., 390
Camdereli, A.Z., 66, 97, 100
Cameras, traditional and digital, historical sales of, 220
Campo, K., 184, 218
Capital intensity, 26, 27
  correlation with inventory turnover, 38, 41
Capital intensity (CI), 3, 26, 27, 29, 32–37, 40, 43, 49, 51
Caro, F., 6, 204, 228, 237–259, 387, 388, 391
Carpenter, G., 266
Carrefour, 149, 169, 170
Cashback allowances, consumer rebate in automotive industry, 356
Cassidy, A., 54
Catalog channel, 12–15, 19, 21, 22, 341
  customers from, shipment of merchandise to, 19
Catalogs, as advertising mechanism, 21
Category captain
  selection, points for, 157
  use of, 148
Category captainship
  adverse effect of, 169
  benefits of, 164
  relationships, basis for, 171
Category captainship, in retail industry
  categories of existing research on
    antitrust concerns, 167–168
    delegation of assortment selection decision, 158–162
    delegation of pricing decisions, 152–157
    emergence of, 162–167
  future research directions, 170–171
  impact of, 168–170
  implementations in practice, 148–151
Category management (CM). *See also* Category captainship; Product variety, on retail shelf, management of
  decentralized assortment control in, 207
  ECR component, 265
Cattin, P., 310
Cell phone tracking, 80
CGS (cost of goods sold)$_{sit}$, components of, 33
Chang, D., 31
Chang, P.L., 332
Channel coordination, 9, 66, 156
Channel rebates. *See* Retailer rebates
Chen, F., 179, 327, 328, 333, 342
Chen, H., 4, 29, 30
Chen, K.D., 295
Chen, M.S., 332
Chen, X., 355
Cheng, F., 391
Chiang, J., 212
Children furnishings, concept of brand retail marketing, 12
Chintagunta, P.K., 212, 215, 274
Choi, S.C., 153
Chong, J-K., 213, 270, 295
Christopher, M., 184, 246
Circa 1990, 2
Clark, A.J., 326
Clearance, and markdown optimization, 3, 21
Clearance markdown
  algorithms, 406
  management, components of, 389
  pricing system, 389
Clearance pricing in retail chains
  discrete price changes, 398–402
    power function form, solution for, 399–402
  mathematical models for, 389–390
  model specifications and optimality conditions, 392–398
    model formulation, 393–398
  numerical examples, 402–405
  related research, 390–392
  trends in, 388–389
Cohen, M.A., 333, 334, 339
Colgate, as category captain in oral care category, 149
Collier, D.A., 339
Competition
  duopoly, 211
  markdown competition
    managerial implications, 420–422
    model, 411–412
    monopolistic retailer, 412–414
Competitive collusion, 167–168
Competitive exclusion, 150, 151, 155, 161, 164, 167, 168, 171
Computing technology, application in retail business, 2
Consideration sets
  effect on expected profit, 313
  use of, 313–314
Constant-split property
  consequence of, 363
  rationale behind, 364

Consumer(s)
basket shopping, 177, 207, 209, 210
behavior, basket effect of, 230
choice, 6, 272, 315
education, 160, 164
heterogeneity of, 280
model, for study of product variety management on retail shelf, 271–274
package goods, purchasing behavior for, 301
purchase behavior model, basis of, 215
purchase decision, modeling, 296–302
substitution behavior of, 6, 177, 266
substitution between, heterogeneity in patterns of, 283
substitution, consumer-driven, 182–184
Consumer rebate
modeling, 350
pull price promotions, 353
in supply chain, 349–384
together with retailer rebate, 356–360
use of, 353
Consumer segments, 230, 265, 279, 287
discrimination between, 230
Containers, 14, 18, 19, 22, 23
integer number of, 22
Continuous review inventory systems, 341–343
Conversion rate, 124, 128–131, 133, 141–143
Converting incoming traffic into sales, 113
Conwood, antitrust row with UST, 150
Cookware essentials, concept of brand retail marketing, 12
Cooper, L.G., 213, 215
Cornuejols, G., 279
Corporate social responsibility, 257–259
Corstjens and Doyle model, for shelf space allocation, 181
Corstjens, J., 153, 169
Corstjens, M., 169, 181, 226, 268
Cortsen, D.S., 266
Costs
holding and penalty costs, 324, 325, 329, 331
inter-node transportation, 332
inventory costs, 181, 190, 192, 201, 209, 246, 249, 330, 332, 338, 354, 392, 393, 396
reductions, opportunities for, 21
Coughlan, A.T., 265
CPFR (collaborative planning, forecasting and replenishment) programs, 69
Cross-channel optimization, in supply chain planning, 3, 21–22
Cross channel pricing tradeoffs, 23
Cross-docking policy, 324
Cross-price sensitivity, 153, 156, 157, 162, 164
Current priority allocation (CPA), 328, 329
Customer(s)
choice behavior of, assumptions characterizing, 187
clustering of, 295, 315
heterogeneity, 7, 202, 267, 312–313, 315
impact on optimal assortments and expected profits, 312
from internet and catalog channels shipment of merchandise to, 19
misplacement of products by, 58
“no purchase” option of, reasons for, 297
retention, 267, 269
store sales *vs.* customer entrances, 64
substitution, patterns with respect to, 176
variety-seeking behavior, 182
Customer co-production, 121
Customer preferences, optimization of assortments for
DVD player data base, illustrative application for
comparing model’s predictions to retailer’s sales data, 310–311
impact of customer preference structure, 312–314
optimal assortment *vs.* expected revenue of retailer’s assortment, 311–312
model description
modeling consumer’s purchase decision, 296–302
optimal assortment, properties of, 306–308
optimization problem, solving of, 308–309
retailer’s assortment optimization, 302–306
Cycle counts, 63, 97, 104
Cycle stock, 39, 101

D
Dada, M., 86, 92, 170, 354, 372
Dalton, D., 61
Daniels, R.L., 120

Data
financial data, 26, 30, 31, 51, 119, 124, 129, 139
inventory data, 4, 7, 28–30, 54, 64, 67, 69, 83, 127, 217, 218, 224, 320
labor data, 117, 124, 125, 129, 131
payroll data, 123–128
point-of-sale data, 207, 211
traffic data, 107, 121, 124, 125, 129, 131, 140, 143
Dayton Hudson, 13
DCs. *See* Distribution centers (DCs)
De Groote, X., 179, 201
De Kok, A.G., 63, 328, 338, 339
Dealer incentives, retailer rebate in automotive industry, 356
Decision making process, retailer's, 22
Decisions, tactical and strategic, impact of execution problems on, 66. *See also Specific types*
Deep discount drug stores, growth of, 266
DeHoratius, N., 53–75, 81, 93, 95–99, 103–105, 143, 218
De-listing of firms, 34
Demand. *See also* Unmet demand
censoring, 80–93, 106, 107
estimation, 4, 67, 82–84, 211–219
estimation, improvement of, 67
fluctuation, 114
forecasting, 301
modeling of, 321–323
non-stationary, 92, 326
pooling, effects on product variety, 26
substitution, 91–92, 205
uncertainty, incorporation of, 287
Demand models
consumer driven substitution, 183–184
exogenous
assortment planning under, 196–201
demand estimation of substitution rates in, 217–219
locational choice model, 189–190
MNL model, 184–186
Demeester, L., 31
Demoralization, of employees, 61
Dempster, A.P., 213
Denend, L., 20
Desai, P., 179
Design-to-sales, 242
Desrochers, D.M., 150, 167–169
Dhar, S.K., 148, 228
Dhebar, A., 390, 396
Diks, E.B., 328, 339
Diseconomies of scale, 40, 202
Display effect (effect of facings) on sales, 287–288
Distribution centers (DCs)
execution problems in, 54–58
merchandise handling capabilities at, 14
shipments from DCs to stores, 20
Distribution planning, and inventory management, 3, 20–21
Disutility, measure for, 287
Dobson, G., 179, 271, 274, 288, 295
Dogru, M.K., 329
Dong, L., 336
Downs, B., 180
Doyle, P., 153, 181, 226, 268
Dreze, X., 178, 354
Drop-shipping channel, 203
Dudey, M., 409, 410
Duopoly competition, 211
DVD player data base, illustrative application for
comparing model's predictions to retailer's sales data, 310–311
impact of customer preference structure, 312–314
optimal assortment *vs.* expected revenue of retailer's assortment, 311–312
Dynamic assortment, 190, 204–206, 230, 243, 244, 250–253
Dynamic assortment planning, 190, 204–206, 230
Dynamic markdown competition, 409, 410. *See also* Markdown competition
Dynamic pricing, 92, 108, 390–392, 406, 409
Dynamic pricing models, variety of, 390
Dynamic substitution, 191, 193, 201, 202, 294
variety under, 202

E

Echelon stock, 325, 326, 342
of warehouse, allocation of, 326
Economic order quantity (EOQ) model
cost function in, 191
for incorporation of economies of scale, 179
for inventory levels decisions at Tanishq, 225
on inventory turnover, 27
replenishment costs as explained by, 39
Economies of scale and scope
arguments of, 45, 49
diminishing, 44–45

drivers of, 40
effects of, 38
factors contributing to, 26
hindrances to, 38, 40
realization in transportation costs, 39
in retail setting, 27
Efficient consumer response (ECR) category management as component of, 265
El-Ansary, A.I., 265
Electronic data interchange (EDI), 20, 140, 323
Eliashberg, J., 179, 391
Elmaghraby, W., 170, 390
Emma, C.K., 65
Emmelhainz, L.W., 53, 184, 266, 273, 274
Emmelhainz, M.A., 53, 184, 266, 273, 274
Empirical research on retail labor, 119, 123–139
Employee error, 57–60
examples of, 57
Employee incentives, 68
Employees
demoralization of, 61
nonconformance among, misplacement of products due to, 57
role in product availability management efforts, 62–63
employee turnover and training, 60–61
employee workload, 61–62
Employee turnover, 4, 58, 60–61, 68, 71, 73, 74, 124, 128, 135–139
and phantom products, relationship between, 71
problems of, 61
and training, influence on product availability management, 60–61
Endogeneity issues between labor and store performance, 124
EOQ. *See* Economic order quantity (EOQ) model
Eppen, G.D., 26, 39, 326, 327, 331
Epple, D., 61
Erlenkotter, D., 279
Estimation and empirical testing, 301
Estimators, ordinary least squares (OLS) estimators, 73
Evers, J., 410
Execution
poor
drivers of, 58
sources of, 56
problems, strategies for reduction of occurrence, 68
and product availability management in retail
factors exacerbating problems, 58–63
influence on inventory planning, 63–66
problems in, 54–58
research opportunities in, 67–69
Exogenous demand models
assortment planning under
Kök and Fisher model, 197–201
Smith and Agrawal model, 196–197
substitution rates in, demand estimation of, 217–218
Expectation-Maximization(EM) algorithm, for correction of missing data, 213
Expenses
labor expenses, 113, 116, 124, 138
payroll expenses, 61, 62, 68, 72, 74, 113, 128, 135
store expenses, 5, 138, 139
wage expenses, 124
Externality costs, 122

F

Fader, P.S., 205, 212
Fall season, 14, 296
Fama, E.F., 30
Farris, P., 181
Fashion-basic products, 240, 243
Fashion products, 241, 337–338, 411
model with ‘demand trajectory’ for, 411
Fashion trends, 204, 254, 256–257
Fast fashion, 6, 237–259, 388
Federgruen, A., 7, 319, 326, 333, 338
Feng, Y., 390, 409
FIFO (first in first out), inventory valuation method, 32
Firms
de-listing of, 34
lifecycle, effects on model estimation, 51
size (*see also* Inventory turnover performance, effects of firm size and sales growth rate on)
correlation with inventory turnover, 27, 38, 39
effect of inventory turnover on, 38–40
Fisher, M.L., 6, 26, 60, 119, 124, 125, 127, 128, 141, 143, 175–231, 245–248, 255, 270, 295, 332, 337, 391
“Fixture fill” minimum on-hand inventory, 393
Flagship brand, 17, 237
Fleisch, E., 66

Flexibility
labor flexibility, 120, 135–139
volume flexibility, 115, 138, 139
Flores, B.E., 54
Forecast accuracy, 67, 181, 257
sensitivity of shelf space allocation models to, 181–182
Forecast errors, 121, 129, 133, 134, 339
Freeland, K., 268
French, K.R., 30
French, S., 276
Furniture retail, importance of DC in, 14

G
Gallego, G., 390, 396, 409, 412
Galway, L.A., 54
Gamma Corporation
drivers of inventory record inaccuracy in, research analysis on, 69
inventory record inaccuracy in, 55–56
The Gap, 6, 13, 23, 178, 409
The Gap stores, 13
Gaukler, G.M., 66, 100
Gaur, V., 3, 25–51, 201, 270, 391
Gavirneni, S., 333
General Mills, as category captain in baking ingredients and mixes category, 149
Gerchak, Y., 339, 390
Gershwin, S.B., 66, 93, 97, 101, 103
Gerstner, E., 353
GFR [Gaur, Fisher and Raman (2005)]
on inventory turnover performance of U.S. retailers, 26, 34, 49
use of, 27
re-test of hypotheses in, 44
Gilbert, S.M., 170
Gilligian, T., 152
Glasserman, P., 339
GMROI. *See* Gross margin return on inventory (GMROI)
Goyal, S.K., 180
Grabner, J., 184
Graves, S.C., 328, 342
Graves, S.G., 63
Greenberger, R.S., 150
Greene, W.H., 213
Green, P.E., 295, 310
Grocery industry, Albert Heijn, replenishment system at, 197
Gross margin (GM), 3, 26–30, 32, 34–37, 43, 49, 51, 133, 175, 243, 387, 389
correlation with inventory turnover, 31
Gross margin return on inventory (GMROI), 243, 244
Gruca, T.S., 158, 165, 213
Gruen, T.W., 53, 93, 148, 183, 184, 266
Guadagni, P.M., 185, 212, 213
Guar, V., 294, 303, 304
Gul, F., 421
Gupta, S., 6, 212, 215, 265–289

H
Ha, A., 339
Hadley, G., 180
Hall, R.V., 59
Hanks, C.H., 54
Hanson, W., 296
Hardie, B.G.S., 205, 212
Harris, B., 276
Hart, M.K., 54
Hauser, J.R., 266, 273
Hausman, W.H., 158, 203, 231, 243, 295
Hayek, F., 421
Hayen, R., 31
Hedonic products, 182
Henig, M., 339
Hennes and Mauritz (H&M), 6, 237, 239–241, 243, 244, 253, 258
Herer, Y.T., 335
Hess, J.D., 353
Ho, T.-H, 209
Hoch, S.J., 182
Holding costs, 197, 304, 324, 328, 392, 406
Hollinger, R.C., 58
Home furnishings
concept of brand retail marketing, 11
supply chains, characteristics of, 13
Honhon, D., 159, 165, 195, 201, 270, 294, 303, 304
Hotelling, H., 178
Hotelling model, 178. *See also* Locational choice model
Huchzermeier, A., 170
Huffman, C., 182
Human discomfort index (HDI), 216
Huson, M., 31

I
Ide, E.A., 209
Iglehart, D.L., 65, 93, 101, 102
Imai, K., 272, 280
Imputed cost of labor, 134
Inditex, 237, 239, 244, 258

Industry practices around workforce management, 116
Infinitesimal perturbation analysis, 335
Information sharing alliance, 331–333
Information systems for retailers, Circa 1990, 2
Inspection policy, 65, 103, 104
In-store experience, 4, 113, 114, 143
In-store retail technologies, 140
In-store visibility, 79–81, 106
Internet channel, 12, 13
  customers from, shipment of merchandise to, 19
  deeper price markdowns in, 21
Internet retailing, drop-shipping channel in, 79, 203
Internet sessions, 294
Intertemporal pricing, 390
Inventory
  agreements, vendor-managed, 231
  allocation
    policies used at warehouse, 323–324
    solution methodologies for issue of, 326–329
  as asset of retail firm, 26
  costs, 4, 22, 181, 190, 192, 201, 209, 246, 249, 294, 330, 332, 338, 354, 392, 393, 396
  counts, optimal frequency of, 65
  imbalance, 392
  levels
    and product availability, relationship between, 67
    role in product availability management, 59–60
  ordering decision, 325 (s*ee also* Inventory decisions)
  pooling, 421
  productivity
    drivers of, 51
    of retail firms, benchmarking of, 26
  productivity performance, 26
  record accuracy, high, in Arrow Electronics, mechanism for, 59 (*see also* Inventory record inaccuracy (IRI))
  replenishment, 60
  shrinkage, 62
  system and actual, absolute value difference between, 55
  team, responsibility of, 17
  theory, newsboy model in, 49
  valuation methods, 32
  vendor managed, 1
Inventory decisions, 256
  aggregate-level, modeling of, 50–51
  factors involved in, 177
Inventory management, 20–21
  of multiple products, 180
  in retail trade, importance of improvement in, 25
Inventory management decisions, 21. *See also* Retail supply chain management, Multi–location inventory models for
Inventory models
  general periodic review, 324–326
    additional issues, 338–339
    batch ordering, 329–330
    decentralized environments, 331–333
    fashion products, 337–338
    lateral pooling, 333–336
    lost sales, 330–331
    solution methodologies, 326–329
    transportation issues, 338
  multi-item, 180
  multi-location, for retail supply chain management, 319
    modeling issues, 320–324
Inventory planning
  and assortment, 5–7
  and assortment selection
    in decentralized supply chains, 203–204
    dynamic, 204–206
    under exogenous demand models, 196–201
    under locational choice, 201–203
    with MNL: van Ryzin and Mahajan model, 190–195
    models with multiple categories, 208–211
  effect of inventory record inaccuracy on, 63–64
  effect of misplaced products on, 64–65
  research focus on, 2–3
Inventory record inaccuracy (IRI), 4, 53, 55–58, 60, 62–69, 79–108
  effect on inventory planning, 63–64
  impact of, mitigation of, 66
  of SKU, 69
  theft as source of, 66
Inventory turnover (IT)
  adjusted (AIT), 29
  correlation with firm size, 27, 31
  correlation with firm size and sales ratio, 38
  effect of
    firm size on, 38–40
    sales ratio on, 41–43

Inventory turnover (IT) (*cont.*)
performance, across firms, differences in, 49
in performance analysis, 4
performance, effects of firm size and sales growth rate on, 26
adjusted inventory turnover, 34, 37–38
data description, 31–36
directions for future research, 49–51
firm size effect on, 38–40
hypotheses, 38–43
literature on, 28–31
model, 43
results, 43–49
sales ratio effect on, 41–43
variation in, 26
variation of, 34
Invoice accuracy, 63
IPACE (information, price, assortment, convenience and entertainment) model, for retail shopping decisions, 296
Irion, J., 181
Ittner, C.D., 60
Iyer, A.V., 170, 245, 246, 249
Iyogun, P., 338

J
Jackson, P.L., 246, 326, 338
Jain, D.C., 274
Jennifer Convertibles, Inc, 50
Jewelry
Indian, 177 (s*ee also* Tanishq)
market, India's, 223
products, 45
Johnson, P.L., 266, 273
Joint fixed costs, 288
Joint inventory and pricing decisions, literature on, 170
Joint pricing and assortment planning, 229
Joint replenishment effect, 321
Joint replenishment problem, 338
Jonsson, H., 326, 335
Just-in-time (JIT)
manufacturing, 59
principles, adoption of, 29, 31

K
Kadane, J.B., 269, 273
Kahn, B.E., 182, 300
Kalish, S., 179, 271, 273, 274, 288, 295, 390, 391, 396
Kalyanam, K., 21, 23, 296, 355, 391, 395
Kamakura, W.A., 274
Kamien, M.I., 394
Kang, Y., 66, 93, 97, 101, 103
Kaplan-Meier estimator, 90
Kapuscinski, R., 249, 333, 338, 354
Karlin, S., 271
Karmarkar, U.S., 221, 334
Kenny, D., 178
Kesavan, S., 3, 25–51, 108, 113–143
Keskinocak, P., 170, 355, 390
Kim, S.Y., 170
Kleywegt, A., 338
Koenig, S., 351
Kohli, R., 202, 295
Kok, A., 6, 63, 158, 270, 294, 295, 303, 328, 338, 339
Kök and Fisher model, for inventory planning and assortment selection, 197–201
Kouvelis, P., 228
Krafcik, J.F., 59
Krajewski, L.J., 54
Kreps, D., 414
Krieger, A.M., 295
Krishna, A., 273
Krishnan, H., 249, 354
Krishnan, K.S., 334
Kuhn-Tucker theorem, 413
Kurtulus, M., 5, 147–171, 203, 204

L
Labor Hours
base hours, 117
forecasted hours, 115, 118
full-time equivalent, 137
sales associate hours, 123
stockroom hours, 123
Labor-mix
flexible labor-mix, 138
part-time labor mix, 139
temporary labor mix, 139
Labor scheduling tools, 140
Lagrangian relaxation approach, 197
Lal, R., 354, 410
Lancaster, K., 179, 189
Langton, L., 58
Lariviere, M.A., 86, 88, 391
Lateral pooling, 333–336
Lattin, J.M., 209, 301
Laudon, K.C., 54
Lazear, E.P., 390, 391
Lead time(s), 17, 198, 246, 250

design-to-shelf, at Mango (Spain), World Co. (Japan), and Zara (Spain), 204
types of, 323
for upholstery or fabrics, 18
Lean production system, 68
Lee, H.L., 95, 271, 331–333
Lee, S.M., 31
Lee and Wrangler, as category captain jeans category, 149
Lehmann, D.R., 266, 300
Lerman, S.R., 185, 210, 298, 299
Levitan, R.E., 153
Levy, M., 219, 228
Li, C.-L., 355
Lieberman, M.B., 31
LIFO (last in first out), inventory valuation method, 32
Lin, C.T., 332
Lippman, S.A., 180, 335
Liquidation options, 14
Little, J.D.C., 185, 212, 214, 276
Little's Law, 327
Liu, P., 184
Locational choice demand model, assortment planning under, 201–203
Locational choice model, 6, 177, 183, 189–190, 203, 206, 270
Lockers, off-site, 21
Logistics planning, of supply chain, 18–20
Lost sales, 54, 59, 65–67, 81–84, 90, 93, 98, 99, 103, 104, 107, 121, 129, 131–133, 180, 187, 195, 198, 205, 209, 214, 304, 315, 322, 323, 327, 330–331, 352
causes of, 66
due to misplaced products, 54, 56
estimate of, 54, 67
modeling, complexity of, 323
for unmet store demands, 323
Louviere, J., 273
Luenberger, D.G., 417, 423
Lundholm, R., 421

M

Maddah, B., 193
Mahajan, S., 158, 178, 180, 185, 190–194, 197, 199, 201, 203, 204, 207, 213, 270, 271, 294, 295, 303, 304, 306
Makridakis, S., 41
Malhotra, A., 29
Manchanda, P., 209
Mango (Spain), design-to-shelf lead time at, 204
Mantrala, M.K., 391
Manufacturer-retailer partnerships, 170
Manufacturers
non-captain, 150, 151, 155, 157, 161–164, 166–168
non-partnering, 170
as Stackelberg leader, 156
Manufacturing, 18, 19, 29, 30, 54, 59, 60, 68, 119–124, 141, 143, 245, 247, 257, 288
Manufacturing process, time taken for, 18
MARK, decision support system, 391
Markdown competition, 9, 409–423
managerial implications, 420–422
model, 411–412
monopolistic retailer, 412–414
Markdown optimization, in supply chain planning, 21
Markdown planning, 8, 21
importance in pricing (*see* Clearance pricing in retail chains)
Marketing
channels, direct-to-consumer, 12
cross-channel, 21–22
trade promotions in, literature on, 170, 353, 354
Markov decision process (MDP), 96, 103
Markov-modulated demand, 92
Martello, S., 22
Martinez-de-Albeniz, 6
Mathur, K., 338
Maximum likelihood estimates (MLE), 82–84, 212, 217
McBride, R.D., 202, 271, 274, 295
McCardle, K.F., 180, 335
McClain, J.O., 54
McCullogh, R., 272, 280
McCutcheon, C., 54
McFadden, D., 212
McGavin, E.J., 323, 327, 330, 338
McGillivray, A.R., 180
McGuire, T.W., 153
McPartland, M., 276
Mean absolute deviation (MAD), 216
Menzies, D., 369
Merchandise
categories of, 17, 135, 149, 178, 207, 214
optimal pricing policy for, properties of, 406
paths for, 15
strategic trends in retail, 388–389

Merchandising constraints, 293
Merrick, A., 387
Mervyns, 13
METRIC approximation, 341
Micro-merchandising, 340
Mierzwinski, E., 54
Miller, C.M., 194, 254, 270, 271, 274
Millet, I., 54
Millman, H., 349
Mismatch
  execution mismatch, 126, 128
  labor-traffic mismatch, 130
  planning mismatch, 126, 128
Misplaced products, 53, 54
  effect on inventory planning, 4, 64–65
  influence on product availability management, 56–58, 63–67, 71
Misplacement, 53, 57, 59, 65–68, 81, 98, 100, 102
Mixed integer program (MIP), 182
MNL model. *See* Multinomial logit (MNL) model
Mobley, W., 61
Model description, for optimization of assortments suiting diverse customer preferences
  modeling consumer's purchase decision, 296–302
  optimal assortment, properties of, 306–308
  optimization problem, solving of, 308–309
  retailer's assortment optimization, 302–306
Model estimation, effects of firm lifecycle and bankruptcies on, 51
Modeling
  of aggregate-level inventory decisions, 50
  of demand, 81–84, 180, 216, 270
  issues, in retail supply chain management
    allocation policies used at warehouse, 323–324
    key decision, 320–321
    lead times, 323
    modeling demand, 321–323
Models
  for clearance pricing in retail chains
    formulation, 392–398
    mathematical models for, 389–390
    specifications and optimality conditions, 392–398
  of markdown competition, 9, 411–412 (*see also Specific models*)
Mondschein, S.V., 390
Monopolist retailer, 411–414
  with high inventory, 414, 421
  optimal strategy for, 307, 414
  shaping of demand by, 421
Monopolization, 150
Moorthy, S., 79
Morey, R.C., 65, 101, 102
Morgan, L.O., 150, 168, 288
Mowday, R., 61
Muckstadt, J., 326, 333, 339
Muller, E., 396
Multi-item inventory models, 180
Multilocation inventory models, complexity of, 323
Multinomial logit (MNL) model, 158, 184–186, 188–190, 194, 195, 202, 203, 210, 229, 271, 298
  assortment planning with, 183, 190–195
  as choice model, 83, 180, 183, 194, 210, 231, 270, 293–295, 297, 303, 310
  classical, 271
  for customer's selection of product and retailer, 293
  demand estimation of, 212–216
  with homogeneous expected utilities, 294
  MCI model, as alternative to, 213
  optimal assortments under, 202
  predictive accuracy of utilities in, 310
  substitution with, 180
Multiplicative competitive interactions (MCI) model, 213
Multi-store, assortment planning in, 231
Mussa, M., 179
Myopic allocation
  method, 326
  policy, 326, 329

N

Naert, P., 181, 268
Nahmias, S., 7, 67, 82, 271, 319, 330
Nakanishi, M., 213, 215
Nanda, D., 31
Narasimhan, C., 353, 391
National Retail Federation, 8
Negative binomial distribution (NBD), 82, 197, 304, 322, 328, 330, 342
Nelson, P., 271
Nelson, R., 61
Nemhauser, G., 297
Nested logit model, 186, 194, 295
Netessine, S., 29, 31, 126, 128, 131, 140, 179, 180, 372, 391, 411
New products, assessment of market potential of, 22

Newsboy model, in inventory theory, 48
Newsvendor distribution, 83, 91
Newsvendor model, 8, 26, 29
  cost function from, 192
  for rebates in retail (*see* Rebates, in supply chain, manufacturer-to-retailer *vs.* manufacturer-to-consumer)
Nguyen, D., 185
Nie, W.D., 54
Non-basket shopper, 210
Non-captain manufacturers, 150, 151, 155, 157, 161–164, 166–168
Nonparametric models, 81, 89–91, 93
Nonperishable inventory, 81, 86–88, 90
Non-stationary demand, 92, 326
Noonan, P.S., 180
No-purchase decision, 273, 275, 280
  modeling, 278
  operationalization of, 274
  utility of, 186, 273
Nunes, J.C., 178, 255, 265, 269

O
Old Navy, 13, 238, 243
Olenick, D., 351
Online, 25, 69, 80, 143, 180, 206, 230, 238, 239, 243, 256, 257, 311, 350, 369
Online shopping, 350
Operations management, 7, 8, 26, 28, 49, 53, 99, 107, 119, 123–125, 139, 268, 354
  focus of research in, 53
Optimal assortment, 6, 7, 191, 192, 194, 197, 200, 202, 203, 206, 209, 230, 231, 268, 270–272, 281, 283–286, 294–296, 311–312
  properties of, 306–309 (s*ee also* Assortments optimization, for diverse customer preferences)
  selection, 195
  sensitivity to input assumptions, 315
Optimal container packing, 22
Optimal discrete pricing, 398, 401–402
Optimal initial price, 404
Optimal inventory, and maximum profit, determination of, 397–398
Optimal inventory policy, 66
Optimal joint inspection, and replenishment policy, 65
Optimal labor plan, 134
Optimal pricing policy for merchandise, properties of, 406
Optimal stocking, 99, 100, 192
Optimal stocking policy, 107, 329
Optimization model
  additional retailer constraints for, 315
  discussion of, 279–280
Ordering decision, 22, 325, 326, 329
Order-up-to policy, 103, 181, 334
Ordinary least squares (OLS) estimators, 43, 44, 73
Oren, S., 390, 391, 396
Ornamentation assortment, 13
Outlet stores, 13–15
  deeper price markdowns in, 21
Out-of-stocks (OOS), consumer response to, 184
Overstreet, T., 152
Over time, 3, 4, 22, 26, 27, 31, 40, 50, 65, 79, 95, 97, 106, 114, 119, 120, 136, 212, 241, 245, 246, 251, 411–412
  retailer, 49

P
Packaging, redesigning for cost saving, 20
Pantumsinchai, P., 338
Parlar, M., 180, 335
Partially observed Markov decision process (POMDP), 96, 99, 101, 103–105
Partnerships, manufacturer-retailer, 170
Pashigian, B.P., 390
Past priority allocation (PPA), 328
Patel, N.R., 334
Pazgal, A., 409
Peak hour traffic, 132
Penalty cost, 84, 85, 94, 271, 324, 325, 329, 331
Penalty for disutility, 6, 267, 287
Pentico, D., 269
Perfect competition, 7, 296, 315
Performance
  across retail stores, 58
  differences among firms and overtime, causes of, 4
  of retailers, importance of inventory in, 25, 69
Perishable inventory, 81, 84–86, 88–91
Permanent assortment reductions (PAR), consumer response to, 184, 218
Petruzzi, N.C., 170, 354, 372
Phantom products
  and employee turnover, relationship between, 73
  and employee workload, relationship between, 73

Phantom products (*cont.*)
high percentage of, 62
and store manager turnover, relationship between, 73
and training, relationship between, 73
Phantom stockout, 56
Physical inventory, 93, 98, 101, 102, 104
"Pick and pack" warehouse, 19, 20
Pieters, R., 182
Plambeck, E., 20, 88, 92
Planning
aggregate planning, 119, 120
long-term planning, 117
short-term planning, 117
Planning processes, in supply chain, 3
clearance and markdown optimization, 21
cross-channel optimization, 21–22
distribution planning and inventory management, 20–21
logistics planning, 18–20
product design and assortment planning, 17–18
sourcing and vendor selection, 18
Point-of-sale (POS) data, 79–81, 92, 103, 107, 124, 127, 129, 131, 207, 211, 253
Point-of-sale (POS) scanner systems, 7, 320
Porteus, E.L., 8, 86, 88, 322, 349–384, 391
Postponement, 208, 245
Predictions, top-down and bottom-up, 16
Price changes, discrete, 398–402
optimal discrete pricing, 401–402
solution for power function form, 399–400
Price lookup (PLU) codes, 57
Price optimization, integration into retail supply chain decisions, 8–9
Pricing, 253–255
clearance pricing in retail chains
difference from other types of retail pricing decisions, 392–393
discrete price changes, 398–402
mathematical models for, 389–390
model specifications and optimality conditions, 392–393
numerical examples, 402–405
related research, 390–392
trends in, 388–389
dynamic pricing models, variety of, 390
joint inventory and pricing decisions, literature on, 170
optimal pricing policy for merchandise, properties of, 406
problem, 287
solutions, continuous and discrete, comparison of, 402
Pricing decisions, 21, 92
delegation of, 5, 151–157
strategic, 298, 336
Pricing strategy, 9, 248, 253–255
Probabilistic inventory record, 66
Probit demand model, household parameter estimates of, 281
Process design, impact of, 68
Product(s)
misplacement (*see* Misplaced products)
positioning, 179
seasonal, 228, 254, 303
Product attractiveness, 251
Product availability
impact of store execution on, 67
and inventory levels, relationship between, 66, 67
Product availability management, 53
influence of product variety on, 60
research opportunities, 67–69
retail execution problems, 54
factors exacerbating, 58–63
incorporation into existing research streams, 65–66
influence on inventory planning, 63–66
inventory record inaccuracy, 55–56
misplaced products, 56
root causes of, 56–58
role of execution in (*see* Retail execution problems, in product availability management)
Product choices, narrowing of, 298
Product design, 3, 13, 17–18, 256, 257
and assortment planning, 3, 13, 17–18
process architecture, 17
Product introduction, 167, 169, 182, 241, 243, 259
Productivity
labor productivity, 124, 127, 132, 136–138
store productivity, 125, 137
Product line design, 158, 178–179
Product variety
effect on supply chain structures, 203
influence on product availability management, 60
and inventory levels, effect on store sales in Borders Group, 73–75
perception of, 182

and product line design, 178–179
on retail shelf, management of, 265
assortment and stocking problem, 274–280
assortment problem, solution technique for, 283
consumer model, 271–274
future work, 286–288
household scanner panel data, description of, 280–283
literature on, 268–271
optimal assortment, 283–286
Profitability, influence of shrinkage of products on, 62
Profit loss, for incorrect consideration set assumption, 314
Profit margin, 27, 127, 128, 135, 136, 159, 165, 203, 280, 302, 305, 359, 361, 363, 367
Profits
under adjacent/random substitution structure, 209
basket profits, 207, 211
expected, effect of consideration sets on, 313
maximum with optimal inventory determination of, 397–398
reduced, due to misplaced products, 54
Promotion planning, 170
Promotions, cross channel impacts of, 23
Proschan, F., 357
Pryor, K., 338
Pull price promotions. *See* Consumer rebates
Purchase decision(s), 140, 208, 212. *See also* No-purchase decision
about grocery and health-and-beauty products, 266–267
as involving substitution, 177
modeling of, 296–302
stages of, 301
Purchase-incidence, 215, 216, 270
Push price promotions. *See* Retailer rebates

Q
Quality
conformance quality, 127, 128
service quality, 118, 121, 124, 125, 127–129
Quelch, J.A., 178
Quick response, 6, 208, 241–250, 255

R
Radio-frequency identification (RFID), 4, 66, 80, 93, 94, 105–107, 141
Rajagopalan, S., 29
Rajan, A., 390, 391, 396
Rajaram, K., 180, 201, 337
Raju, J.S., 354
Rakesh, A., 390, 391, 396
Ralph Lauren, 13
Raman, A., 26, 30, 54–56, 58–62, 64–75, 93, 98, 106, 124, 125, 127, 128, 141, 143, 204, 218, 245–247, 255
Random yield, 95
Rao, R.C., 354, 410
Rao, S., 391
Rao, V.R.K., 334
Rappold, J.A., 339
Ray, S., 391
Reason, J., 58
Rebates
in supply chain, manufacturer-to-retailer *vs.* manufacturer-to-consumer, 349–384
consumer and retailer rebates together, 356–360
consumer rebate only, 362–365
literature on, 353–356
numerical examples, 365–368
retailer rebate only, 360–362
types of, 351–353, 356, 365, 366, 368
Record inaccuracy. *See also* Inventory record inaccuracy
alternative solution to, 66
shortages caused by, protection against, 65
Redistribution decision, 335
Redman, T., 54
Reinman, M., 338
Rekik, Y., 66, 94, 97, 100, 101
Relative sales per title (RST), 223, 226
Replenishment
attribute-focused rather than product-ocused, 228
costs, 39
joint replenishment problem, 338
lead time, 93, 321, 323, 329, 330, 334, 337, 341
policy, 65, 81, 101–103, 177, 342
Requirements
labor requirements, 119, 120, 129, 140, 142
labor requirements in manufacturing, 119
minimum labor requirements, 117–118
workload requirements, 117

Resale price maintenance (RPM), 152, 153, 171
Re-stocking, 63, 284
Retail assortments, optimization for diverse customer preferences
  DVD player data base, illustrative application for
    comparing model's predictions to retailer's sales data, 310–311
    impact of customer preference structure, 312–314
    optimal assortment *vs.* expected revenue of retailer's assortment, 311–312
  model description
    modeling consumer's purchase decision, 296–302
    optimal assortment, properties of, 306–308
    optimization problem, solving of, 308–309
    retailer's assortment optimization, 302–306
Retail category management. *See* Category captainship; Product variety on retail shelf, management of
Retail chains, clearance pricing in
  discrete price changes, 398–402
    power function form, solution for, 399–402
  mathematical models for, 389–390
  model specifications and optimality conditions, 392–398
    model formulation, 393–398
  numerical examples, 402–405
  related research, 390–392
  trends in, 388–389
Retailer(s)
  customer's choice of, 295
  decision making process, 22, 389
  growth rate of, factors restricting, 41
  market potentials of, 42
  market strength, sensitivity to, 307–308
  monopolistic, 411–414
  non-identical, 326, 327, 329, 330, 334, 335
  overall objective function for, 276
  perfect competition among, 7, 296
Retailer assortment and stocking problem
  basic formulation, 274–277
  modeling no purchase, 277–278
  optimization model and some special cases, 279–280
  reformulation, 278–279
Retailer decisions
  assortment decision, 265
  inventory decision, 265
  strategic, basis for, 294
Retailer rebates
  push price promotions, 353
  in supply chain, 349, 352, 354, 355
  use of, 349, 354, 361, 368
Retail execution problems, in product availability management
  factors exacerbating
    employee effort, 62–63
    employee turnover and training, 60–61
    employee workload, 61–62
    inventory levels, 59–60
    product variety, 60
  influence on inventory planning
    incorporation of execution problems into existing research streams, 65–66
    inventory record inaccuracy, 63–64
    misplaced products, effect of, 64–65
  root causes of, 4, 56–58
Retail firm, assets of, 25
Retail industry, average employee turnover in, 60–61
Retail inventory management. *See* Retail supply chain management, Multilocation inventory models, complexity of
Retail management, academic research focus in, 2–9
Retail master calendar, 16
Retail performance, suboptimal, causes of, 66
Retail shopping decisions, iPACE model for, 296
Retail store
  associates, 4
  employees, 57–59, 62, 63, 68, 122, 127, 137
  labor, 61
  managers, 93, 108, 117
  performance, 122, 123, 125
Retail store demand, modeling of, 322
Retail stores, aggregate echelon inventory positions of, 325
Retail supply chain
  decisions, integration of price optimization into, 8–9
  practices, empirical studies of, 3–5
Retail supply chain management, crucial areas of, 2–3
Retail supply chain management, Multilocation inventory models, complexity of
  modeling issues
    allocation policies used at warehouse, 323–324

key decision, 320–321
lead times, 323
modeling demand, 321–323
periodic review inventory model, general
additional issues, 338–339
batch ordering, 329–330
decentralized environments, 331–333
fashion products, 337–338
lateral pooling, 333–336
lost sales, 330–331
solution methodologies, 326–329
transportation issues, 338
Retail Workbench, at Santa Clara university, 2
Return on assets (ROA), effect of JIT adoption on, 31
RFID. *See* Radio-frequency identification (RFID)
RFID technology, in reducing execution errors, 66
Rhee, B.-D., 355
Ricadela, A., 351
Rinehart, R.F., 54
Risk pooling
advantage of, 321
benefits of, 203, 331
disadvantages of, 326
Roberts, J.H., 298, 301
Robinson, L.W., 334, 335
Rosen, S., 179
Rossi, P.E., 272, 280
Ross Products, as category captain for Safeway in infant formula category, 149
Rout, W., 54
Rudi, N., 180, 335, 336
Ruiz-Diaz, F., 276
Rumyantsev, S., 29, 31
Russell, G.J., 209, 274, 391
Ryan, J.K., 229

S

Sack, K., 25, 26
Safety stock, 39, 65, 305, 330
Sales contraction region, 42, 43, 47, 49
Sales expansion region, 42, 43, 47, 49
Sales growth on inventory turnover, effect of volatility in, 48
Sales growth rate, 3, 25–51
and inventory turnover, relationship between, 26
Sales ratio
correlation with inventory turnover, 27, 38
on inventory turnover, effect of volatility in, 48
negative effect on inventory turnover, 41, 42
positive effect on inventory turnover, 41, 42
regions of, 42
in sales contraction/expansion region, 42
Sales surprise (SS), 3, 27, 29, 34, 37, 38, 49
correlation with inventory turnover, 34
Samroengraja, R., 328
Scarf, H., 84, 88, 326
Schary, P., 184
Scheduling constraints, 22, 133, 134
Schmidt, C.P., 327
Schonberger, R.J., 59
Schrady, D.A., 54
Schrage, L., 26, 39, 326, 327
Schroeder, L., 44
Schwartz, N., 394
Seasonality, 12, 29, 71–74, 82, 92, 114, 139, 223
Sethi, S.P., 391
Shang, K.H., 65
Shapiro, J., 276
Shelf space
allocation models, 181–182
sensitivity to forecast accuracy, 181
availability, 6, 177, 181
constraints, 154, 181, 198, 199, 203, 226, 266, 270, 287
Corstjens and Doyle model for allocation of, 181
and mindspace, battle for, 170
Sheppard, G.M., 54
Sherali, H.D., 287
Sherbrooke, S.C., 341
Shift
lengths, 118, 120, 133
schedules, 120, 123
Shipments, 7, 18–20, 23, 39, 40, 62, 63, 245, 319–321, 323–325, 327–329, 331, 332, 334–336
direct-to-consumer, 14
Shipping
needs, for retailers, 19
to stores, frequency of, 20
Shmueli, G., 269, 273
Shoemaker, R.W., 273
Shopping
basket shopping consumers, 177, 207, 209, 210

Shopping (*cont.*)
decisions, iPACE model for, 296
fixed and variable costs of, 209
non-basket shopper, 210
Shrinkage, 62, 81, 98, 100, 102, 320, 391, 396
Shrinkage of products, influence on store profitability, 62
Shubik, M.J., 153
Shumsky, R., 411
SIC code. *See* Standard industry classification (SIC) code
Siddarth, S., 298
Silver, E.A., 39, 83, 180, 326, 335, 338
Simchi-Levi, D., 338
Simonson, I., 182
Singh, P., 203, 231
Skinner, W., 60
SKU-day-store, 214
SKUs. *See* Stock keeping units (SKUs)
Sloot, L., 269
Slow sellers, 15
Smart shopping carts, 80
Smith and Agrawal model, for inventory planning and assortment selection, 196–197
Smith, S.A., 1–9, 11–23, 67, 82, 188, 196, 209, 230, 255, 270, 278, 293–315, 319–343, 387–407
Song, J.-S., 95, 180, 246
Sourcing, 3, 13, 14, 17–19, 22, 30, 245–250, 255, 257–259, 275
agents, 17
and vendor selection, 3, 18
Space allocation, inter/intra-category, 266
Speranza, M.G., 338
Srinivasan, T.C., 298
Srinivasan, V., 310
Stackelberg leader, 156, 336
Staelin, R., 153, 170
Staffing
overstaffing, 132, 140
understaffing, 131–134, 140
Staffing models
queueing theory based, 139
square-root staffing model, 132, 140
Standard industry classification (SIC) code, 31, 32, 61
Standard & poor's compustat database, 31, 61
Staw, B., 61
Steers, R., 61
Steinberg, R., 179, 391
Steiner, R.L., 150, 167
Stern, L.W., 265
Stock
allocation, 327
cycle, 39, 101
returns, inventory turnover performance with, 30
safety, 39, 65, 305, 330
Stocking
decisions, 6, 82, 180, 333
for retail category management, 286
inventory, back-rooms for, 21
quantity, 53, 192, 197
Stock, J.R., 184, 266, 273, 274
Stock keeping units (SKUs)
at Albert Heijn, groups of, 197, 200, 214, 216, 225–227
at Best Buy, 176, 220–222
at Borders Group Inc, 222–223
categories of, 12
categorization in retail, 97
inter-relationships among, 3, 11
on move, and in stock with Transworld Entertainment, 181
ornamentation, 13
rationalization efforts in General Mills, 149
retailer's breadth/depth, 176
at Tanishq, 223–225
on replenishment, 225
Stockout-based substitution
estimation of, 217
in exogenous demand model, 187–189, 196, 217
Stockouts
consumer response to, 183, 184
forecast accuracy in the presence of, 67
lost revenue due to, 64
phantom stockout, 56
resulting from poor inventory planning/execution, 59
Stokey, N., 390
Storage capacity, 68
Storage needs, additional, 21
Store-bound merchandise, 14
Store execution
impact on product availability, 4
poor, consequences of, 69
poor, examples of, 53
and product variety, relationship between, 75
and storage area size, relationship between, 68
Store inventory, 54, 71, 338

Store manager turnover, and phantom products, relationship between, 73
Store sales
vs. customer entrances, 64
effect of product variety and inventory levels on, study in Borders Group, 73
Stulman, A., 332
Substitution
assortment based, 176, 183, 184, 191, 192, 196, 198, 201, 303
estimation of, 218
behavior, of consumers, 6, 177, 266, 273
between brands, asymmetries in patterns of, 283
consumer driven, 182–184
between consumers, heterogeneity in patterns of, 283
dynamic, 191, 193, 201, 202, 294
variety under, 202
incremental demand arising from, 303–304
involving, 177
rates in exogenous demand models, demand estimation of, 217–218
in retail, classification of, 32
stocking, adjacent, 209
stockout-based, 183, 184, 189, 191, 196, 198, 199, 218, 294, 303, 304
estimation of, 217
Sudharshan, D., 158, 165, 213
Sugrue, P.K., 322
Sukumar, R., 202, 295
Supply chain(s)
decentralized, assortment planning in, 203–204
description, 13–15
performance, opportunities for improvement of, 23
planning processes
clearance and markdown optimization, 21
cross-channel optimization, 21–22
distribution planning and inventory management, 20–21
logistics planning, 18–20
product design and assortment planning, 17–18
sourcing and vendor selection, 18
planning processes, details of, 3
rebates in
consumer and retailer rebates together, 356–360
consumer rebate only, 362–365
literature on, 353–356
numerical examples, 365–368
retailer rebate only, 360–362
structures, effect of product variety on, 203
Supply chain management, constraints on, 13
Supply chain management, multi-location inventory models for
modeling issues
allocation policies used at the warehouse, 323–324
key decision, 320–321
lead times, 323
modeling demand, 321–323
periodic review inventory model, general
additional issues, 338–339
batch ordering, 329–330
decentralized environments, 331–333
fashion products, 337–338
lateral pooling, 333–336
lost sales, 330–331
solution methodologies, 326–329
transportation issues, 338
Supply chain profits
constant-split property of
consequence of, 363
rationale behind, 364
split under consumer rebates, 352
Svoronos, A., 329
Swait, J., 273
Swaminathan, J.M., 66, 97, 100
System inventory, 54, 59, 64, 67
and actual inventory, absolute value difference between, 54, 59

T

Tagaras, G., 334
Tallman, J., 54
Talluri, K., 213, 217, 390, 409
Tang, C.S., 180, 245, 258
Tanishq (Indian jewelry retailer), assortment planning in practice, 177
Target stores, 13
Taylor, T.A., 179, 349, 350, 354
Tayur, S., 63, 339
Tellkamp, C., 66
Theft, as source of inventory record inaccuracy, 66
Third party logistics (TPLs), 15
Time-to-market, 250
Todor, W., 61
Toktay, L.B., 5, 147–171

Ton, Z., 4, 53–75, 81, 93, 98, 127, 128, 131, 135
Top-down analysis, 15
Tractability, assortment planning models for, assumption made in, 188
Traffic counters, 80, 106, 124, 125, 128, 140, 141
"Traffic generators", 276
Traffic to associate ratio, 124
Training, 4, 58, 60–61, 68, 73, 119, 136, 138, 139, 259, 389
  and phantom products, relationship between, 73
Transaction errors, 81, 98, 102
Transportation costs, 39, 332
Transportation issues, in retail supply chain management, 338
Transshipments, 333–336
Trucks, use in delivery of shipments, 14, 19, 23
True customer demand, 129
Tsay, A.A., 8
Turnover
  employee turnovers, 4, 58, 60–61, 68, 71, 73, 74, 124, 128, 135–139
  turnover-performance link, 137
Type of labor, 123, 126

U
Ukovich, W., 338
Uncertainty
  labor uncertainty, 129
  traffic uncertainty, 129
Uniqlo/Fast Retailing, 239, 244
United States Tobacco Co. (UST), antitrust row with Conwood, 150
Unmet demand
  backlogged, 324
  backordering, 328, 330, 332, 342
  lost sales, 330–331
  meeting of, 333
Urban, G.L., 266, 273
Urban, T.L., 181
U.S. retail sector, effects of firm size and sales growth rate on inventory turnover performance in
  adjusted inventory turnover, 34–38
  data description, 31–34
  directions for future research, 49–51
  hypotheses
    effect of firm size on inventory turnover, 38–40
    effect of sales ratio on inventory turnover, 41–43
  literature on, 28–31
  model, 43
  results, 43–49
Utility, of purchasing from retailer, 298, 299

V
Vaidyanathan, R., 175–231
Van Dijk, A., 268
Van Donselaar, K., 326
Van Dyck, D.A., 272, 280
Van Herpen, E., 182
Van Ryzin, G., 158, 178, 180, 185, 190–194, 197, 200, 207, 213, 217, 219, 254, 270, 271, 294, 295, 303, 304, 306, 390, 396, 409, 412
Van Ryzin and Mahajan model, 190–195, 201, 203, 204
  demand process in, 213
Variability
  customer arrival process variability, 121
  inter-day traffic variability, 129–131
  intra-day traffic variability, 129
  traffic variability, 129–131, 135
Variables, for each retailing segment, summary statistics of, 34, 35
Variety. *See also* Product variety
  consumers' perception of, 177, 230
  perception of, 182, 229
Vendor capacity planning, 18
Vendor managed inventory, 1, 69, 170, 231
Vendors
  brand management by, 6
  social compliance by, 18
Vendor selection, 3, 18
Venkataramanan, M.A., 265–289
Verrijdt, J.H.C.M., 339
Video monitoring, 80
Vilcassim, N.J., 274
Villas-Boas, J.M., 158
Virtual allocation policy, 324, 328
Visual and marketing group, 17
Vives, X., 153
Vlachos, D., 334
Volatility, effect in sales ratio on inventory turnover, 48

W
Walmart, 20
Walter, C., 184

Wang, Y., 153, 155–158, 162
Warehouse
  allocation policies used at, 323–324
  clubs, growth of, 266
  demand
    approximation of, 327, 330
    variability, 330
  echelon stock of, allocation of, 326
  holding costs, 328
  periodic replenishment at, 327
  "pick and pack", 19, 20
  space, sharing of, 13–14
  stock-less, 326
Wecker, W.E., 67, 81, 82
Weeks-of-supply (WOS), 21
Weibull distribution, 83, 84, 88
Weitz, B.A., 219, 228
Whang, S., 409–423
Wharton research data services (WRDS), 31
Wheelwright, S.C., 41
White, E., 60
Whitin, T.M., 180
Whybark, D.C., 54
Wijngaard, J., 326
Wilson, L.W., 54
Winston, W.L., 391
Winter, S., 61
Wittink, D.R., 310
Woellert, L., 54
Wolfe, H.B., 391
Woolsey, G., 54
Workers
  full-time workers, 114, 115, 117, 118, 120, 135, 138, 139
  part-time workers, 114, 115, 117–119, 139
  seasonal workers, 114, 115, 118, 119
  temporary workers, 123, 135, 138
Workforce management
  software, 116
  solutions, 4, 141
  technologies, 114
  tools, 115
Working capital management, 4
Workload, 61, 62, 68, 117, 128, 140, 141
World Co. (Japan), design-to-shelf lead time at, 204
Wrangler and Lee, as category captain jeans category, 149
Wu, C.F.J., 213

X
Xiao, B., 409

Y
Yano, C., 170
Ye, J., 170
Young, S.T., 54

Z
Zara (Spain), design-to-shelf lead time at, 204
Zara/Inditex, 239
Zero-balance walks, 99, 102
Zhao, H., 336
Zhao, W., 390
Zheng, Y.-S., 390
Zheng, Y.S., 327, 338, 342, 390
Zinn, W., 184
Zipkin, P.H., 81, 93, 246, 326, 338
Zotteri, G., 67
Zufryden, F.S., 202, 274, 295