UNIVERSITY OF CALIFORNIA

Los Angeles

Essays on Supply Chains Facing Competition from Gray Markets

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy in Management

by

Foad Iravani

2012

ABSTRACT OF THE DISSERTATION

Essay on Supply Chains Facing Competition from Gray Markets

by

Foad Iravani

Doctor of Philosophy in Management
University of California, Los Angeles, 2012
Professor Reza Ahmadi, Chair

This dissertation comprises of three chapters. The first chapter describes the development and implementation of a hierarchical framework for organizing the process for producing tax software at a leading tax software company in the United States. Every year, companies that produce commercial tax preparation software struggle with thousands of state and federal changes to tax laws and forms. Three competitors dominate the market with its short selling season, and release delays slash profits. Tax authorities issue updates August-December, and all changes must be processed and incorporated before year end. Systematic resource allocation and process management are crucial yet problematic due to the volume and complexity of changes, brief production timeframe, and feedback loops for bug resolution. A leading tax software provider asked us to formulate systematic approaches for managing process flow and staffing development stages with the goal of releasing the new version on time at minimum cost. To that end, we develop deterministic models in chapter 1 that partition tax forms into dedicated groups and determine staffing levels. Parti-

tioning tax forms into groups simplifies workflow management and staffing decisions. To provide a range of resource configurations, we develop two modeling approaches. Numerical experiments show that our models capture the salient features of the process and that our heuristics perform well. Implementing our models reduced company overtime hours by 31% and total resource costs by 13%.

The second and third chapters of the dissertation focus on supply chains that face competition from gray markets. Manufacturers in many industries have been challenged with the resale of their products in unauthorized distribution channels. Also known as parallel importation, gray markets are primarily driven by price differentials. When manufacturers release their products in different markets, they choose the price in each market based on consumer purchase power, sensitivity to price changes, and the overall economic conditions. This practice of price discrimination enables manufacturers to take advantage of differences between markets and maximize profit. However, price discrimination can potentially lead to the emergence of gray markets. Gray marketers buy products in the markets with lower prices and import them to markets with higher prices to sell below manufacturer price, thereby undermining the pricing structure of manufacturers and damaging brand value. The rapid growth and implications of gray markets have made it imperative for companies react to the diversion of their products to gray markets properly and take gray markets into consideration when making important strategic decisions.

Chapter 2 analyzes the impact of parallel importation on a price-setting manufacturer that serves two markets with uncertain demand and characterizes the appropriate policy that the manufacturer should adopt against parallel importation. We show that adjusting prices is more effective in controlling gray market activity than reducing product availability, and that parallel importation forces the manufacturer to reduce the price gap while demand uncertainty forces the manufacturer to lower

prices. We illustrate that ignoring demand uncertainty can take a significant toll on the manufacturer's profit. Furthermore, we explore the impact of market conditions (such as market base, price sensitivity, and demand uncertainty) and product characteristics (such as ”fashion” vs. ”commodity”) on the manufacturer's policy. We also provide managerial insights about the value of strategic decision-making by comparing the optimal policy to the uniform pricing policy that has been adopted by some companies.

Chapter 3 extends the Stackelberg game to analyze the role of providing service as a non-price mechanism in coping with parallel importation in a deterministic setting. We observe that the price and service competition leads to a Prisoner's Dilemma equilibrium: both players would be better off if the parallel importer does not offer service. We show that parallel importation forces the manufacturer to provide more service in both markets. Although the manufacturer achieves higher profits with providing service, the price gap may be higher or lower than when no service is provided. We find that a little service can go a long way; even if the contribution of service to total revenue in the absence of gray markets is not very large, the manufacturer can significantly increase total profits by providing service when facing gray market activities. Also, when consumers become indifferent between the manufacturer and the parallel importer, the manufacturer uses the service mechanism to differentiate herself from the parallel importer. However, when the parallel importer free rides on manufacturer service, the manufacturer provides lower service. We also consider the manufacturer's service policy towards customers who buy the product from the gray market, and show that the manufacturer may achieve higher profit by allowing more parallel imports and charging gray market customers a service fee.

The dissertation of Foad Iravani is approved.

Rakesh Sarin

Sriram Dasu

Guillaume Roels

Christiane Barz

Reza Ahmadi, Committee Chair

University of California, Los Angeles

2012

In Memory of My Father, Aliakbar,

May He Rest In Peace

Contents

List of Figures

List of Tables

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Professor Reza Ahmadi for his guidance and support throughout my PhD study. It was indeed a pleasure for me to work under his supervision and mentorship. His patience, encouragement, availability, valuable comments, and openness during our discussions were greatly instrumental in accomplishing my research. I also would like to thank Professor Sriram Dasu for his sincere support and encouragement. I have benefited from his guidance and insightful comments a lot. He taught me how to think broadly and see things from a different perspective.

I would like to thank Professor Rakesh Sarin, Professor Guillaume Roels and Professor Christiane Barz for agreeing to serve on my dissertation committee and providing me with useful comments about research and teaching. I am grateful to my friends in the DOTM PhD Program at Anderson, Morvarid Rahmani, Jaehyung An, Dimitrios Andritsos, Boyoun Choi, George Georgiadis, Aparupa Das Gupta, Priya Mittal, Sandeep Rath, Paul Rebeiz, and Wei Zhang for their encouragement and good memories that we have had together. I sincerely thank the Harold and Pauline Price Center for Entrepreneurship and the Easton Technology Leadership Program at the Anderson School for their generous financial support. Also, I would like to thank Lydia Heyman who works tirelessly and beyond her responsibilities to help students succeed in the PhD program.

I wish to acknowledge the sincere support and encouragement from Hamed Mamani who has been my best friend for a very long time. He truly symbolizes a genuine and trustworthy friend and I wish all the best for him and his family.

Last but not least, I extend my deepest gratitude to my family for supporting me throughout my graduate studies and enduring my long absence from home. My mother has dedicated her life to taking care of my siblings and I and helping us be successful. Whatever I have accomplished I owe it to her.

Chapter 1 is a version of Iravani, F., S. Dasu, and R. Ahmadi. 2012. A Hierarchical Framework for Organizing a Software Development Process. *Accepted for publication in Operations Research*.

Chapter 2 is a version of Iravani, F., H. Mamani, and R. Ahmadi. 2012. Coping with Gray Markets: The Impact of Market Conditions and Product Characteristics. *working paper*.

Chapter 3 is a version of Iravani F., S. Dasu, and R. Ahmadi. 2012. Beyond Price Mechanisms: How Much Can Service Help Manage the Competition from Gray Markets. *working paper*.

VITA

2003	Bachelor of Science, Industrial Engineering Sharif University of Technology, Tehran, Iran
2007	Master of Applied Science, Mechanical and Industrial Engineering University of Toronto, Toronto, Canada
2007–2012	PhD Student The Anderson School of Management University of California, Los Angeles

PUBLICATIONS

1. Iravani, F., S. Dasu, and R. Ahmadi. 2012. A Hierarchical Framework for Organizing a Software Development Process. *Accepted for publication in Operations Research*

2. Iravani, F., B. Balcioglu. 2008. On Priority Queues with Impatient Customers. *Queueing Systems.* 58(4) 239-260.

3. Iravani, F., B. Balcioglu. 2008. Approximations for the M/GI/N+GI Type Call Center. *Queueing Systems.* 58(2) 137-153.

Chapter 1

A Hierarchical Framework for Organizing a Software Development Process

1.1 Introduction

Consumer tax preparation applications comprise a profitable niche in software development, an industry which the U.S. Bureau of Labor Statistics projects to remain the third-fastest growing in the American economy through the next decade. Growth specifically in tax preparation software has been bolstered by Internal Revenue Service (IRS) efforts to achieve an 80% electronic filing rate for major returns by 2013. Reflecting those efforts, 2011 sales of consumer tax programs grew by as much as 20% with 40 million taxpayers using them to file their returns (Electronic Tax Administration Advisory Committee, 2011).

The idea for this project grew from conversations with a software engineer who is studying for an MBA while working for one of the nation's largest tax preparation software providers. The engineer characterized the market as fiercely competitive, subject to a short and fixed sales window, and committed to a product that becomes obsolete every year. Three companies dominate this high-pressure arena, racing one another each year to incorporate changes to laws and forms, test their new versions, and release bug-free products for the upcoming tax season. Obviously, early release confers an advantage on a tax software development company (henceforth TSDC) by helping it maximize market share. Thus, delays can result in significant losses.

The development process for tax software is complex and consists of multiples stages that incorporate thousands of revisions. Some changes are trivial while others command significant developer time. Adding to the challenge is the fact that state and federal authorities only begin announcing their changes in early August and continue doing so through December, while mid-December marks the start of the tax software sales season. Each stage in the development process contains built-in feedback loops that interrupt workflow in order to correct errors. To release the application on time and control development costs, therefore, TSDC must effectively manage the process and accurately determine staffing levels. However, in the course of analyzing and observing a development cycle, we noted that tasks were assigned on an ad hoc basis, and staffing levels were subject to the vagaries of individual power and influence. Not surprisingly, the firm's bottom line chronically suffered from significant overtime costs, and TSDC found it difficult to achieve a timely release.

Clearly, the large number of forms calls for a staffing plan driven by an effectively organized development effort that simplifies day-to-day operations management, improves coordination among different stages, and facilitates information flow. To that end, we propose a model that restructures the development effort by sorting tax forms into multiple independent groups and determines staffing levels throughout the process. In answer to the challenge of making the two decisions simultaneously, we employ a hierarchical approach in which we first form the groups and then allocate sufficient resources (we use the terms *resource* and *employee* interchangeably) to ensure timely completion. The decision to create groups is supported by studies in software development (e.g., Brooks 1975, Cusumano 1997) that demonstrate benefits such as increased effectiveness and enhanced quality arising from the division of tasks into groups.

Although we develop a hierarchical planning framework to organize tax software development, our approach could also be applied to development processes in other highly-regulated industries that periodically revise a product, face tight deadlines, and involve processing requirements similar to those we encounter here (e.g., Ahmadi et al. 2001). Every year, for example, the Centers for Medicare & Medicaid Services publishes regulatory updates to payment rules, standard assessments, and resource utilization group and case mix calculations. Companies producing Medicare and Medicaid billing software must incorporate these changes before October as efficiently as they can. Aerospace, cellular communication, healthcare, and enterprise resource planning (ERP) face similar development challenges. Development teams in these disciplines typically work in parallel, sharing a common deadline to complete their design and development tasks. Thus, task assignment and resource allocation are widespread issues.

The remainder of this chapter is organized as follows. In Section 1.2, we review the relevant literature. In Section 1.3, we describe the problem in detail, explain our modeling assumptions, and propose approximations to capture the effect of feedback loops on process completion time. We introduce our notation and formulate our models in Section 1.4, describe their solution procedures in Section 1.5, and investigate the performance of the hierarchical approach and the solution procedures using numerical experiments in Section 1.6. Section 1.7 explores the implementation of our models at TSDC. Section 1.8 concludes the chapter with a summary of our research.

1.2 Literature Review

The development process for tax preparation software has elements of reentrant flow shops (Graves et al. 1983) and other product development projects. On one hand, the process is repetitive because the software is produced each year and each version encompasses multiple jobs, all facing the same deadline. On the other hand, task requirements vary significantly each year and TSDC cannot just repeat the same process. Every version of the application, therefore, can be thought of as a project.

In addition to traditional project management techniques (Tavares 1998), several other approaches have been proposed for reducing the duration of product development projects by changing the sequence of development activities, overlapping activities, and changing the flow of information among developers (Krishnan et al. 1997, Smith and Eppinger 1997, Carrascosa et al. 1998, Loch and Terwiesch 1998, Roemer et al. 2000, Ahmadi et al. 2001, Roemer and Ahmadi 2004). In our setting, however, the sequence of activities is fairly straightforward, offering minimal opportunity for modification or increased overlap. Therefore, we seek to achieve the desired duration of the development process by creating groups to facilitate the flow of the process and then staffing the stages properly.

Partitioning the development effort and allocating tasks among groups is a common software industry practice that substantially influences project duration, software quality, development cost, and reusability. Cusumano (1997) and Cusumano et al. (2003) survey techniques employed by leading software companies to partition and manage large projects. Such studies explore ways to divide complex tasks into manageable modules (Shaw and Clements 2006). In our problem, we define an index for grouping tasks based on similarities between the amount of work they require (also used in Ahmadi and Matsuo 2000). We then staff the groups to ensure timely completion.

In software engineering, an interesting dilemma is whether to assign one or two developers to work on a module. Empirical evidence suggests that assigning multiple developers increases the time and effort needed to develop a module but decreases the time and effort needed to integrate it. Dawande et al. (2008) develop a mathematical model to find conditions under which one approach supersedes the other. We assume that jobs can be divided among the developers, which is an approximation for tractability.

Browning and Ramasesh (2007) provide a comprehensive review of network-based process models for managing product development activities, and they cite very few papers that are concerned with resource allocation. Joglekar and Ford (2005) study ways to dynamically shift a finite pool of resources across different stages of the process, using a procedure that is considerably simpler than ours. Though we are not concerned with dynamically reassigning resources, Joglekar and Ford intriguingly suggest that complexity diminishes the value of dynamic resource allocation.

In software development, another class of resource allocation problems addresses the optimal allocation of resources among competing priorities. Ji et al. (2005) explore optimal allocation between software construction and debugging to maximize quality. Kumar et al. (2006) look at the tradeoff between the benefits of adding a new feature and the risk of introducing new bugs with it.

Queuing network models also have been proposed for staffing projects. Adler et al. (1992) consider multiple product development projects that proceed through nodes representing departments or functional capabilities. Queuing models assume that the design facility is continually receiving design projects (Adler et al. 1992) or maintenance requests (Asundi and Sarkar 2005, Kulkarni et al. 2009, Feng et al. 2006), each with its own deadline. Although we have multiple tasks in the process, they are all part of a single project.

Resource scheduling problems also arise during software execution (Hos and Shin 1997) where the challenge is to complete a job on time while allocating the elements of the job in real time to a set of resources. Complexity in these problems stems from the nature of the precedence relationships and interdependencies among tasks. In our problem, the flow patterns and precedence relationships are simple. The challenge stems from the large number of tasks that need to be performed.

Kekre et al. (2009) address an interesting problem with features similar to this work. They analyze the workforce management of multistage check-clearing operations at a commercial bank. They use simulation–optimization to capture the tradeoff between efficiency and the risk of delayed checks resulting from excessive workforce reduction. Our problem involves staffing the stages of a process and deciding to which group each form should be assigned. The very large number of forms produces a correspondingly large number of scenarios for assigning forms to groups and allocating resources, even for three to five groups. Therefore, simulation–optimization is not well-suited to solving practical instances of our problem, though one could use the technique to explore different staffing scenarios for a fixed assignment of forms to groups.

1.3 Problem Characteristics and Modeling Assumptions

Maintaining a tax preparation application encompasses multiple processes throughout the year. Some processes, such as maintaining the software engine at the core of the application, proceed independently of revisions to the tax code. Others are concentrated into the quarter before the impending tax season. These processes arise from the IRS and each state taxing authority independently constructing tax laws and forms. The number of changes is immense; they are released dynamically, starting in August and continuing into the tax software selling season; and their implementation encompasses highly variable degrees of difficulty. Five major stages dominate the workflow (see Figure 1.1).

1. The Image Development Group (IDG) evaluates all tax form and document changes and creates an image of each form.

2. Calculation (CALC) elucidates and tests the computations that underlie each form. CALC is the most important stage and carries the most amount of work.

3. Electronic Filing (EF) develops electronic versions of the forms, employing functions and macros based on the structure of each form and the fields it contains.

4. The Interview team designs the user interface that guides the consumer through the software.

5. Integration and Final Test (I&FT) incorporate the forms into the application and put each component through exhaustive trials. Integration also designs the buttons, menus, and toolbars. Final Test sends each error back to the team that introduced it.

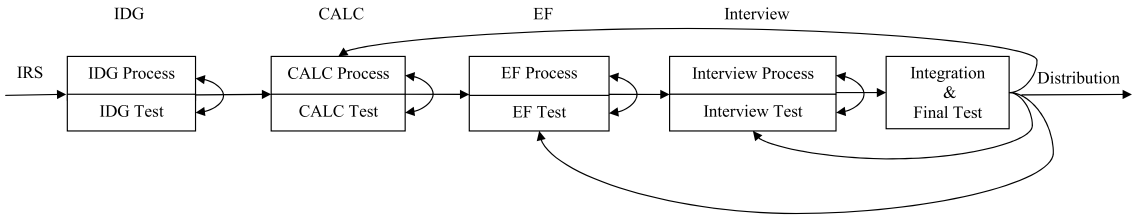


Figure 1.1: Tax software process line

The backward arcs in Figure 1.1 indicate feedback from I&FT to earlier, upstream stages. When an error is found in a form during I&FT, it is returned to the appropriate stage—CALC, EF, or Interview—for correction. Each stage, furthermore, consists of two substages, the first for processing forms and the second for testing the forms internally. The internal tests may return a form to their corresponding process for rework. Taking into account the number of changes to be incorporated and their dynamic introduction, workflow management and coordination becomes a formidable task.

1.3.1 Modeling Assumptions

To effectively control this development effort, we need to sort the forms into manageable groups, staff each group, and develop rules for sequencing the tasks. The work volume, dynamic arrivals, feedback loops, and processing time variability make it nearly impossible to identify good sequencing rules. Hence, we develop models that support tactical decisions about grouping and staffing, and we ignore operational issues such as sequencing and scheduling.

We make two simplifications while developing the models, one based on work forecasts and the other on downstream stage idle times during rework. Having observed that the system as a whole is never idle due to lack of work, we can disregard pattern details associated with forms arriving sporadically. Instead, we base grouping and staffing decisions on work forecasts. If the forecast were to change significantly during the season, TSDC could always revisit the models and alter staffing levels. TSDC managers who participated in our study, validated this simplification based on their past experience, and we obtained objective validation via our computational experiments.

The second simplification concerns feedback loops, which increase the amount of work at each stage, introduce additional uncertainty into processing time requirements, and may temporarily put a stage out of action while a stage downstream creates a rework loop. Given our interest in completion time, the latter impact is the most problematic. The likelihood of inserted idle times depends on the initial amount of work; when the amount of work is high, inserted idle times are unlikely. In the absence of inserted idle times, we are able to approximate the effect of feedback loops with a no-loops process, provided we suitably modify the processing times at each stage.

1.3.2 Approximating Feedback Loops

In the upper half of Figure 1.2, we show an alternate depiction of the development process that separates the sub-stages. In the lower half, we depict the no-loops approximation to the original process. Because I&FT never sends a form back to IDG, we eliminated the IDG process and its internal test from the figure.

In our approximation, the processing time distribution of a form at each stage is determined by the original processing time distribution of the form and the distribution of the number of times the form revisits the stage. In the original process, let $\rho_k(i,n)$ represent the probability that the number of times form i visits stage k is n, and let $f_k(i,t)$ denote the distribution of the processing time t of form i at stage

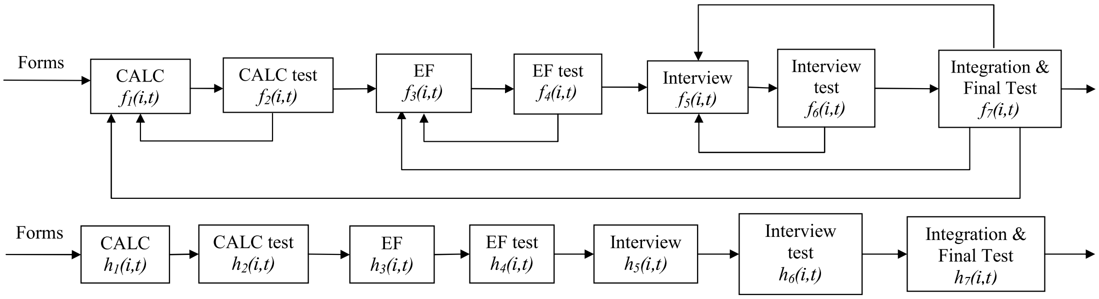


Figure 1.2: The original process and the no-loops approximation

k. In the no-loops process, the processing time distribution for form i at stage k is defined as $h_k(i,t) = \sum_n \rho_k(i,n) f_k^{(n)}(i,t)$, in which $f_k^{(n)}(i,t)$ is the n–fold convolution of $f_k(i,t)$. In Section 1.10, we show that for a two-stage system with deterministic processing times, the completion time for the no-loops process converges to that of the original process as the number of forms becomes large. Our numerical experiments also show that the approximation performs well.

1.4 Models

Based on the foregoing assumptions, we first formulate the Monolithic Grouping and Resource Allocation Model (MGRAM) that simultaneously sorts tax forms into groups and allocates resources to the groups. We find that MGRAM is strongly NP-hard and solvers such as Industrial Lingo are consequently ineffective in solving moderate-size problems that involve more than 500 forms. Although realistic instances of the monolithic model cannot be solved to optimality in reasonable amounts of time, discussing the formulation of the monolithic model is useful for setting the stage for the heuristic development in Section 1.4.2.

We then formulate three models and combine them into a hierarchical approach to heuristically solve the monolithic model. The three models we formulate calculate indices that measure the similarity between processing times for forms, use the indices to assign forms to groups, and allocate resources to the groups with the goal of releasing the software on time. This approach is similar to the classic hierarchical production planning approach in Hax and Candea (1984) if one pictures the groups of forms as families of similar items. While Hax and Candea focus on disaggregating the production plan of product types to product families and items, we are concerned with staffing the production process. The three models used in the hierarchical approach (see Figure 1.3) are the Grouping Index Model (GIM), the Grouping Model (GM), and the Resource Allocation Model (RAM).

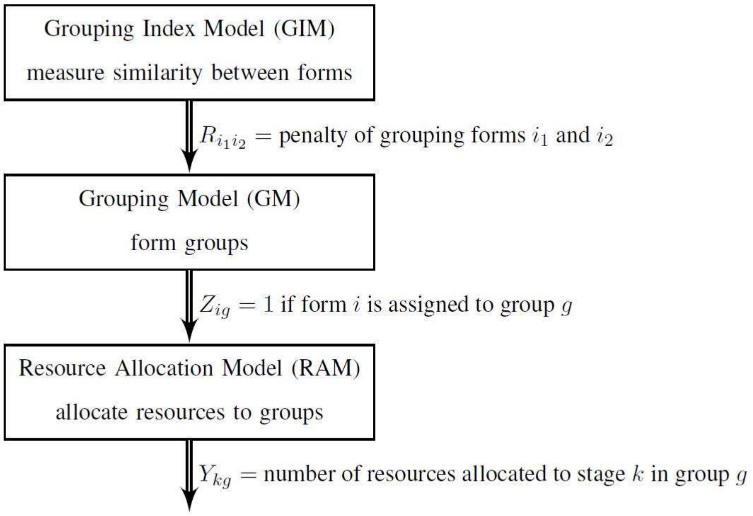

Grouping Index Model (GIM)
measure similarity between forms
$R_{i_1 i_2}$ = penalty of grouping forms i_1 and i_2
Grouping Model (GM)
form groups
$Z_{ig} = 1$ if form i is assigned to group g
Resource Allocation Model (RAM)
allocate resources to groups
Y_{kg} = number of resources allocated to stage k in group g

Figure 1.3: Models of the hierarchical approach for grouping and resource allocation

1.4.1 Monolithic Grouping and Resource Allocation Model (MGRAM)

Using two approaches to formulating MGRAM, we obtain an upper value (MGRAM_1) and a lower value (MGRAM_2) for the optimal number of resources. The two approaches differ in their method for approximating the time to complete the process. Let i index the set of forms $\mathcal{I} = \{1, \ldots, I\}$, k index the set of process stages $\mathcal{K} = \{1, \ldots, K\}$, and g index the set of groups $\mathcal{G} = \{1, \ldots, G\}$ to be formed. Define P_{ik} as the processing time of form i at stage k, which in fact represents the total time needed to perform a number of divisible tasks for this form. If Y_{kg} is the number of resources allocated to stage k of group g, and form i is processed in group g, then the *effective* processing time of form i at stage k is roughly $\dfrac{P_{ik}}{Y_{kg}}$. Let $T_{ig} = \max_{k \in \mathcal{K}} \dfrac{P_{ik}}{Y_{kg}}$ be the maximum effective processing time of form i across all stages in group g. A simple yet tractable approximation for the time to complete all the forms in group g is $\sum_i T_{ig}$ in which the summation encompasses all forms that are assigned to group g. We find that this estimate, motivated by Proposition 1 in the following paragraph, continues to be accurate in our numerical experiments even with feedback loops. Section 1.9 provides proofs for all propositions.

Proposition 1.1. *Suppose $G = 1$ and $Y_k = 1$ for all $k \in \mathcal{K}$. Then, the time to complete the processing of forms is at most* $\sum_{i \in \mathcal{I}} max_{k \in \mathcal{K}} \{P_{ik}\} + (K-1) max_{i \in \mathcal{I}} \{max_{k \in \mathcal{K}} \{P_{ik}\}\}$ *and* $\liminf_{I \longrightarrow \infty} \frac{\sum_{i \in \mathcal{I}} max_{k \in \mathcal{K}} \{P_{ik}\}}{\sum_{i \in \mathcal{I}} max_{k \in \mathcal{K}} \{P_{ik}\} + (K-1) max_{i \in \mathcal{I}} \{max_{k \in \mathcal{K}} \{P_{ik}\}\}} = 1.$

As the number of forms becomes large relative to the number of stages, $(K - 1)\max_{i \in \mathcal{I}} \{\max_{k \in \mathcal{K}} \{P_{ik}\}\}$ becomes relatively small and the completion time asymptotically approaches $\sum_{i \in \mathcal{I}} \max_k \{P_{ik}\}$. In our models, each group processes at least 1,000 forms, which allows us to use $\sum_{i \in \mathcal{I}} T_{ig}$ in MGRAM$_1$ to approximate the time to complete all forms in group g. We set decision variable Z_{ig} equal to 1 if form i is assigned to group g, but otherwise equal to 0. Thus, we formulate MGRAM$_1$ as follows:

$$(\text{MGRAM}_1) \quad \min \sum_{k \in \mathcal{K}} \sum_{g \in \mathcal{G}} w_k Y_{kg} \tag{1.1}$$

$s.t.$

$$T_{ig} \geq \frac{P_{ik}Z_{ig}}{Y_{kg}} \qquad \forall i \in \mathcal{I}, \quad \forall k \in \mathcal{K}, \quad \forall g \in \mathcal{G}, \tag{1.2}$$

$$\sum_{i \in \mathcal{I}} T_{ig}Z_{ig} \leq D \qquad \forall g \in \mathcal{G}, \tag{1.3}$$

$$\sum_{g \in \mathcal{G}} Z_{ig} = 1 \qquad \forall i \in \mathcal{I}, \tag{1.4}$$

$$Y_{kg} \geq 1 \text{ and integer} \quad \forall k \in \mathcal{K}, \quad \forall g \in \mathcal{G}, \tag{1.5}$$

$$Z_{ig} \in \{0, 1\} \qquad \forall i \in \mathcal{I}, \quad \forall g \in \mathcal{G}. \tag{1.6}$$

Objective function (1.1) minimizes total resource cost. Constraint (1.2) defines the maximum effective processing time of form i in group g. Constraint (1.3) ensures that the sum of the maximum effective processing time across all stages of all forms is less than the deadline in each group. Constraint (1.4) guarantees that each form is assigned to one group only. Finally, constraints (1.5) and (1.6) ensure that resources are positive integers and that Z_{ig} is a binary variable.

For MGRAM$_2$, we require that the total effective processing time for all forms *at each stage* be less than the time to deadline. In a flow shop, the maximum total effective processing time across all stages is a lower value for the completion time. The resources prescribed by MGRAM$_1$ will always be greater than those prescribed by MGRAM$_2$. The formulation of MGRAM$_2$ is the same as MGRAM$_1$, except that we replace constraints (1.2) and (1.3) with:

$$\sum_{i \in \mathcal{I}} \frac{P_{ik}Z_{ig}}{Y_{kg}} \leq D, \quad \forall k \in \mathcal{K}, \quad \forall g \in \mathcal{G}.$$

With Proposition 2, we assert the complexity of the two MGRAMs.

Proposition 1.2. *MGRAM$_1$ and MGRAM$_2$ are strongly NP-hard.*

1.4.2 Hierarchical Grouping and Resource Allocation Models

In this section, we formulate the grouping and resource allocation models that constitute our hierarchical approach to solving the MGRAMs heuristically. We also apply the models to estimating completion times. Our hierarchical models, GRAM$_1$ and GRAM$_2$, correspond to the monolithic models, MGRAM$_1$ and MGRAM$_2$. Beginning with the Grouping Index Model (GIM), we develop an index of similarity between two forms. We then use this index in the Grouping Model (GM) to assign forms to groups. Once the groups are formed, we use the Resource Allocation Models (RAM$_1$ and RAM$_2$ corresponding to GRAM$_1$ and GRAM$_2$) to determine staffing levels.

Grouping Index Model (GIM)

The first component of our hierarchical framework provides a means of measuring the similarity between forms. The idea behind the GIM is the notion that if form i_1 and form i_2 require proportional processing times in each stage (i.e., $\dfrac{P_{i_1k}}{P_{i_2k}}$ is the same for all stages k) then it is suitable to assign these forms to the same group. Suppose that forms i_1 and i_2 are to be processed in one group by Y_k resources at stage k (g is suppressed) and define

$$\Lambda_{i_1} = \max_{k\in\mathcal{K}} \left\{ \frac{P_{i_1k}}{Y_k} \right\}, \quad \Lambda_{i_2} = \max_{k\in\mathcal{K}} \left\{ \frac{P_{i_2k}}{Y_k} \right\}.$$

We define *balance loss* as the total idle time arising from the difference between the maximum effective processing times and the effective processing times at other stages, which is equal to $\sum_{k\in\mathcal{K}}\left\{\left(\Lambda_{i_1}-P_{i_1k}/Y_k\right)+\left(\Lambda_{i_2}-P_{i_2k}/Y_k\right)\right\}Y_k$ and can be written as

$$\left(\Lambda_{i_1}+\Lambda_{i_2}\right)\sum_{k\in\mathcal{K}}Y_k-\sum_{k\in\mathcal{K}}\left(P_{i_1k}+P_{i_2k}\right). \tag{1.7}$$

The smaller the balance loss, the more suitable it would be to place forms i_1 and i_2 in the same group. With decision variables Y_k and Λ_i, we define the GIM for forms i_1 and i_2 as

$$\text{(GIM)}\quad \min\ \left(\Lambda_{i_1}+\Lambda_{i_2}\right)\sum_{k\in\mathcal{K}}Y_k \tag{1.8}$$

$$s.t.$$

$$\Lambda_i\geq\frac{P_{ik}}{Y_k}\qquad i\in\{i_1,i_2\},\qquad\forall k\in\mathcal{K}, \tag{1.9}$$

$$Y_k\geq 1\qquad\forall k\in\mathcal{K}.$$

The GIM addresses the following question: *Assuming we can allocate unlimited resources, what is the minimum balance loss we will incur if we assign forms i_1 and i_2 to the same group?* Note that (1.8) is the variable part of (1.7). Also, resources are allowed to take real values because the GIM is not concerned with resource allocation.

Let $\Lambda^*_{i_1},\Lambda^*_{i_2},Y^*_k$ be the optimal solution of (1.8). We define the penalty for placing forms i_1 and i_2 in the same group to be $R_{i_1i_2}=d_{i_1i_2}-\min_{b\neq i_1}d_{i_1b}+d_{i_2i_1}-\min_{b\neq i_2}d_{i_2b}$ in which $d_{i_1i_2}=\Lambda^*_{i_1}\sum_{k\in\mathcal{K}}Y^*_k-\sum_{k\in\mathcal{K}}P_{i_1k}$ is the proportion of idle times attributable to form i_1, $d_{i_2i_1}=\Lambda^*_{i_2}\sum_{k\in\mathcal{K}}Y^*_k-\sum_{k\in\mathcal{K}}P_{i_2k}$ is the proportion of idle times attributable to form i_2, and d_{i_1b} (d_{i_2b}) is the value of $d_{i_1i_2}$ ($d_{i_2i_1}$) when (1.8) is solved for form i_1 (i_2) and form $b\in\mathcal{I}$ such that $b\neq i_1$ ($b\neq i_2$).

Ideally, we would formulate and solve the GIM and minimize the balance loss for all combinations of forms. Because this would make the GIM and the GM exceedingly difficult, we find the balance loss for pairs of forms, instead, and use the sum of pairwise losses as a surrogate for the actual balance loss when more than two forms are assigned to the same group. To alleviate the issue of double counting, we subtract the terms $\min_{b \neq i_1} d_{i_1 b}$ and $\min_{b \neq i_2} d_{i_2 b}$.

Grouping Model (GM)

Using the indices obtained from the GIM, we formulate the GM and assign forms to groups. The lower the penalty $R_{i_1 i_2}$, the better it is to have forms i_1 and i_2 in the same group. By defining

$$X_{i_1 i_2} = \begin{cases} 1 & \text{if forms } \ i_1 \ \text{ and } \ i_2 \neq i_1 \ \text{ are assigned to the same group,} \\ 0 & \text{otherwise} \end{cases}$$

and the total processing time of form i as $P_i = \sum_{k \in \mathcal{K}} P_{ik}$, we formulate the GM with decision variables $X_{i_1 i_2}$ and Z_{ig} as

$$\text{(GM)} \min \sum_{i_1=1}^{I-1} \sum_{i_2=i_1+1}^{I} R_{i_1 i_2} X_{i_1 i_2} \tag{1.10}$$

$s.t.$

$$X_{i_1 i_2} \geq Z_{i_1 g} + Z_{i_2 g} - 1 \qquad \forall g \in \mathcal{G}, \ \forall i_1, i_2 \in \mathcal{I}, \ i_1 \neq i_2, \tag{1.11}$$

$$\sum_{g \in \mathcal{G}} Z_{ig} = 1 \qquad \forall i \in \mathcal{I}, \tag{1.12}$$

$$\sum_{i \in \mathcal{I}} P_i Z_{ig} \leq Q \qquad \forall g \in \mathcal{G}, \tag{1.13}$$

$$X_{i_1 i_2} \in \{0,1\} \qquad \forall i_1, i_2 \in \mathcal{I}, \;\; i_1 \neq i_2, \tag{1.14}$$

$$Z_{ig} \in \{0,1\} \qquad \forall i \in \mathcal{I}, \;\; \forall g \in \mathcal{G}, \tag{1.15}$$

in which $Q = (1+\delta) \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} P_{ik}/G$.

Constraint (1.13) requires that the total processing time of forms in each group not exceed the average total processing time per group by more than a predefined fraction δ, which is a parameter that controls the feasibility of the GM. If δ is too low, the GM may be infeasible. If δ is too large, then (1.13) becomes a redundant constraint. Although the value of δ need not be unique, our numerical investigations indicate that 0.25 is reasonable for three to five groups. However, when the number of groups increases, δ should be increased to ensure feasibility. We can now attest to the complexity of the GM with Proposition 3.

Proposition 1.3. *The GM is strongly NP-complete.*

Given that the GM is a hard problem, we use a decomposition procedure described in Section 1.5 to heuristically solve it.

Resource Allocation Models (RAM)

For the last component of the hierarchical framework, we formulate RAM_1 and RAM_2 to find upper and lower values for staffing levels that allow each group to finish processing their assigned forms on time at minimum expense. The formulation of RAM_1 and RAM_2 is the same for all groups. Therefore, we suppress g in decision variables Y_{kg} and T_{ig}, and, with a little abuse of notation, use $\mathcal{I}$ in this section for the set of forms that is assigned to the same group. RAM$_1$ is formulated as

$$
\begin{aligned}
(\text{RAM}_1) \quad \min \quad & \sum_{k \in \mathcal{K}} w_k Y_k \\
s.t. \quad & \\
& T_i \geq \frac{P_{ik}}{Y_k} && \forall k \in \mathcal{K}, \quad \forall i \in \mathcal{I}, && (1.16) \\
& \sum_{i \in \mathcal{I}} T_i \leq D, && && (1.17) \\
& Y_k \geq 1 \text{ and integer}, && \forall k \in \mathcal{K}.
\end{aligned}
$$

We obtain RAM$_2$ from RAM$_1$ by replacing (1.16) and (1.17) with $\sum_{i \in \mathcal{I}} P_{ik}/Y_k \leq D$ for all stages k. The following proposition states that RAM$_1$ is a hard problem.

Proposition 1.4. *RAM$_1$ is binary NP-hard.*

In Section 1.5, we propose a pseudo-polynomial time algorithm to solve RAM$_1$, and show that, unlike RAM$_1$, RAM$_2$ is an easy problem.

1.4.3 Process Line Separation Model (PLSM)

The hierarchical models we have described thus far assume that TSDC will acquire additional resources if necessary. Since hiring new employees requires time to interview candidates and train new hires, TSDC managers were also concerned with managing their existing resources. More specifically, they were interested in a model that would distribute the existing resources to two major process lines: one line for processing all federal forms and one line for processing all state forms. The Process Line Separation Model (PLSM) addresses TSDC's concern.

Let M_k be the number of existing resources at stage k. Also, let $\mathcal{I}_{\mathcal{S}}$ and $\mathcal{I}_{\mathcal{F}}$ denote the set of federal and state forms indexed by i_s and i_f. We use Y_{k_s} and Y_{k_f} to denote the number of resources allocated to stage k to process state and federal forms. With decision variables Y_{k_s}, Y_{k_f}, T_{i_s}, and T_{i_f}, the PLSM is formulated as

$$\text{(PLSM)} \quad \min \ \max \left\{ \sum_{i_s \in \mathcal{I}_{\mathcal{S}}} T_{i_s}, \sum_{i_f \in \mathcal{I}_{\mathcal{F}}} T_{i_f} \right\} \tag{1.18}$$

$$s.t.$$

$$T_{i_s} \geq \frac{P_{i_s k}}{Y_{k_s}} \qquad \forall k \in \mathcal{K}, \quad \forall i_s \in \mathcal{I}_{\mathcal{S}}, \tag{1.19}$$

$$T_{i_f} \geq \frac{P_{i_f k}}{Y_{k_f}} \qquad \forall k \in \mathcal{K}, \quad \forall i_f \in \mathcal{I}_{\mathcal{F}}, \tag{1.20}$$

$$Y_{k_s} + Y_{k_f} \leq M_k \qquad \forall k \in \mathcal{K}, \tag{1.21}$$

$$Y_{k_s}, Y_{k_f} \geq 1 \text{ and integer} \qquad \forall k \in \mathcal{K}.$$

Objective function (1.18) minimizes the approximate process completion time. Constraints (1.19) and (1.20) define the maximum effective processing time of state and federal forms. Constraint (1.21) represents resource availability.

Proposition 1.5. *PLSM is binary NP-hard.*

1.5 Solution Procedures

In this section, we describe the solution procedures for the hierarchical models and the PLSM.

1.5.1 GIM Solution

The GIM can be solved by defining $\beta = \frac{\Lambda_{i_1}}{\Lambda_{i_1}+\Lambda_{i_2}}$ and finding Y_k from (1.9) as follows:

$$\begin{aligned} \min \ \ (\Lambda_{i_1} + \Lambda_{i_2}) \sum_{k \in \mathcal{K}} Y_k &\equiv \min\nolimits_{\Lambda_{i_1}, \Lambda_{i_2} > 0} (\Lambda_{i_1} + \Lambda_{i_2}) \sum_{k \in \mathcal{K}} \max\left(P_{i_1 k}/\Lambda_{i_1}, P_{i_2 k}/\Lambda_{i_2}\right) \\ &\equiv \min\nolimits_{0<\beta<1} \sum_{k \in \mathcal{K}} \max \ \left(P_{i_1 k}/\beta, P_{i_2 k}/(1-\beta)\right). \end{aligned}$$

The function $\max\left(P_{i_1 k}/\beta, P_{i_2 k}/(1-\beta)\right)$ is strictly convex in $\beta \in (0,1)$. Because the objective function is the sum of strictly convex functions, we easily find the optimal solution.

1.5.2 GM Solution

To solve the GM, we propose a decomposition procedure that provides a lower bound on the objective function and offers feasible solutions. If we relax (1.11) by positive Lagrangian multipliers $\gamma_{i_1 i_2 g}$, we get

$$\begin{aligned} L(\gamma) = \ \min \ & \sum_{i_1=1}^{I-1} \sum_{i_2=i_1+1}^{I} \left(R_{i_1 i_2} - \sum_{g \in \mathcal{G}} \gamma_{i_1 i_2 g} \right) X_{i_1 i_2} \\ & + \sum_{g \in \mathcal{G}} \sum_{i_1=1}^{I} \left(\sum_{i_2=1}^{i_1-1} \gamma_{i_2 i_1 g} + \sum_{i_2=i_1+1}^{I} \gamma_{i_1 i_2 g} \right) Z_{i_1 g} - \sum_{g \in \mathcal{G}} \sum_{i_1=1}^{I-1} \sum_{i_2=i_1+1}^{I} \gamma_{i_1 i_2 g} \\ & s.t. \\ & \qquad (1.12), (1.13), (1.14), (1.15). \end{aligned}$$

For a vector of Lagrangian multipliers, γ, with $\gamma \geq 0$, the math program that defines the function $L(\gamma)$ can be decomposed into two math programs:

$$L_1(\gamma) = \min \sum_{i_1=1}^{I-1} \sum_{i_2=i_1+1}^{I} \left(R_{i_1 i_2} - \sum_{g \in \mathcal{G}} \gamma_{i_1 i_2 g} \right) X_{i_1 i_2} \qquad L_2(\gamma) = \min \sum_{g \in \mathcal{G}} \sum_{i_1=1}^{I} \theta_{i_1 g} Z_{i_1 g}$$

$$s.t. \qquad\qquad s.t.$$

$$(1.14), \qquad\qquad (1.12), (1.13), (1.15),$$

in which $\theta_{i_1 g} = \sum_{i_2=1}^{i_1-1} \gamma_{i_2 i_1 g} + \sum_{i_2=i_1+1}^{I} \gamma_{i_1 i_2 g}$.

In the math program that defines the function $L_1(\gamma)$, the optimal value of each $X_{i_1 i_2}$ is equal to 0 if its coefficient in the objective function is positive; otherwise it is equal to 1. The math program that defines the function $L_2(\gamma)$ is a packing-by-cost variation of the bin packing problem in which a set of items should be packed in bins (groups) of the same capacity to minimize the total packing cost. We solve the bin packing subproblem using a dynamic program in which $V_i(U_1, U_2, ..., U_G)$ denotes the minimum cost of assigning form i to one of the groups with sufficient capacity, given that the remaining capacity of bin (group) g is U_g and forms 1 to $i-1$ are already assigned. The value functions are calculated as follows:

$$V_i(U_1, U_2, ..., U_G) = \min_{\{g \in \mathcal{G} | U_g \geq P_i\}} \{\theta_{ig} + V_{i+1}(U_1, U_2, ..., U_g - P_i, ..., U_G)\} \quad \forall i \in \mathcal{I}.$$

To ensure feasibility, we set $V_i(U_1, U_2, ..., U_G) = +\infty$, if $U_g < P_i$ for all $g \in \mathcal{G}$. The optimal solution to $L_2(\gamma)$ is $V_1(Q, Q, ..., Q)$ and the complexity of the procedure is $O\left(IQ^G\right)$. This complexity is practically manageable because the number of groups does not exceed five.

Since $L(\gamma)$ is a lower bound on (1.10), we solve the following math program, using a subgradient optimization algorithm (Fisher 1981, 1985) to find the best of such lower bounds:

$$\begin{aligned} \max \quad & L(\gamma) \\ s.t. \quad & \\ & \gamma_{i_1 i_2 g} \geq 0, \quad \forall i_1, i_2 \in \mathcal{I},\ \forall g \in \mathcal{G}. \end{aligned}$$

In addition to obtaining a lower bound on the objective function, we also use the decomposition procedure to find feasible solutions to the GM. In each iteration of the subgradient optimization algorithm, we can construct a feasible solution to the GM by using the optimal values of $Z_{i_1 g}$ and $Z_{i_2 g}$ and setting $X_{i_1 i_2} = \max_{g \in \mathcal{G}}\{Z_{i_1 g} + Z_{i_2 g} - 1\}$. These feasible grouping scenarios become inputs to the RAMs for obtaining the optimal resource allocation.

1.5.3 RAM$_1$ and RAM$_2$ Solutions

We first explain how to solve RAM$_2$ because it is an easy problem. The optimal number of resources at stage k in RAM$_2$ is $Y_k^* = \left\lceil \sum_{i \in \mathcal{I}} P_{ik}/D \right\rceil$, for which $\lceil x \rceil$ is the smallest integer larger than or equal to x.

RAM$_1$ can be transformed to a shortest path problem. First, we find a lower value Y_k^{min} and an upper value Y_k^{max} for the optimal solution, $Y_k^*, k \in \mathcal{K}$, to limit the size of the network. The lower value for Y_k^* can be found by replacing (1.16) with $\sum_{i \in \mathcal{I}} T_i \geq \sum_{i \in \mathcal{I}} P_{ik} \big/ Y_k$ and enforcing (1.17). Thus, $Y_k^* \geq Y_k^{min} = \left\lceil \sum_{i \in \mathcal{I}} P_{ik} \big/ D \right\rceil$ for all stages $k \in \mathcal{K}$. For the upper value, set $P_{ik} = \max_{k \in \mathcal{K}}\{P_{ik}\}$ for all stages $k \in \mathcal{K}$ to inflate the processing times of form $i \in \mathcal{I}$ to its maximum processing time across all stages. Then, if $Y_k = \overline{Y} = \sum_{i \in \mathcal{I}} \max_{k \in \mathcal{K}}\{P_{ik}\} \big/ D$ for all stages $k \in \mathcal{K}$, the deadline will be met. Therefore, $\sum_{k \in \mathcal{K}} w_k \overline{Y} \geq \sum_{k \in \mathcal{K}} w_k Y_k^*$, which means $Y_k^* \leq Y_k^{max} = \left\lceil \left(\overline{Y} \sum_{k \in \mathcal{K}} w_k - \sum_{r \in \mathcal{K}, r \neq k} w_r Y_r^{min} \right) \big/ w_k \right\rceil$ for all stages $k \in \mathcal{K}$.

The network for RAM$_1$ has K layers and each node is represented by (k, Y_k, E) for which $E = \sum_{i\in\mathcal{I}} T_i - D$ given resources $Y_1, \ldots, Y_K$. The network is constructed using the following steps.

Step 0. Generate start (0) and finish (F) nodes.

Step 1. Generate nodes $(1, Y_1, E)$ starting from $Y_1 = Y_1^{min}$ and increasing by 1 while setting $Y_r = Y_r^{min}$ for stages $r > 1$. The cost of arc $(0) \to (1, Y_1, E)$ is $(Y_1 - Y_1^{min})\,\omega_1$. Stop adding new nodes if $Y_1 = Y_1^{max}$ or $E \leq 0$. If $E \leq 0$ occurs first, connect the last node to node F with cost 0.

Step 2. In layers $k = 2, \ldots, K$, generate the children of nodes in layer $k - 1$ with $E > 0$ in the same manner as Step 1. To be more specific, the value of E in node (k, Y_k, E) is computed by setting $Y_r = Y_r^{min}$ for stages $r > k$ and setting Y_r for stages $r \leq k$ equal to their values on the path $(1, Y_1, E) \to (2, Y_2, E) \to \cdots \to (k-1, Y_{k-1}, E) \to (k, Y_k, E)$. The cost of arc $(k-1, Y_{k-1}, E) \to (k, Y_k, E)$ is $(Y_k - Y_k^{min})\,\omega_k$. In layer K, the cost of $(K, Y_K, E) \to F$ will be a very large number if $E > 0$ and 0 otherwise.

Step 3. Compute the shortest path of the network. The optimal solution to RAM$_1$ is the sum of the cost of the shortest path and the cost of allocating $Y_k^{\min}$ to each stage. The optimal resource configuration can be determined by traversing the shortest path.

The complexity of the size of the network is $O\left((\max_{k\in\mathcal{K}}\{Y_k^{max} - Y_k^{min}\})^K\right)$. Section 1.11 provides an example of the network construction.

1.5.4 PLSM Solution

To solve the PLSM, we propose a heuristic solution that relaxes the integrality constraint on Y_{kg}, after which we establish the worst-case performance of the heuristic.

Let $\widetilde{Y}_{k_s}$ and $\widetilde{Y}_{k_f}$ be the solution to the PLSM when the integrality constraints are relaxed. We find an integer solution using the following algorithm.

Step 1. For all stages $k \in \mathcal{K}$, set $Y_{k_s} = \lfloor \widetilde{Y}_{k_s} \rfloor$ and $Y_{k_f} = \lfloor \widetilde{Y}_{k_f} \rfloor$, for which $\lfloor x \rfloor$ is the greatest integer less than or equal to x, and define $\overline{M}_k = \max\big(0, M_k - Y_{k_s} - Y_{k_f}\big)$ as the number of remaining available resources in stage $k \in \mathcal{K}$.

Step 2. Find the first stage k with $\overline{M}_k > 0$. First compute TS_k, the objective function of the PLSM for $\big(Y_{k_s} + 1, Y_{k_f}\big)$. Then compute TF_k, the objective function of the PLSM for $\big(Y_{k_s}, Y_{k_f} + 1\big)$. If $TS_k < TF_k$, let $Y_{k_s} := Y_{k_s} + 1$; otherwise, let $Y_{k_f} := Y_{k_f} + 1$. Update $\overline{M}_k$ and repeat this step until $\overline{M}_k = 0$. Finally, find the next stage with $\overline{M}_k > 0$ and repeat this step.

Let ϕ^H be the objective function value of the above heuristic. Proposition 6 places a bound on the worst-case performance of the heuristic.

Proposition 1.6. *If ϕ^* denotes the optimal value of the PLSM, then* $\dfrac{\phi^H}{\phi^*} \leq 2$.

1.6 Numerical Experiments on Performance

In this section, we report the results of our numerical experiments for evaluating the proposed hierarchical procedure. We chose the parameters for the experiments based our observations at TSDC. We found that historically about 50% of the forms are easy to process, 20% are difficult to process, and the remaining 30% are moderately difficult to process. Tables 1.1 and 1.2 shows the processing time intervals (in hours) for these three categories, with the last row displaying the scaled resource costs in thousands of dollars.

For our experiments, we calculated the number of working hours in the time interval between the start of the project on August 1^{st} and the finish on December

Table 1.1: Processing time intervals for form categories and resource costs at process stages

	Processing Time Intervals (hours)				
Form Categories	IDG Process	CALC Process	EF Process	Interview Process	Integration
Easy	[2 5]	[3 6]	[2 5]	[1 3]	[0.5 1.5]
Fairly Hard	[5 12]	[8 12]	[4 7]	[2 5]	[1 1.5]
Hard	[12 20]	[12 24]	[6 9]	[4 7]	[1.5 2]
Cost ($1000)	50	60	40	35	50

Table 1.2: Processing time intervals for form categories and resource costs at test stages

	Processing Time Intervals (hours)				
Form Categories	IDG Test	CALC Test	EF Test	Interview Test	Final Test
Easy	[1 2]	[1 2]	[0.5 1]	[0.5 1]	[1 2]
Fairly Hard	[1.5 2.5]	[2 3]	[0.5 1.5]	[0.5 1.5]	[2 3]
Hard	[2 3]	[3 4]	[1 1.5]	[1 1.5]	[3 4]
Cost ($1000)	10	15	10	10	15

15^{th}. In every experiment, we considered different combinations of group G and stage K, and we generated 90 instances of the models for each combination. We evaluated the quality of the proposed hierarchical approach by means of two major comparisons. First, we compared the solution of the hierarchical models to the solution of the monolithic models, for small to moderate instances that we could solve optimally. Second, we compared the solution of the hierarchical models to the solution obtained from simulation–optimization.

1.6.1 GM Lagrangian Heuristic Performance

In this section, we evaluate the contribution of the Lagrangian decomposition heuristic to solving the Grouping Model (GM). Because the solution to the GM depends on the value of δ in (1.13), we also vary δ. We report Ave(LH/LB), which is the average ratio of the grouping penalty of the Lagrangian heuristic solution to the lower bound on the optimal grouping penalty. Note that each iteration of the subgradient method provides a heuristic solution to the GM. We use the best solution over the 35 iterations of the subgradient method.

Table 1.3 shows the results of the experiments. One can see that the average of Ave(LH/LB) is lowest when $\delta = 0.25$, which supports our statement that for 3 to 5 groups, the value of δ should be around 0.25. The values of Ave(LH/LB) for $\delta = 0.25$ show that the difference between the heuristic solution and the lower bound is on average 2% and has a 95% confidence interval of [1.8% 2.2%]. Note that this is a conservative estimate of the quality of the Lagrangian heuristic since we do not have the optimal solution to the GM. Therefore, the performance of the solution procedure for the GM is quite good.

1.6.2 Hierarchical vs. Monolithic

Table 1.4 displays the average ratio of the total workforce cost of the hierarchical models to that of the monolithic models, Ave(HA/MO), calculated from 810 problem instances. The bottom three rows of the table provide the grand averages and 95% confidence intervals (CI) for Ave(HA/MO). The grand averages indicate that the total workforce cost of the hierarchical models are on average 2.07% greater than the total workforce cost of the monolithic models. We also conducted an experiment to compare the performance of the hierarchical models to that of the monolithic models

Table 1.3: Performance of the Lagrangian heuristic for the GM

G	K	I	Ave(LH/LB)						
			$\delta = 0.10$	$\delta = 0.15$	$\delta = 0.20$	$\delta = 0.25$	$\delta = 0.30$	$\delta = 0.35$	$\delta = 0.40$
3	8	500	1.061	1.044	1.006	1.044	1.041	1.007	1.045
	8	750	1.020	1.032	1.031	1.009	1.015	1.028	1.027
	8	1000	1.025	1.002	1.030	1.002	1.032	1.007	1.032
	9	500	1.066	1.038	1.039	1.009	1.046	1.052	1.003
	9	750	1.027	1.020	1.009	1.043	1.053	1.008	1.033
	9	1000	1.064	1.015	1.022	1.003	1.032	1.006	1.038
	10	500	1.064	1.014	1.026	1.011	1.034	1.007	1.007
	10	750	1.005	1.057	1.026	1.017	1.030	1.022	1.041
	10	1000	1.005	1.038	1.019	1.020	1.026	1.052	1.050
4	8	500	1.043	1.021	1.048	1.018	1.036	1.050	1.041
	8	750	1.011	1.026	1.020	1.024	1.020	1.004	1.044
	8	1000	1.027	1.005	1.025	1.017	1.015	1.056	1.054
	9	500	1.069	1.034	1.020	1.018	1.013	1.017	1.056
	9	750	1.017	1.006	1.011	1.035	1.017	1.028	1.052
	9	1000	1.012	1.014	1.035	1.027	1.029	1.032	1.064
	10	500	1.049	1.022	1.013	1.021	1.048	1.018	1.032
	10	750	1.050	1.002	1.033	1.042	1.003	1.035	1.009
	10	1000	1.052	1.059	1.045	1.032	1.010	1.058	1.009
5	8	500	1.021	1.040	1.033	1.002	1.015	1.026	1.036
	8	750	1.005	1.015	1.034	1.028	1.019	1.023	1.017
	8	1000	1.019	1.002	1.038	1.008	1.000	1.011	1.007
	9	500	1.024	1.008	1.001	1.010	1.003	1.055	1.055
	9	750	1.009	1.000	1.037	1.018	1.027	1.015	1.061
	9	1000	1.075	1.009	1.011	1.007	1.003	1.003	1.030
	10	500	1.054	1.020	1.020	1.020	1.033	1.027	1.063
	10	750	1.062	1.013	1.028	1.020	1.036	1.032	1.008
	10	1000	1.008	1.023	1.030	1.041	1.025	1.044	1.052
95% CI (lower)			1.032	1.019	1.024	1.018	1.023	1.024	1.033
Average			1.035	1.021	1.026	1.020	1.025	1.027	1.036
95% CI (higher)			1.038	1.024	1.027	1.022	1.026	1.029	1.038

when the problem size grows. Figure 1.4 shows Ave(HA/MO) for 10 instances of the hierarchical models when $G = 5$, $K = 10$, and I increases to 500. The results indicate that the quality of the hierarchical solution does not deteriorate when we increase the size of the models. Therefore, the hierarchical models provide a good heuristic solution to the complex monolithic models.

Table 1.4: The hierarchical models vs. the monolithic models

G	K	Ave(HA/MO)					
		GRAM$_1$/MGRAM$_1$			GRAM$_2$/MGRAM$_2$		
		$I = 50$	$I = 75$	$I = 100$	$I = 50$	$I = 75$	$I = 100$
3	8	1.022	1.014	1.023	1.011	1.024	1.035
	9	1.011	1.021	1.032	1.016	1.031	1.046
	10	1.024	1.009	1.019	1.034	1.030	1.016
4	8	1.007	1.027	1.024	1.013	1.020	1.005
	9	1.005	1.010	1.012	1.024	1.024	1.040
	10	1.021	1.026	1.024	1.013	1.013	1.015
5	8	1.024	1.028	1.038	1.012	1.022	1.025
	9	1.021	1.027	1.027	1.009	1.008	1.025
	10	1.028	1.016	1.016	1.015	1.021	1.008
CI (lower)		1.017	1.019	1.022	1.015	1.020	1.022
Average		1.018	1.020	1.024	1.016	1.022	1.024
CI (upper)		1.019	1.021	1.025	1.017	1.023	1.025

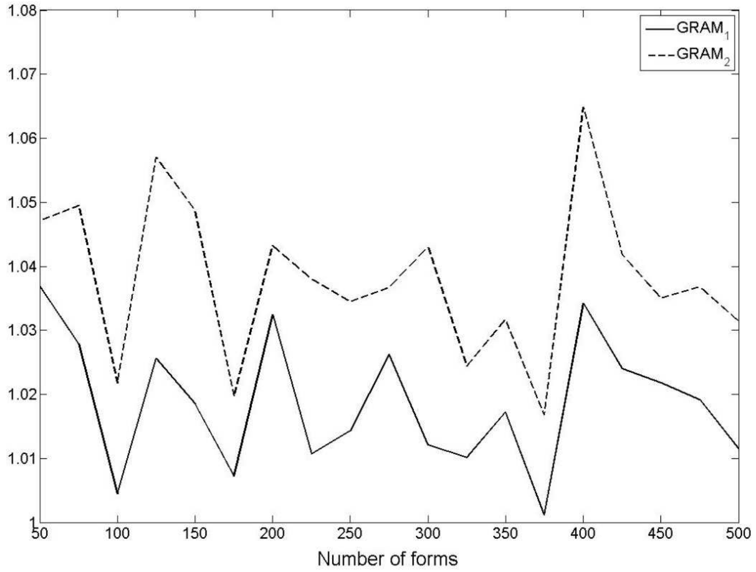


Figure 1.4: Ratio of the solution of the hierarchical models to the solution of the monolithic models

1.6.3 Hierarchical vs. Simulation-Optimization

Next, we compared the result of our hierarchical approach to the simulation–optimization approach in a fashion similar to that used in Kekre et al. (2009). We used Arena simulation software to capture the operational particulars of the system and to compute the exact completion time of the process. The arrival times of the forms, processing times, and probability of feedback are all random numbers. Because our hierarchical models approximate the completion time, arrival times, and feedback loops, the simulation model scheduled overtime when was needed to meet the deadline. The overtime cost was assumed to be 50% higher than the cost of regular time.

For simulation–optimization, we used Arena's OptQuest in an iterative manner. OptQuest generated a vector of resource configuration and assignment of forms to groups, and Arena evaluated the project completion time and the total workforce cost. We repeated this process and used the default options of OptQuest to stop simulation–optimization.

Table 1.5 reports the average ratio of the total workforce cost of the hierarchical models to the total workforce cost obtained from simulation–optimization for small instances, Ave (HA/SO). We should note that the ratios for GRAM_2 are higher than the ratios for GRAM_1, because GRAM_2 allocates fewer resources to each stage than GRAM_1 and consequently results in more overtime. Although simulation–optimization achieves a lower overall cost than GRAM_1, the total workforce cost of GRAM_1 is on average only 2.57% greater than the total workforce cost obtained from simulation–optimization.

Although simulation–optimization appears to provide better solutions for smaller instances, it is unfortunately computationally very intensive. We were unable to use simulation–optimization due to the large number of binary variables that we

Table 1.5: The hierarchical models vs. simulation–optimization

G	K	Ave(HA/SO)					
		GRAM$_1$/SO			GRAM$_2$/SO		
		$I = 50$	$I = 75$	$I = 100$	$I = 50$	$I = 75$	$I = 100$
3	8	1.021	1.021	1.037	1.103	1.064	1.105
	9	1.029	1.023	1.033	1.066	1.065	1.062
	10	1.009	1.016	1.026	1.027	1.022	1.058
4	8	1.042	1.010	1.033	1.053	1.092	1.065
	9	1.032	1.033	1.029	1.062	1.086	1.065
	10	1.043	1.022	1.024	1.042	1.033	1.063
5	8	1.024	1.025	1.025	1.042	1.102	1.049
	9	1.016	1.033	1.018	1.045	1.030	1.034
	10	1.014	1.036	1.021	1.076	1.041	1.107
CI (lower)		1.024	1.022	1.025	1.053	1.055	1.063
Average		1.026	1.024	1.027	1.057	1.059	1.068
CI (upper)		1.027	1.025	1.028	1.060	1.062	1.070

encountered in our problem. Figure 1.5 shows the ratio of simulation–optimization runtime to the runtime of the hierarchical models for different number of forms, $G = 3$, and $K = 8$. The ratio of the runtime grows almost exponentially with the number of forms. Furthermore, the number of decision variables exceeded the current capabilities of commercial software such as OptQuest and prohibited us from using this methodology in practice.

We also evaluated the computation time of the hierarchical models. Table 1.6 report the average time (in seconds) it takes to solve 90 instances of the hierarchical models for each combination of G, K, and I. The average runtime is 84.5 seconds for GRAM$_1$ and 47.8 seconds for GRAM$_2$. These computation times are quite satisfactory for tactical planning in practice.

Table 1.6: Computation time of hierarchical models

G	K	I	Average CPU Time (seconds)	
			GRAM$_1$	GRAM$_2$
3	8	500	48.6	31.0
	8	750	66.2	31.1
	8	1000	79.7	48.0
	9	500	85.4	49.3
	9	750	87.7	53.5
	9	1000	112.8	54.5
	10	500	51.1	50.0
	10	750	58.5	61.3
	10	1000	86.2	76.4
4	8	500	73.0	33.5
	8	750	83.9	39.5
	8	1000	118.6	50.0
	9	500	77.3	41.0
	9	750	94.7	46.3
	9	1000	129.4	64.1
	10	500	65.2	45.6
	10	750	68.6	51.6
	10	1000	70.4	62.5
5	8	500	79.8	44.4
	8	750	91.2	52.3
	8	1000	106.6	55.5
	9	500	58.5	31.2
	9	750	73.0	36.8
	9	1000	100.1	41.8
	10	500	74.3	32.2
	10	750	93.3	40.3
	10	1000	110.8	49.2
CI (lower)			82.0	46.4
Average			84.5	47.8
CI (higher)			87.0	49.1

1.6.4 Effect of Approximations

In addition to the aforementioned experiment, we also used simulation–optimization to examine the effect of approximations used by the hierarchical procedure on total workforce cost. In particular, we were interested to know what percentage of the

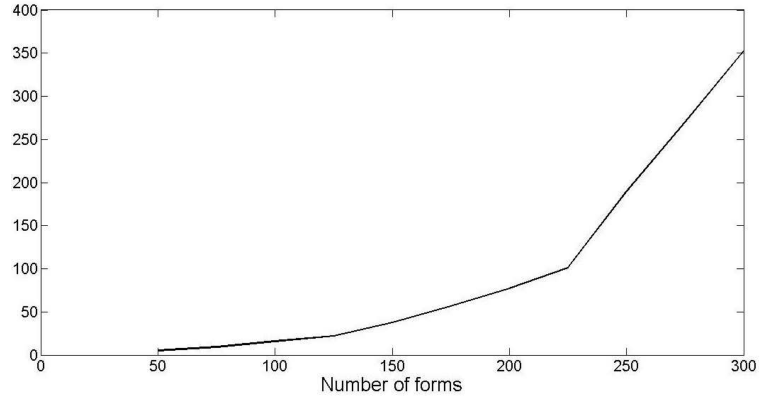


Figure 1.5: Ratio of simulation–optimization runtime to hierarchical approach runtime

error in the total workforce cost of GRAM_1 was due to approximating feedback loops and what percentage was due to approximating the process completion time with $\max_{g\in\mathcal{G}}\{\sum_{i\in\mathcal{I}} T_{ig}Z_{ig}\}$. For this purpose, we conducted two variations of simulation–optimization for the GRAM_1 instances in Table 1.5 and calculated the ratio, λ, for each variation. The denominator of λ in both variations was the average difference between the total workforce cost of GRAM_1 and the total workforce cost of the simulation–optimization experiment in Section 1.6.3.

In the first variation of our simulation–optimization, we eliminated the feedback loops and applied the approximation described in Section 1.3.2 to the process. We then used the difference between the total workforce cost of the new simulation–optimization experiment and the total workforce cost of GRAM_1 as the numerator of λ, and found that 37.6% ± 2.3% (95% confidence interval) of the average difference between the total workforce cost of GRAM_1 and the simulation–optimization in Section 1.6.3 was due to the feedback loop approximation.

In the second variation, we applied the approximation for feedback loops and approximated dynamic arrivals with the availability of all forms when the process started. We used the difference between the total workforce cost of this simulation–

optimization experiment and the total workforce cost of GRAM_1 as the numerator of λ. The resulting ratio indicated that $52.3\% \pm 2.9\%$ (95% confidence interval) of the average difference between the total workforce costs in Section 1.6.3 was due to approximating the process completion time in GRAM_1.

In summary, we conclude that GRAM_1 generates very good solutions for organizing the complex development process, and that the solutions of GRAM_1 and GRAM_2 provide TSDC managers with a range of resource configurations to finish the processing of forms in all groups by the deadline.

1.7 Implementation

Prior efforts at process improvement had rendered TSDC receptive to our analytical approach. Upon examining our analysis and considering the potential benefits to TSDC, the Vice President of Operations directed the groups to implement our models during the 2010 development period. Process managers cooperated fully and provided us with the necessary data, and the vice president removed internal obstacles throughout the project, making her trust and commitment essential to testing and implement our solutions in a real-world situation.

1.7.1 Simulations

Prior to putting our hierarchical models into practice at TSDC, we evaluated the quality of the solutions they evoked by collecting historical data and then developing a simulation model. To estimate the expected number of forms, we fit arrival data from the past several decades to a linear regression model. To generate arrival patterns, we used the average cumulative arrivals of forms over the past three years (2007-

2009). Then, we randomly assigned arrival dates to them such that their cumulative arrivals matched the historical percentage of arrivals. To estimate processing time distributions at each stage, we designed a questionnaire (see Section 1.12) asking employees and managers to estimate the following:

- Minimum, average, and maximum processing times needed to complete the forms at each stage;

- Percentage of forms that take less than 25%, between 25% and 50%, between 50% and 75%, and above 75% of the maximum processing time; and

- Percentage of forms that fail each internal test and Final Test and return for rework.

Subsequently, we developed a simulation model that accounted for uncertainty in the arrival and processing times of forms, feedback loops for addressing errors (bug fixes), and TSDC rules for sequencing forms and scheduling overtime policies. At the time of our study, TSDC employed two sequencing rules: (1) among available forms, the form with the shortest total processing time, $\min_{i\in\mathcal{I}}\left\{\sum_{k\in\mathcal{K}} P_{ik}\right\}$, was assigned to a resource first; and (2) at each stage, forms were assigned to the resource with the least amount of work to do. TSDC overtime policy required managers to update their estimates of total remaining work at the end of each week. They determined the amount of overtime for the upcoming weeks by calculating the difference between new and old estimates and dividing it by the number of weeks remaining until the deadline.

To validate the simulation model, we applied a procedure similar to that used in Kekre et al. (2009). We ran the simulation model using historical data on staffing levels for a two-year period, and the model provided values for the total amount of work remaining at the end of each week. Next, we compared these weekly values with the actual total amount of remaining work at the end of each week provided by management. We constructed a 95% paired$-t$ confidence interval $\overline{\chi} \pm t_{(N-1,0.975)}\sqrt{S_{\chi}^{2}/N}$ in which $\overline{\chi}$ and S_{χ} denote the average and sample standard deviation of the difference between actual and simulated amount of work remaining, and $t_{(N-1,0.975)}$ denotes the 0.975 critical value for the t distribution with $N-1=42$ degrees of freedom. Because the confidence interval was $[-0.012, 0.021]$ and included 0, we could not reject the null hypothesis that the average of the total work to do at the end of each week obtained from the simulation model was the same as the actual average. Thus we concluded that our simulation model was a good representation of the actual process.

We then conducted the runs test (Black 2011) at the 5% significance level to discover whether or not errors in the simulated total amount of work to do at the end of each week were random. We arrived at a test statistic value of -1.31, which falls between -1.96 and 1.96. Therefore, we could not reject the null hypothesis that the errors were random.

1.7.2 Two-Phase Implementation

TSDC managers requested that we implement the models in two phases. The goal of phase 1 was to assess the potential benefits of adopting the models' recommendations without altering the workforce level. The goal of phase 2 was to implement these recommendations by hiring and relocating employees during the 2010 production season.

Phase 1 entailed analyzing the benefits of optimally dividing the existing workforce into two designated groups: one for all state forms and one for all federal forms. We were given 2009 TSDC data for the number of resources at each stage, M_k, and the processing times of each form, P_{ik}. Because P_{ik} already included all processing and reprocessing times, we did not apply our approximation procedure for feedback loops in this phase. We used the PLSM to allocate resources to these two predefined groups. Next, we incorporated the actual arrival pattern of forms in 2009 and the solution from the PLSM into the simulation model and evaluated the total time needed to process the forms in each of the two groups. We compared the simulated process completion time to the actual completion time in 2009 and found that TSDC could have reduced the completion time by 23.5%.

Phase 2 involved fully implementing our models throughout the 2010 tax season. As mentioned earlier, a regression model generated the number of forms and questionnaires collected processing time estimates. We followed the procedure in Section 1.3.2 to calculate the no-loops approximate processing times at each stage. From discussions about the relative importance of resource pooling, communication, and coordinating and controlling the workflow, the process managers elected to use two groups for processing all forms. We used GRAM$_1$ and GRAM$_2$ to generate a minimum and maximum staffing level for each group at each stage. TSDC proceeded to base hiring and relocation decisions for each stage on the average staffing level obtained from GRAM$_1$ and GRAM$_2$.

In accordance with our suggested new configuration, TSDC relocated approximately 12% of the 237 employees to new job assignments and hired 8 new employees (a 3% increase in the workforce). Classified into three major skill categories—programmers, testers, and accountants—employees can be relocated within their category but not across categories. Therefore, of the 12% who were relocated, 8.5% were relocated from other teams working on the same product and 3.5% came from a different product line. a different product line.

1.7.3 Outcomes

To measure the savings obtained by implementing the hierarchical models, TSDC compared overtime and total cost figures for the 2010 tax year to those from 2009. The company discovered that it enjoyed a 25.7% reduction in overtime and an 11.3% reduction in total workforce cost. The high cost of engaging eight new employees notwithstanding, the hiring and relocation decisions helped reduce overtime payments, which ultimately reduced total workforce cost. The savings proved even greater when we considered the total amount of work before and after release. In 2009, total overtime was 59,439 hours, and the ratio of total overtime to regular hours was 22.5%. In 2010, total overtime declined by 31.6% to 40,656 hours, and the ratio to regular hours was 15.4%. The total workforce cost in 2010 was 13.6% lower than that of 2009, roughly $960,000 in cost savings. Precise workforce costs being confidential, TSDC did confirm that the savings were not the product of other factors such as employee turnover or changing product demand, workforce skills, or structure of the software. In fact, the actual amount of work TSDC accomplished in 2010 was 1.8% greater than that in 2009 due to changes in the forms. Therefore, we can claim that the savings were the result of using our decision-making tools.

At the end of the season, we reapplied our hierarchical models and used the simulation model to understand the contribution different decisions made to the savings. We found that with two groups and hiring not allowed, the savings in total overtime and workforce cost would have been reduced to 22.8% and $573,000. We also found that with one group processing all forms and hiring allowed, the savings would have reduced to 21.9% and $687,000. It is important to note that there are a lot of intangible and non-monetary benefits in organizing the development process by creating groups. Finally, if relocating employees was not allowed, the savings would have been

18.5% and $660,000. Motivated by the savings in overtime and total resource costs, TSDC decided to implement our models every year.

In addition to saving money and completing the software on time, our models helped TSDC resolve some long-standing internal disagreements about task assignment and workforce management. Some managers were particularly amenable to our proposed solutions because they did not involve issuing pink slips. Also, establishing two processing lines promoted healthy competition between groups to complete tasks sooner with less overtime.

One challenge to implementation was estimating processing times. At the end of the season, we found that the actual total amount of work was 3.6% larger than our estimates. This error was due to estimating the number and processing times of forms and approximating the effect of feedback loops. When we reapplied our models at the end of the 2010 season using the actual processing times, we found that the total workforce cost savings could have been $1,124,000. In an effort to improve accuracy, TSDC decided to improve the system for recording processing times and rework iterations.

TSDC managers are now contemplating expanding our modeling framework to development processes for other product lines. Because employees with certain skills (e.g., programmers) can work in across products, a consolidated workforce management system could help TSDC utilize its personnel more efficiently.

1.8 Conclusion

The survival of a commercial tax preparation application depends on developers being able to accurately update thousands of individual forms under extremely tight release deadlines. Studying the software development process at one large company, we observed that dynamically-arriving form changes, variable processing times, feedback loops, and high task volume make process management a formidable undertaking. The same challenges are faced by many companies servicing highly-regulated domains that require them to routinely upgrade complex software applications.

We used an approximation to capture the effect of feedback loops on completion time. To help the company manage the development process more effectively, we then introduced a hierarchical framework that focused on assigning tax forms to groups and allocating resources to meet the release deadline. Numerical experiments supported our modeling assumptions and attested to the excellent quality of the hierarchical models and solution procedures. Implementing our models reduced overtime hours by 31% and total workforce cost by 13% or around $1 million. The software company successfully delivered the application on time, even though the amount of work performed was greater than in the previous year. We hope to study other product lines at the company and expand our models to manage more processes.

1.9 Appendix A: Proofs

Proof of Proposition 1.1. It is not difficult to see that the optimal schedule is a permutation schedule; i.e., the sequence of processing forms is the same in all stages. An upper value on the completion time is obtained by inflating the processing time of each form at each stage to its maximum processing time across stages. In this case, the form with the largest inflated processing time and all the forms after it are processed in stages $2, \cdots, K$ with no inserted idle time. Therefore, the completion time is equal to the sum of the processing times plus the transition time of the form with the largest inflated processing time in stages $2, \cdots, K$, which is $\sum_{i \in \mathcal{I}} \max_{k \in \mathcal{K}} \{P_{ik}\} +$

$(K-1)\max_{i\in\mathcal{I}}\{\max_{k\in\mathcal{K}}\{P_{ik}\}\}$. This upper value is independent of the sequence of forms and the ratio of $\sum_{i\in\mathcal{I}}\max_{k\in\mathcal{K}}\{P_{ik}\}+(K-1)\max_{i\in\mathcal{I}}\{\max_{k\in\mathcal{K}}\{P_{ik}\}\}$ to $\sum_{i\in\mathcal{I}}\max_{k\in\mathcal{K}}\{P_{ik}\}$ approaches 1 when the number of forms is sufficiently large, because K is much smaller than I. □

Proof of Proposition 2.2. We show that MGRAM$_1$ and MGRAM$_2$ are as hard as the 3-partition problem, which is known to be strongly NP-hard (Garey and Johnson 1979). Assume that we are given a general instance of the 3-partition problem consisting of an index set $A=(1,2,...,3m)$, positive elements a_i for $i=1,2,...,3m$, and a positive integer B such that $B/4<a_i<B/2$ and $\sum_{i=1}^{3m}a_i=mB$. We now introduce a specific instance of MGRAM$_1$ and MGRAM$_2$ as follows: $K=1$, $G=m$, $I=3m$, $P_{ik}=a_i$ for all i and k, $D=B$, and $w_k=1$ for all k. We shall show that the optimal solution of MGRAM$_1$ and MGRAM$_2$ takes value m if and only if the $3m$ elements of A can be partitioned into m disjoint subsets $A_1,A_2,...,A_m$ such that $\sum_{i\in A_r}a_i=B$ for $r=1,...,m$. If the 3-partition problem has a solution, then it is easy to see that the elements of each subset A_r could be assigned to one of the m groups and that the total effective processing time in each group would be equal to D. In this case, each group would have one resource assigned to it: $Y_{kg}=1$ for all k and g and $\sum_{k\in\mathcal{K}}\sum_{g\in\mathcal{G}}Y_{kg}=m$. If the 3-partition problem does not have a solution, then there is at least one subset A_r such that the total effective processing time in that group would be greater than D. To meet the deadline, D, more than one resource would have to be assigned to this group. All other subsets would require one unit of resource, making the total resources greater than m. □

Proof of Proposition 2.3. We shall show that the recognition version of the GM is as hard as the 3-partition problem, which is known to be strongly NP-complete. Assume that we are given a general instance of the 3-partition problem consisting of an index set $A = (1, 2, ..., 3m)$, positive elements a_i for $i = 1, 2, ..., 3m$, and a positive integer, B, such that $B/4 < a_i < B/2$ and $\sum_{i=1}^{3m} a_i = mB$. We now introduce a specific instance of the GM as follows: $G = m$, $I = 3m$, $P_i = a_i$ for all i, $R_{i_1 i_2} = 0$ for all i_1 and i_2, and $Q = B$. We shall show that the GM has a feasible solution if and only if the $3m$ elements of A can be partitioned into m disjoint subsets $A_1, A_2, ..., A_m$ such that $\sum_{i \in A_r} a_i = B$ for $r = 1, \ldots, m$. If the 3-partition problem has a solution, then it is easy to see that the elements of each subset A_r could be assigned to each of the m groups and that the total processing time of each group would be equal to B. If the 3-partition problem does not have a solution, then there is at least one subset A_r such that the total processing time of that group would be greater than B and it is easily seen that the GM has no feasible solution. □

Proof of Proposition 2.4. We shall show that RAM$_1$ is as hard as the equal-size, equal-number-of-items partition problem, which is known to be binary NP-hard. Assume that we are given a general instance of the equal-size, equal-number-of-items partition problem consisting of an index set $A = (1, 2, ..., m)$, in which m is even and elements a_i for $i = 1, 2, ..., m$ are positive. Consider the following instance of RAM$_1$ as follows: $K = m$ and $P_{ik} = 2a_i$ if $i = k$; otherwise $P_{ik} = a_i$. Also, $D = \frac{3}{2} \sum_{r \in A} a_r$, $w_k = 1$ and $1 \leq Y_k \leq 2$ for all k. Finally, we consider an objective value of $D = \frac{3}{2} \sum_{r \in A} a_r$. Note that in any feasible solution to RAM$_1$, there exists a subset of stages, B, such that $Y_k = 2$ for all $k \in B$ and $Y_k = 1$ for all $k \notin B$. If there exists a partition, $B \subset A$, such that $\sum_{r \in B} a_r = \sum_{r \in A-B} a_r$ and $|B| = m/2$, then setting $Y_k = 2$ for all $k \in B$ and $Y_k = 1$ for all $k \in A - B$ generates a feasible solution to RAM$_1$ with the objective value of $D = \frac{3}{2} \sum_{r \in A} a_r$. If no partition exists and there is a solution to RAM$_1$ such that its objective value is less than or equal to $D = \frac{3}{2} \sum_{r \in A} a_r$, then it is easy to see that there should exist a subset $C \subset A$ such that $|C| < |A|/2$. Setting $Y_k = 2$ for all $k \in C$ and $Y_k = 1$ for all $k \in A - C$ does not provide a feasible solution to RAM$_1$. □

Proof of Proposition 1.5. We show that PLSM is as hard as the equal-size, equal-number-of-items partition problem with set $A = (1, 2, ..., m)$ and m even. Consider the following instance of PLSM as follows: $I_S = I_F = m$, $K = m$, $m_k = 3$, and $P_{ik} = 2a_i$ if $i = k$; otherwise, $P_{ik} = a_i$. Suppose the objective value is $\frac{3}{2}\sum_{r\in A} a_r$. If there exists a partition $B \subset A$, such that $\sum_{r\in B} a_r = \sum_{r\in A-B} a_r$ and $|B| = m/2$, then in each feasible solution to PLSM, each stage in the federal or state process line will use either one or two resources. Let B be the set of stages in the state process line that use two resources and let $A - B$ be the set of stages in the federal process line that use two resources. Then, the total maximum amount of work is given by $\sum_{r\in B} a_r + 2\sum_{r\in A-B} a_r$ for the state process line and $2\sum_{r\in B} a_r + \sum_{r\in A-B} a_r$ for the federal process line. Consequently, the objective value will be $\frac{3}{2}\sum_{r\in A} a_r$. Conversely, if no partition exists, we assume by contradiction that there is a solution to PLSM with the objective value of $\frac{3}{2}\sum_{r\in A} a_r$. Then, we would have:

$$\sum_{r\in B} a_r + 2\sum_{r\in A-B} a_r \leq \frac{3}{2}\sum_{r\in A} a_r,$$

$$2\sum_{r\in B} a_r + \sum_{r\in A-B} a_r \leq \frac{3}{2}\sum_{r\in A} a_r,$$

which lead to a contradiction. □

Proof of Proposition 1.6. Let ϕ^f be the value of PLSM when $Y_{k_s} = \left\lfloor \widetilde{Y}_{k_s} \right\rfloor$, $Y_{k_f} = \left\lfloor \widetilde{Y}_{k_f} \right\rfloor$. Similarly, let ϕ^c be the value of PLSM when $Y_{k_s} = \left\lceil \widetilde{Y}_{k_s} \right\rceil$ and $Y_{k_f} = \left\lceil \widetilde{Y}_{k_f} \right\rceil$. Since the objective function is non-decreasing in the number of resources, we can write:

$$\frac{\phi^H}{\phi^*} \leq \frac{\phi^f}{\phi^c} \leq \frac{\sum\limits_{i_s \in \mathcal{I}_\mathcal{S}} \max_{k \in \mathcal{K}} \left\{ P_{i_s k} / \left\lfloor \widetilde{Y}_{k_s} \right\rfloor \right\} + \sum\limits_{i_f \in \mathcal{I}_\mathcal{F}} \max_{k \in \mathcal{K}} \left\{ P_{i_f k} / \left\lfloor \widetilde{Y}_{k_f} \right\rfloor \right\}}{\sum\limits_{i_s \in \mathcal{I}_\mathcal{S}} \max_{k \in \mathcal{K}} \left\{ P_{i_s k} / \left\lceil \widetilde{Y}_{k_s} \right\rceil \right\} + \sum\limits_{i_f \in \mathcal{I}_\mathcal{F}} \max_{k \in \mathcal{K}} \left\{ P_{i_f k} / \left\lceil \widetilde{Y}_{k_f} \right\rceil \right\}}.$$

For each $i_s \in \mathcal{I}_\mathcal{S}$, consider $\eta_{i_s} = \frac{\max_{k \in \mathcal{K}} \left\{ P_{i_s k} / \left\lfloor \widetilde{Y}_{k_s} \right\rfloor \right\}}{\max_{k \in \mathcal{K}} \left\{ P_{i_s k} / \left\lceil \widetilde{Y}_{k_s} \right\rceil \right\}}$ and let k' and k'' be the stages that determine the maximum in the numerator and denominator, respectively. If $k' = k''$, then $\eta_{i_s} \leq 2$. If $k' \neq k''$, then $\left\lceil \widetilde{Y}_{k''_s} \right\rceil \leq P_{i_s k''_s} \left\lceil \widetilde{Y}_{k'_s} \right\rceil / P_{i_s k'_s}$. Thus, $\eta_{i_s} = P_{i_s k'_s} \left\lceil \widetilde{Y}_{k''_s} \right\rceil / \left\lfloor \widetilde{Y}_{k'_s} \right\rfloor P_{i_s k''_s} \leq 2$. The inequality also holds if we define η_{i_f} in a similar manner for federal forms. Therefore, $\phi^H / \phi^* \leq 2$. □

1.10 Appendix B: A Special Case of the Feedback Loops Approximation

Consider a process with two single-resource stages and deterministic processing times P_1 and P_2. When a form visits stage 2 for the n-th time, it returns to stage 1 for reprocessing with probability α if $n \leq \zeta$, and leaves the process with certainty if $n = \zeta + 1$, in which $1 \leq \zeta < \infty$. Therefore, if the number of times a form returns to stage 1 is r, then its processing times in the no-loops system will be $(r+1)P_1$ and $(r+1)P_2$. Let C^L_{max} and C^{NL}_{max} denote the completion time in the system with loops and in the no-loops system, respectively.

Proposition 1.7. *As the number of forms increases, the completion time of the no-loops process approaches the completion time of the process with loops for any realization of feedback loops; i.e.,* $\pi = \lim\limits_{I \to \infty} \frac{C^{NL}_{max}}{C^L_{max}} = 1$.

Proof of Proposition 1.7. Assume that the forms are indexed according to the order of process; i.e., the first form is form 1, the second form is form 2, and so on. We first consider the case of $P_1 \geq P_2$. Let t^L denote the time when resource 1 finishes processing the last form that leaves the process with loops behind empty. Then $C^L_{max} = t^L + P_2$. Denote with t^{NL} the time when resource 1 completes processing form I in the no-loops system. Then $t^{NL} \leq t^L \leq t^{NL} + \zeta P_2$. The inequalities hold because in the system with loops, the return of forms may generate idles times for resource 2, and the total idle time is at most ζp_2. In the no-loops process, either form I does not wait for resource 2, or it must wait for resource 2 behind forms with inflated processing times. In both cases, we can write $C^{NL}_{max} \leq t^{NL} + (1 + 2\zeta) P_2$. Putting all the inequalities together and noting that $C^{NL}_{max} \geq t^{NL}$, we have $C^L_{max} - (1 + \zeta)P_2 \leq C^{NL}_{max} \leq C^L_{max} + 2\zeta P_2$, which means $\pi = 1$.

Now, we consider the case of $P_2 > P_1$. Let φ_1 and φ_2 be the number of times forms 1 and 2 return. For forms 2,...,I, let j be the total number of forms that return and $n_1, ..., n_\zeta$ be the number of forms that return 1,...,ζ times. Clearly $\sum_{\nu=1}^{\zeta} n_\nu = j$. Also, define $\mathcal{A} = P_1 + \sum_{\nu=1}^{\zeta} n_\nu (\nu + 1) P_2 + (I - j) P_2$. In both systems, the earliest time to complete processing all forms is when resource 2 works continuously. Therefore, $C^L_{max} \geq \mathcal{A}$ and $C^{NL}_{max} \geq \mathcal{A} + \varphi_1 P_1$. Since the return of a form may generate idle times at stage 1 and forms return at most ζ times, we have $C^L_{max} \leq \mathcal{A} + \zeta P_1$. Now suppose in the no-loops system some forms generate idle times for resource 2 between processing consecutive forms. The largest idle time is generated when form 2 makes additional $\zeta - \varphi_2$ returns, because processing form 2 at stage 1 starts when stage 2 starts processing form 1. Therefore, $C^{NL}_{max} \leq \mathcal{A} + (1 + \varphi_1 + \zeta) P_1 + (\zeta - \varphi_2) P_2$. Because C^L_{max} and C^{NL}_{max} are bounded by linear functions of $\mathcal{A}$, and $\mathcal{A} \longrightarrow \infty$ when $I \longrightarrow \infty$, we have $\pi = 1$. $\square$

1.11 Appendix C: An Example of the Shortest Path Algorithm for RAM$_1$

Figure 1.6 illustrates a small example of the network construction for RAM$_1$. Due to limited space, we only considered five stages and did not generate all the nodes to $Y^{\max}$ (see Table 1.7 for the data). Five forms with processing times are listed in columns two through six. The seventh column shows the cost of hiring one employee for each stage. The eighth and ninth columns show the lower and upper values for each stage. The deadline is 20, hence $\overline{Y} = 3.85$. The shortest path is shown with dashed arcs. Thus the optimal solution is $(3, 4, 4, 4, 4)$, which incurs 675 units of cost.

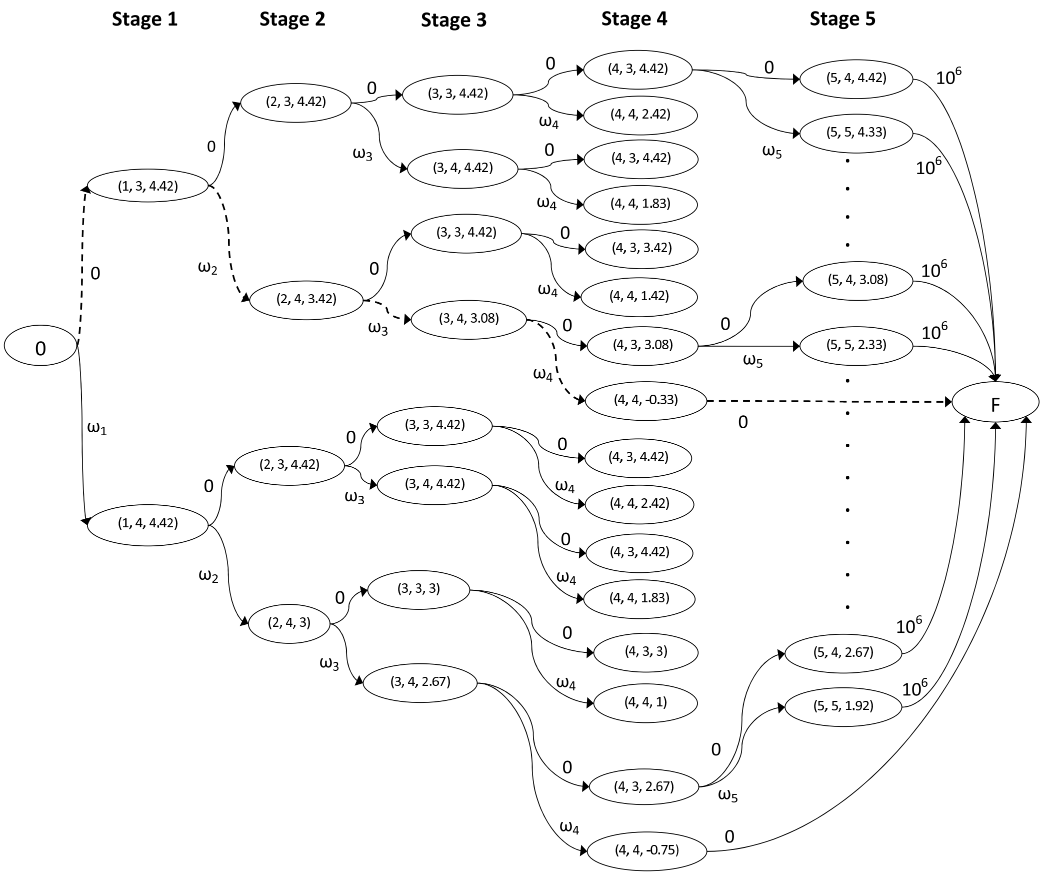


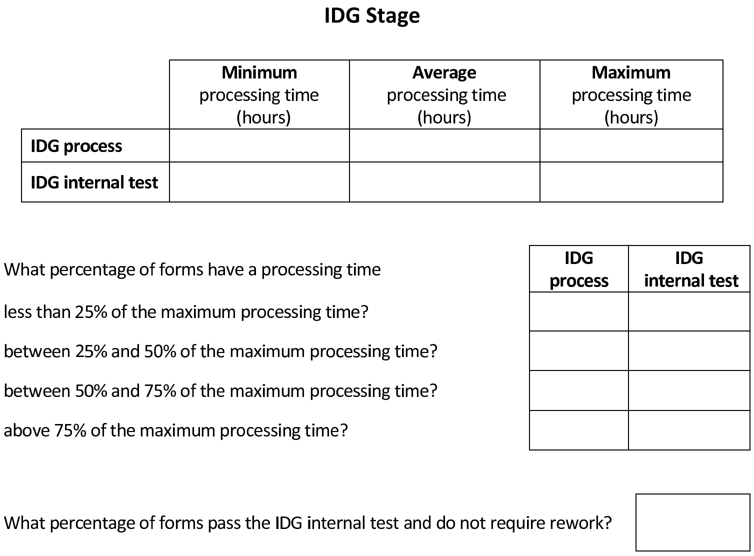
Figure 1.6: An example of constructing a network for RAM$_1$

Table 1.7: Data for the network example

Stage	Form 1	Form 2	Form 3	Form 4	Form 5	w_k	$Y_k^{\min}$	$Y_k^{\max}$
1	10	12	10	14	6	25	3	8
2	14	9	10	17	7	30	3	7
3	14	10	13	12	11	45	3	6
4	13	16	15	10	5	35	3	7
5	12	13	14	7	15	40	4	7

1.12 Appendix D: Questionnaires

By way of two questionnaires, TSDC managers provided a number of the estimates we used in our models. Figure 1.7 shows the questionnaire for obtaining processing time distributions and percentage of forms that require rework after internal tests. Though the example presents the questionnaire for the Image Development Group (IDG), we asked the same questions of all stages. Figure 1.8 shows the questions we asked for estimating the distribution of the destination of feedback loops from Integration & Final Test and for estimating the percentage of forms requiring one or more rounds of rework.



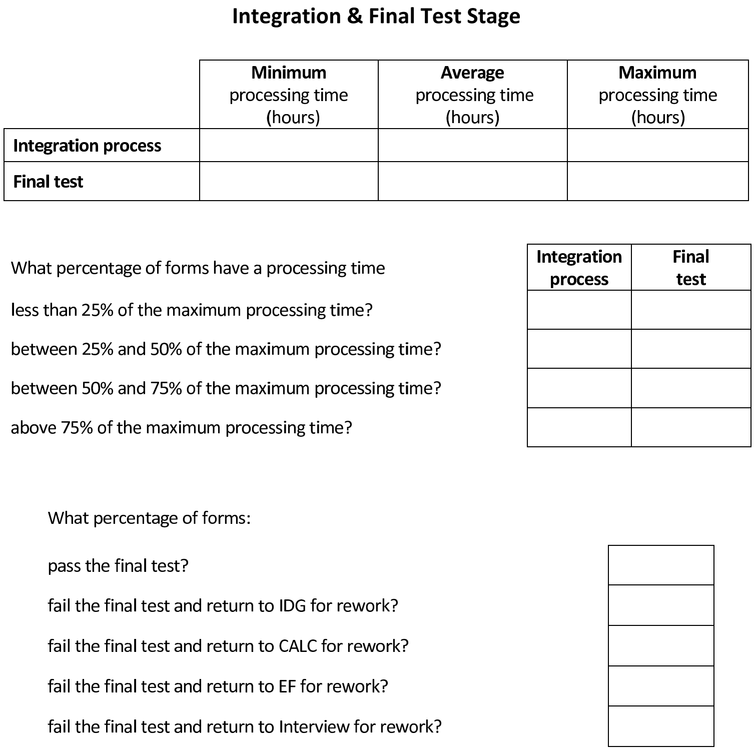
IDG Stage

	Minimum processing time (hours)	**Average** processing time (hours)	**Maximum** processing time (hours)
IDG process			
IDG internal test			

What percentage of forms have a processing time	**IDG process**	**IDG internal test**
less than 25% of the maximum processing time?		
between 25% and 50% of the maximum processing time?		
between 50% and 75% of the maximum processing time?		
above 75% of the maximum processing time?		

What percentage of forms pass the IDG internal test and do not require rework? □

Figure 1.7: Questionnaire for estimating the processing time distribution and rework probabilities



Integration & Final Test Stage

	Minimum processing time (hours)	**Average** processing time (hours)	**Maximum** processing time (hours)
Integration process			
Final test			

What percentage of forms have a processing time	**Integration process**	**Final test**
less than 25% of the maximum processing time?		
between 25% and 50% of the maximum processing time?		
between 50% and 75% of the maximum processing time?		
above 75% of the maximum processing time?		

What percentage of forms:	
pass the final test?	
fail the final test and return to IDG for rework?	
fail the final test and return to CALC for rework?	
fail the final test and return to EF for rework?	
fail the final test and return to Interview for rework?	

Figure 1.8: Questionnaire for estimating the rework probability distribution at the Final Test

Chapter 2

Coping with Gray Markets: The Impact of Market Conditions and Product Characteristics

2.1 Introduction

Manufacturers around the world confront new pressures with the trade of their brand name products in unauthorized distribution channels known as gray markets. Gray markets primarily emerge when manufacturers offer their products in different markets at different prices. Price differentials may motivate enterprises or individuals to buy products from authorized distributors in markets with a lower price and sell them in markets with a higher price. Gray market channels may operate in the same market as the authorized distributors, or bring parallel imports from another market.

Each year products worth billions of dollars are diverted to gray markets. In the IT industry alone, the approximate value of gray market products was \$58 billion dollars and accounted for 5 to 30 percent of total IT sales, according to a 2008 study conducted jointly by KPMG and The Alliance for Gray Market and Counterfeit Abatement (AGMA). In the pharmaceutical industry, 20% of the products sold in the United Kingdom are parallel imports (Kanavos and Holmes, 2005). In communications, nearly 1 million iPhones were unlocked in 2007 and used on unauthorized carriers worldwide (*New York Times*, 2008). International versions of college textbooks, drinks, cigarettes, automobile parts, luxury watches, jewelry, electronics, chocolates, and perfumes are among the numerous products that are traded in gray markets (Schonfeld, 2010).

Unlike black markets, products traded in gray markets are genuine. In the United States, gray markets are usually legal under the first-sale doctrine. Growing numbers of efficient global logistics networks help gray markets reach more customers faster. Advancing web technology and a rapidly growing online retail sector also boost gray markets. To name only a few, Amazon, eBay, Alibaba, Kmart, and Costco are among the retailers known to have sold gray goods (Bucklin, 1993; Schonfeld, 2010).

As to benefit and harm, opinions about gray markets are mixed, depending on one's perspective. Manufacturers generally consider gray markets harmful because products diverted to gray markets end up competing with those sold by authorized distributors, and unauthorized channels get a free ride from expensive advertising and other manufacturer efforts to increase sales. Also, brand value may erode as products become available to segments that the manufacturer deliberately avoided. Gray markets, however, can benefit manufacturers by generating a new stream of demand and providing a means for manufacturers to deter their competitors.

The existing literature on gray markets largely focuses on pricing decisions in deterministic settings. In this chapter, we consider a manufacturer that operates in two markets with uncertain demand under the threat of competition from a parallel importer. While consumer demand can be accurately estimated for some products or markets, in many cases manufacturers are challenged both with gray markets and uncertainty in demand. For instance, when the iPhone 4S was released, demand was much higher than expected and customers of some carriers had to wait more than three weeks to get the iPhone (*Wall Street Journal*, 2011). *PCWorld* (2011) reports that diversion to gray markets is a serious concern for the iPhone 4S. To the best of our knowledge, this chapter is the first work that analyzes price and quantity decisions of a manufacturer that faces demand uncertainty and parallel importation.

In our setting, if the manufacturer were to charge different prices across the markets, the parallel importer could buy the product in the low-price market and transfer it to sell in the high-price market. The manufacturer can control gray market activities through two levers: price and quantity. Consumers base their purchase decisions on perception and price, comparing the offering of the manufacturer to that of the parallel importer. They perceive gray markets to be inferior to authorized channels, valuing instead the peace of mind they get when they buy a product from an authorized distributer. This lower perception can also be attributed to characteristics of the product under consideration, with gray markets for some products being less attractive than others.

This chapter builds on and extends Ahmadi and Yang (2000) and makes three main contributions to the literature. First, we extend the literature on gray markets by incorporating demand uncertainty and production quantity decisions. Second, we explore the impact of market conditions and product characteristics on the high-level reaction of the manufacturer in response to gray market activities, such as ignoring, allowing, or blocking parallel imports – referred to as the manufacturer's policy hereafter. Third, we examine the value of making price and quantity decisions strategically in the presence of parallel importation. In doing so, we compare the manufacturer's profit under the strategic decision-making scenario to the profit of the myopic uniform pricing policy, which has been used by some companies, such as TAG Heuer and Christian Dior (Antia et al., 2004), and charges the same price in both markets to eliminate parallel importation entirely. We measure the value of strategic decisions under different market conditions and product characteristics.

To be more specific, we address the following questions in this research.

1. How does the presence of the parallel importer and demand uncertainty change the manufacturer's price and quantity decisions?

2. Is adjusting prices a more effective tool in controlling gray market activities or reducing the availability of the product?

3. How do price and quantity decisions define the manufacturer's policy against parallel importation such as ignoring, blocking, or allowing parallel imports?

4. What are the impacts of market conditions (such as market base, consumer price sensitivity, and demand uncertainty) and product characteristics (such as "fashion" or "commodity") on the manufacturer's policy?

5. When, if at all, should the manufacturer leave (enter) a market, when faced with the risk of parallel importation from one market to another?

6. How significant is the added value of making price and quantity decisions strategically instead of following a uniform pricing policy? Are there situations in which the uniform pricing policy serves as a good alternative to the strategic policy?

2.2 Literature Review

Despite the ubiquity of gray markets and their operational and marketing implications, this topic occupies a relatively small niche in the interface of marketing and operations management literature. Existing marketing and economics research into gray markets can be divided into two groups of studies. The first group includes empirical studies and qualitative discussions about gray markets. Myers (1999) surveys organizational, control, and market specific factors that lead to the emergence of gray markets. Banerji (1990), Maskus (2000), and Ganslandt and Maskus (2004) provide empirical evidence of gray market activities as well as an overview of the policy debate. Antia et al. (2004) discuss the impact of different policies on gray markets and methods trademark owners should employ to cope with them. Dutta et al. (1999) highlight the importance of business efficiency in territorial restriction policies.

The second group of studies includes analytical models for the price decision and whether or not gray markets should be deterred. Dutta et al. (1994) use an economic model to study the optimal policy towards retailers selling across their territories (bootlegging) and show the optimal policy is to tolerate some level of bootlegging. Bucklin (1993) examines the claims made by trademark owners and gray market dealers and draws public policy implications. Li and Maskus (2006) find that parallel imports inhibit innovation and diminish expected welfare if the manufacturer deters parallel imports with a high wholesale price. Richardson (2000) analyzes an economic model of countries deciding whether to prohibit gray markets or not. Matsushima and Matsumura (2010) and Chen (2009) use economic models to explore the ramifications of parallel imports for intellectual property holders and manufacturers. Results from these studies indicate that manufacturers should tolerate some level of territorial restriction violation. Coughlan and Soberman (1998) observe that gray markets can lead to higher profits when competing manufacturers sell products through retailers in a differentiated market with two types of customers who have different sensitivity to price and services. Xiao et al. (2011) show that whether the manufacturer sells directly or through a retailer is critical to determining the increase or reduction in manufacturer profit due to parallel importation. Shulman (2012) shows that competition among authorized retailers may lead to a prisoner's dilemma situation in which retailers divert to gray markets even if it does not increase total sales. Autrey et al. (2012) consider two firms that engage in a Cournot competition in a domestic market and face gray market activities when they enter a foreign market. They show that when the products are close substitutes, it is better to decentralize the management structure in the foreign market. Ahmadi and Yang (2000) (hereafter A&Y) investigate the interaction between a manufacturer and a parallel importer in a deterministic setting with endogenous prices. They show that not only does parallel importation increase total sales, but it can also increase manufacturer profit by serving customers with a low willingness to pay.

We adopt A&Y's framework for modeling parallel importation and market segmentation. However, the research questions we address differentiate our work from theirs (and others mentioned above) in several important directions: (1) we incorporate demand uncertainty, which means that both price and quantity are decision variables, whereas A&Y assume deterministic demand and their only decision is price. We show that ignoring demand uncertainty and only relying on the decisions of a deterministic model adversely impacts the manufacturer's profit; (2) we take into consideration the policy of blocking parallel importation by not entering a market; (3) we provide managerial insights about the adoption of a policy against parallel importation based on the characteristics of the markets and the product; (4) we provide insights about the value of a strategic reaction to parallel importation over using a myopic policy that has been applied in practice.

In the operations management literature, there exists a rich body of research on optimal pricing and quantity decisions with stochastic demand (Petruzzi and Dada, 1999; Chan et al., 2004); however, these studies ignore the potential for gray market activities.

Recently, some effort in the operations management literature has been devoted to analyzing quantity decisions and coordination in supply chains that face gray markets. In these papers, however, either price is exogenous or demand is deterministic. Ahmadi et al. (2012) consider a decentralized supply chain with exogenous pricing in which a retailer could salvage leftover inventory or sell it to the gray market. Altug and van Ryzin (2010) consider a manufacturer selling a product through a large number of retailers that face stochastic demand and sell their excess inventory to an internal gray market. They assume a market-clearing price for the gray market, but an exogenous retail price. Hu et al. (2010) consider a reseller who can place large orders to benefit from a supplier quantity discount offer and divert a portion of the order to a gray market. They show that when the reseller's batch inventory holding cost is high, the gray market improves channel performance. Su and Mukhopadhyay (2011) consider a deterministic setting in which a manufacturer offers a quantity discount to sell a product through one dominant retailer and N fringe retailers. Krishnan et al. (2010) study the impact of gray markets on a decentralized supply chain with one manufacturer and two retailers that may divert the product to the gray market, when demand is assumed to be deterministic. Our work differs from the foregoing in that we analyze the impact of parallel importation on a vertically integrated manufacturer who must set both prices and quantities before demand uncertainty is resolved. By deriving the solution to a game model, we show when it is in the manufacturer's interest to ignore, allow, or block parallel importation.

2.3 Analysis Framework

Consider a manufacturer who sells a single product in two separate markets. The manufacturer chooses price p_1 and quantity q_1 in market 1, and chooses price p_2 and quantity q_2 in market 2. Table 2.1 summarizes the notation used throughout the chapter. The demand in both markets is stochastic and additive, and defined as $\mathcal{D}_i\left(p_i, \epsilon_i\right) = d_i\left(p_i\right) + \epsilon_i = N_i - b_i p_i + \epsilon_i$ in which $d_i(p_i)$ denotes the deterministic component of demand, ϵ_i denotes the stochastic component of demand, N_i denotes the market base, and b_i represents the consumer sensitivity to price change in market i. We assume that ϵ_i takes its value in the interval $[L_i, U_i]$ with the probability density functions $f_i(x)$ and cumulative distributions $F_i(x)$. We denote the expected value and the standard deviation of ϵ_i with μ_i and σ_i, respectively. We assume that the coefficient of variations of ϵ_1 and ϵ_2 are such that the probability of a negative demand realization is negligible in the game model. We also assume that the hazard rate function of ϵ_i, denoted by $r_i(x) = \frac{f_i(x)}{1-F_i(x)}$, satisfies the Increasing Failure Rate (IFR) property; i.e., $\frac{d\,r_i(x)}{d\,x} > 0,\ \ \forall x \in [L_i, U_i],\ i = 1, 2$. This property holds for many common distributions such as Normal and Uniform.

The manufacturer has to determine her* prices and quantities before demand uncertainties are resolved. As depicted in Figure 2.1, after the manufacturer sets her prices, a parallel importer may decide to transfer the product from the low-price to the high-price market if the price gap makes the venture sufficiently profitable. The parallel importer must choose the quantity to buy from the manufacturer in the

*Throughout the chapters, we refer to the manufacturer as a female and to the parallel importer as a male.

Table 2.1: Notations for price and quantity

Parameters	
N_i, b_i, ϵ_i	base, price sensitivity, and demand uncertainty of market $i = 1, 2$
$[L_i, U_i]$, μ_i, σ_i	domain, expected value, and standard deviation of ϵ_i
$F_i(x), r_i(x)$	probability distribution and hazard rate function of ϵ_i
c	manufacturer's unit production cost
c_G	parallel importer's unit transfer cost
ω	consumer's relative perception of parallel imports
Manufacturer's Variables	
p_i, q_i	price and quantity in markets $i = 1, 2$
$\overline{\pi}$	Total profit when there are no parallel imports
π	Total profit in the presence of the parallel importer
π^d	Total profit in the presence of the parallel importer for deterministic demand
Manufacturer's Optimal Variables	
$\widetilde{p_i}$, $\widetilde{q_i}$	when there are no parallel imports
$\widetilde{p_i}^d$, $\widetilde{q_i}^d$	when there are no parallel imports and demand is deterministic
p_i^*, q_i^*	in the presence of the parallel importer
p_i^{*d}, q_i^{*d}	in the presence of the parallel importer for deterministic demand
Parallel Importer's Variables	
q_G, p_G, π_G	quantity, price, and profit

low-price market and then set the selling price in the high-price market.

We make two assumptions about the importer's ordering from the manufacturer. First, we assume that he places his order before other customers. Most gray marketers hire agents to swiftly purchase products, sometimes within a few hours of release. Also, in most situations it is very difficult for a manufacturer to distinguish between orders received from end customers and orders placed by gray market agents, especially when orders are placed through the Internet. Though, one is likely to assume that purchase volume may provide a clue, *New York Times* (2008) reports that more than one million iPhones were sold in gray markets. The second assumption is that the parallel importer makes his decisions based on an estimate of average demand and does not have the capability to estimate the uncertainty he would face. We believe this assumption is reasonable because gray marketers typically have low capital and cannot invest in market research to estimate the parameters of demand distribution. In contrast, most manufacturers that operate in international markets are large companies that have the experience and resources to study markets extensively and collect data. These two assumptions keep the model tractable and allow us to better analyze the impact of parallel importation on the manufacturer. We model this problem in a Stackelberg game framework with the manufacturer as the leader and the parallel importer as the follower. To analyze the impact of the parallel importer, we first consider the case of no parallel imports. We assume that there are no capacity constraints, and that unsatisfied demand is lost. For ease of exposition, we assume holding costs, lost-sales costs, and salvage values are zero. With c denoting the per-unit production cost, the manufacturer's problem can be formulated as

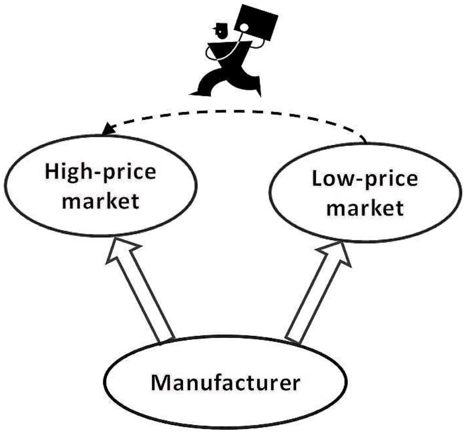


Figure 2.1: The manufacturer serves two markets and the parallel importer transfers the product between markets

$$\max_{p_1,p_2,q_1,q_2} \overline{\pi} = E\Big[p_1 min\Big\{q_1, \mathcal{D}_1\left(p_1, \epsilon_1\right)\Big\} + p_2 min\Big\{q_2, \mathcal{D}_2\left(p_2, \epsilon_2\right)\Big\} - c\left(q_1 + q_2\right)\Big] \quad (2.1)$$

This is the classic price-setting newsvendor problem (Petruzzi and Dada, 1999) in two independent markets and $\overline{\pi}$ is strictly quasiconcave in p_1 and p_2 (Xu et al., 2010). The optimal price, $\widetilde{p_i}$, solves

$$N_i - 2b_i p_i + z_i\left(p_i\right) + cb_i - \int_{L_i}^{z_i(p_i)} F_i\left(x\right) dx = 0, \qquad i = 1, 2 \tag{2.2}$$

in which $z_i\left(\widetilde{p_i}\right) = F_i^{-1}\left(1 - \frac{c}{\widetilde{p_i}}\right)$, and the optimal quantity is $\widetilde{q_i} = d_i(\widetilde{p_i}) + z_i\left(\widetilde{p_i}\right)$.

If the manufacturer's price is larger in one market than in the other, the parallel importer may consider transferring the product to the high-price market for resale. Since we lack *a priori* information as to which monopoly optimal price is higher than the other, we can impose conditions on the model parameters without loss of generality to ensure that $\widetilde{p_2} > \widetilde{p_1}$. The next proposition introduces this condition. All proofs are provided in Section 2.7.

Proposition 2.1. *In the absence of a parallel importer, the optimal price of the second market will be greater than the price of the first market ($\widetilde{p_2} > \widetilde{p_1}$) if and only if*

$$\frac{N_1 + L_1 + \displaystyle\int_{L_1}^{z_1(\widetilde{p_2})} \left(1 - F_1(x)\right) dx}{b_1} < \frac{N_2 + L_2 + \displaystyle\int_{L_2}^{z_2(\widetilde{p_2})} \left(1 - F_2(x)\right) dx}{b_2}. \tag{2.3}$$

Note that in the absence of demand uncertainty, the inequality in (2.3) reduces to $N_1/b_1 < N_2/b_2$, which is simply the equivalent condition when demands are deterministic. The added term $L_i + \int_{L_i}^{z_i} \left(1 - F_i(x)\right) dx$ accounts for the randomness in demand. Proposition 2.1 results in two sufficient conditions, which are provided in the following corollaries.

Corollary 2.1. *Suppose $b_1 = b_2$. If ϵ_2 stochastically dominates ϵ_1 in the first order ($\epsilon_2 \succeq_{s.t.} \epsilon_1$) and $N_2 > N_1$, then $\widetilde{p_2} > \widetilde{p_1}$.*

Corollary 2.2. *Suppose $N_1 = N_2$. If $b_2 < b_1$ and $\epsilon_2 \succeq_{s.t.} \epsilon_1$, then $\widetilde{p_2} > \widetilde{p_1}$.*

Corollary 2.1 states that if price sensitivities are identical in both markets, the manufacturer will charge a higher price in market 2 if it has the larger consumer base and demand stochasticity. Corollary 2.2 shows if the markets bases are same, the price in market 2 will be higher if its consumers are less price sensitive and demand is stochastically larger. We assume for the rest of the analysis that the parameter values are such that the direction of import is from market 1 to market 2.

Next we analyze a case when the parallel importer is present in the high-price market (market 2). By entering the high-price market, the parallel importer engages in a Stackelberg game with the manufacturer. We use backward induction to characterize the equilibrium of this two-stage game. That is, given the values of the manufacturer's price and quantity in both markets, we first derive the best response function of the parallel importer. Then, we characterize the manufacturer's optimal price and quantity decisions, taking into account the parallel importer's reaction.

2.3.1 Parallel Importer's Problem

In this section, we study the parallel importer's problem. For given price and quantity decisions by the manufacturer, define q_G to be the size of the order that the parallel importer places with the manufacturer in market 1; furthermore, let p_G be the gray market price in market 2. We assume that the parallel importer incurs cost of c_G to transfer one unit of the product to market 2. This cost represents the shipping cost and all other costs associated with distributing the product in market 2 (e.g., translating the user manual, repackaging, tariffs).

When there are no parallel imports and the manufacturer sets the price of market 2 at p_2, some customers buy the product and some do not. Once the parallel importer enters market 2 and offers the product at price p_G, the market divides into three segments as depicted in Figure 2.2. The first segment of the market is those customers who continue to buy the product from the manufacturer. The second segment contains customers who buy the product from the parallel importer. Some of these customers initially bought from the manufacturer, but now switch to the parallel importer (the distance between the dashed lines) and some had not considered buying the product before due to the higher price charged by the authorized channel. The third segment contains those who had not bought the product before and continue to refrain from doing so even after the parallel importer enters the market.

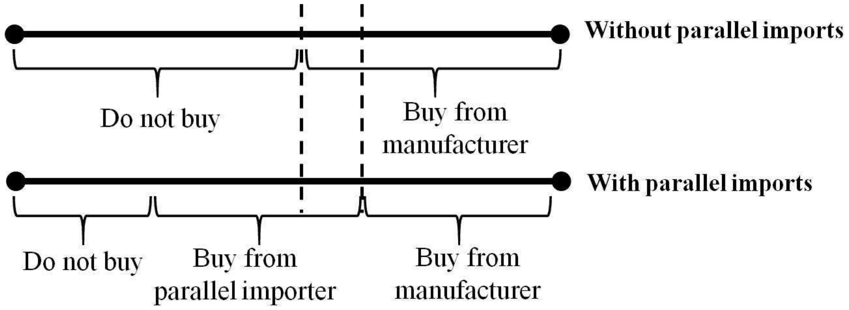


Figure 2.2: Segmentation of market 2 before and after parallel importation

The size of these segments is determined by the prices set by the manufacturer and the parallel importer. Size is also affected by the consumers' relative perception of gray-market products in market 2, whose valuation of parallel imports compared to their valuation of products provided by the manufacturer we denote with $0 < \omega < 1$. A low value of ω implies that consumers strongly prefer to buy the product from the authorized channel, whereas a high value of ω implies that consumers are relatively indifferent between buying from the authorized channel and buying from the gray market.

To determine the market segments, we note that the manufacturer's linear demand model $N_2 - b_2p_2$ is equivalent to assuming that the customers' net utility of consuming the manufacturer's product is equal to $\theta - \frac{b_2p_2}{N_2}$, in which θ is between 0 and 1. Given that the reputation of the parallel importer is lower than that of the manufacturer, we assume that the net utility of consuming parallel imports is $\omega\theta - \frac{b_2 p_G}{N_2}$. Now, if θ_1 is the boundary between the segment that buys from the manufacturer and the segment that buys from the parallel importer, we obtain it by equating the consumption utilities:

$$\theta_1 - \frac{b_2 p_2}{N_2} = \omega\theta_1 - \frac{b_2 p_G}{N_2} \Longrightarrow \theta_1 = \frac{p_2 - p_G}{N_2(1-\omega)} b_2.$$

Similarly, if θ_2 is the boundary between the segment that buys from the parallel importer and the segment that does not buy the product at all, we can write

$$\omega\theta_2 - \frac{b_2 p_G}{N_2} = 0 \Longrightarrow \theta_2 = \frac{b_2 p_G}{\omega N_2}.$$

Therefore, the net deterministic demand for the manufacturer is $N_2(1-\theta_1) = N_2 - \frac{p_2 - p_G}{1-\omega} b_2$ and the net demand for the parallel importer is $N_2(\theta_1 - \theta_2) = \frac{\omega p_2 - p_G}{\omega(1-\omega)} b_2$. However, because the parallel importer buys the product from the manufacturer in market 1, his order quantity is limited by q_1. Therefore, the parallel importer's problem is

$$\max_{p_G} \quad \pi_G = (p_G - p_1 - c_G)\, q_G$$

in which $q_G = min\left(\frac{\omega p_2 - p_G}{\omega(1-\omega)} b_2, q_1\right)$.

Proposition 2.2. *For any pair (p_1, p_2), let ψ be the quantity that maximizes parallel importer's profit, i.e., $\psi = \frac{\omega p_2 - p_1 - c_G}{2\omega(1-\omega)} b_2$. If the manufacturer's supply in market 1 is large enough to fulfill the parallel importer's quantity ψ, i.e., $q_1 > \psi$, the parallel importer's optimal price and quantity are*

$$p_G = \frac{\omega p_2 + p_1 + c_G}{2}, \qquad q_G = max\,(0, \psi)\,. \tag{2.4}$$

Otherwise,

$$p_G = \omega p_2 - \frac{\omega(1-\omega)q_1}{b_2}, \qquad q_G = q_1. \tag{2.5}$$

Equation (2.4) shows the parallel importer's optimal decisions when he is not constrained by manufacturer's quantity. The parallel importer incurs a cost of $p_1 + c_G$ to purchase the product in market 1 and transfer it to market 2. Clearly, if this cost is above the manufacturer's authorized price in market 2, p_2, transferring the product will not be profitable. However, because consumers in market 2 have a lower perception of gray market products, the importer's total purchase and transfer cost should be even lower (below ωp_2) to justify the importer's entry to the competition. As a result, the gray market is profitable if and only if $\omega p_2 > p_1 + c_G$. When the product availability is low, the parallel importer is forced to charge the price in (2.5) to clear his market.

2.3.2 Strategic Manufacturer's Problem

Having obtained the parallel importer's optimal decisions, we present the manufacturer's optimal price and quantity decisions in this section. In doing so, we also characterize the manufacturer's optimal response to gray market activities. Specifically, hereafter we make the following distinction between the manufacturer's *decisions* and her *policy*. We use the term decision to describe the manufacturer's price and quantity values. On the other hand, we define the manufacturer's policy as her high-level reaction to gray market activities as follows. The manufacturer can respond to the parallel importer in one of three ways:

1. ***Ignore the parallel importer.*** Under this policy, the manufacturer continues to use her optimal decisions in the absence of gray markets; i.e., $\widetilde{p_i}$ and $\widetilde{q_i}$. We

will later show that the manufacturer uses this policy only if the gray market can be automatically eliminated by using prices $\widetilde{p_1}$ and $\widetilde{p_2}$. Intuitively this would be the case when consumers' relative perception of parallel imports is sufficiently low ($\omega \ll 1$) or the parallel importer's transfer cost, c_G, is very high. Then, it would be too costly for the gray market to emerge independent of the manufacturer's prices. Thus, the manufacturer can simply ignore the parallel importer. A similar outcome can occur if the optimal prices of $\widetilde{p_1}$ and $\widetilde{p_2}$ are fairly close to each other. In this situation, these prices would render the gray market unprofitable unless ω is extremely high or c_G is extremely small.

2. ***Block parallel imports.*** If the difference between $\widetilde{p_1}$ and $\widetilde{p_2}$ is large enough for the gray market to operate, then the manufacturer can decide to block the parallel importer. Note that the manufacturer has two levers to block the gray market.

   - **Block using prices.** In this case, the manufacturer blocks the importer by altering her prices such that $\omega p_2 = p_1 + c_G$. This could be an effective policy when the price difference in the absence of gray markets ($\widetilde{p_2} - \widetilde{p_1}$), consumers' perception (ω), and importer's transfer costs (c_G) are such that the gray market could (barely) exist; however, a simple and small reduction in the price gap between the two markets could make parallel importation no longer profitable. Alternatively, the manufacture may opt for this policy when the parallel importer could emerge as a strong competitor. This can happen when the consumers' perception of parallel imports is high. The gray market, if allowed, could undercut the manufacturer and gain a significant portion of market 2.

   - **Block using quantity.** In this case, the manufacturer blocks the gray market by simply not offering the product to the parallel importer. Since there is usually no way of identifying the orders placed by the gray market, this policy is equivalent to setting $q_1 = 0$, or exiting/not entering market 1. The manufacturer may choose to use this policy when the consumers' perception of parallel imports is high ($\omega \approx 1$) and the manufacturer's optimal prices in the absence of the gray market are significantly different for the two markets ($\widetilde{p_1} \ll \widetilde{p_2}$). In this situation, blocking the parallel importer using prices simply becomes too costly. The manufacturer would lose significant portions of her profit if she insists on staying in both markets while trying to block the gray market. Therefore, she foregoes the relatively small profit in market 1 entirely to eliminate the parallel importer and only operates in market 2 using price $\widetilde{p_2}$. The impact of parallel importation on market entry/exit decisions has been witnessed by the pharmaceutical industry. For example, Eli Lilly and its European partner Boehringer Ingelheim have decided to delay the launch of the diabetes drug Trajenta in Germany, because they believe it would create opportunities for parallel importation and undermine their price structure in the European Union. Another example is Novartis's decision to stop the marketing of Rasilamlo, a drug for high blood pressure, in Germany (*2020health*, 2011).

3. ***Allow parallel imports.*** Under this policy, the manufacturer allows the importer to enter and resell the product in market 2; i.e., sets prices p_1 and p_2 such that $\omega p_2 - p_1 - c_G > 0$, and set $q_1 > 0$. The manufacturer would opt for this policy when $\widetilde{p_1}$ and $\widetilde{p_2}$ are moderately different and the consumers' perception of parallel imports is neither too high nor too low. Blocking the parallel importer in such a setting requires a relatively significant deviation

from otherwise optimal decisions. Furthermore, because consumers still highly value the authorized channel's products, the parallel importer is not a grave threat to the manufacturer. In this situation, the manufacturer would allow the gray market to emerge simply because the cost of blocking the parallel importer exceeds the cost of allowing parallel imports.

The next proposition describes the impact of parallel imports on the manufacturer's demand, which was also observed in the deterministic setting of A&Y.

Proposition 2.3. *When the parallel importer enters the competition, the manufacturer's demand in market 1 increases by q_G, and her demand in market 2 decreases by ωq_G.*

From Figure 2.2, we see that the manufacturer's demand in market 2 reduces because some consumers switch to the parallel importer. However, the segment of market 2 that buys the product from the parallel importer increases the manufacturer's demand in market 1. Overall, the manufacturer's demand goes up because parallel importation provides the product at a lower price and induces the consumers that have a lower willingness-to-pay to buy the product. The manufacturer could directly offer the product at a discounted price, but doing so through the authorized channel would lead to consumer confusion and severe demand cannibalization. Although the parallel importer also cannibalizes the demand of the authorized channel, this effect is alleviated because the importer is not affiliated with the manufacturer and has a lower reputation in the market.

The increase in total demand does not necessarily translate into higher profits because of the difference between the market prices. As a matter of fact, we will explain shortly that the manufacturer's expected profit is always lower in the presence of parallel importation.

In order to characterize the manufacturer's optimal decisions and policy, we first present an interesting insight gained from our model. Based on Proposition 2.2, if the manufacturer's quantity in market 1 is large enough, the parallel importer will order $\psi = \frac{\omega p_2 - p_1 - c_G}{2\omega(1-\omega)} b_2$, if ψ is positive. If the manufacturer's quantity in market 1 is lower than ψ, then the importer only gets a portion of his order and chooses a market-clearing price. However, as the next proposition shows, if the manufacturer's prices are such that a viable market could exist for the parallel importer, then the second scenario will always be dominated by the first scenario.

Proposition 2.4. *If the manufacturer's prices (p_1, p_2) are such that a gray market could exist, i.e., $\omega p_2 - p_1 - c_G > 0$, then the optimal policy of the manufacturer is to either (a) block the parallel importer using quantity; i.e., $q_1 = 0$, or (b) allow the parallel and provide a large enough quantity in market 1 to fulfill his entire order; i.e., $q_1 \geq \psi$.*

This proposition suggests that a mixed policy of allowing gray market activities through prices, but limiting the size of the gray market through quantity (i.e., allowing partial importation) is not optimal. Put differently, if the manufacturer chooses prices that allow for gray market activities and leave a segment of market 2 to the parallel importer, she will not reduce her quantity in market 1 to restrict the importer's sales. Note that we have assumed that the parallel importer is the first customer who receives the product. We expect that the insight obtained from Proposition 2.3 would even be strengthened if this assumption were to be relaxed. That is, if customers of the authorized channel can receive the product before the entire order of the importer is fulfilled, the manufacturer would have an even lower incentive to reduce the availability of the product in market 1 once parallel importation is allowed. Therefore, this result leads to an interesting insight that, once she decides to allow gray market activities, the manufacturer would not reduce product availability to limit parallel importation.

We can now formulate the Stochastic Stackelberg Game (SSG) and characterize the manufacturer's optimal decisions and policy

$$\begin{aligned}\text{(SSG)} \quad \max_{p_1,p_2,q_1,q_2} \pi \quad &= \quad E\Big[p_1 min\Big\{q_1, \mathcal{D}_1\left(p_1, \epsilon_1\right) + q_G\Big\} + p_2 min\Big\{q_2, \mathcal{D}_2\left(p_2, \epsilon_2\right) - \omega q_G\Big\} \\ & \qquad -c\left(q_1 + q_2\right)\Big]. \end{aligned} \tag{2.6}$$

The next theorem characterizes the solution to the SSG, which describes the structure of the manufacturer's optimal decisions and policy.

Theorem 2.1. *Suppose the manufacturer does not leave market 1. Let* $\widehat{p}_1$ *be the solution to the following equation*

$$\begin{aligned}\omega\left(N_2 - 2b_2\left(\frac{p_1 + c_G}{\omega}\right) + cb_2 + z_2\left(\frac{p_1 + c_G}{\omega}\right) - \int_{L_2}^{z_2\left(\frac{p_1+c_G}{\omega}\right)} F_2(x)\right) \\ + \omega^2\left(N_1 - 2b_1p_1 + cb_1 + z_1\left(p_1\right) - \int_{L_1}^{z_1(p_1)} F_1(x)\right) = 0. \end{aligned} \tag{2.7}$$

Then,

(a) $\widehat{p}_1$ *is unique.*

(b) If $\omega\widetilde{p}_2 - \widetilde{p}_1 - c_G \leq 0$*, then the manufacturer's optimal policy is to ignore the parallel importer. Thus,* $p_1^* = \widetilde{p}_1$ *and* $p_2^* = \widetilde{p}_2$*.*

(c) If $\omega\widetilde{p}_2 - \widetilde{p}_1 - c_G > 0$ *and* $\eta > 0$*, where*

$$\eta = N_1 - 2b_1\widehat{p}_1 + \frac{c_G}{2\omega\left(1-\omega\right)}b_2 + z_1\left(\widehat{p}_1\right) + c\left(b_1 + \frac{b_2}{2\omega}\right) - \int_{L_1}^{z_1(\widehat{p}_1)} F_1\left(x\right), \tag{2.8}$$

then the optimal policy is to block the parallel importer by setting

$$p_1^* = \widehat{p}_1, \quad p_2^* = \frac{\widehat{p}_1 + c_G}{\omega}.$$

(d) *If $\omega\widetilde{p}_2 - \widetilde{p}_1 - c_G > 0$ and $\eta \leq 0$, then it is optimal to allow parallel importation. In this case, p_1^* and p_2^* solve the following system of equations*

$$N_1 - 2b_1p_1 + \frac{2\left(\omega p_2 - p_1\right) - c_G}{2\omega\left(1-\omega\right)}b_2 + c\left(b_1 + \frac{b_2}{2\omega}\right) + z_1\left(p_1\right) - \int_{L_1}^{z_1(p_1)} F_1\left(x\right) = 0,$$

$$N_2 - 2b_2p_2 - \frac{2\left(\omega p_2 - p_1\right) - c_G}{2\left(1-\omega\right)}b_2 + c\frac{b_2}{2} + z_2\left(p_2\right) - \int_{L_2}^{z_2(p_2)} F_2\left(x\right) = 0.$$

(e) *Let p_1^*, p_2^* be the optimal prices obtained above. Then the manufacturer's optimal quantities are*

$$q_1^* = d_1\left(p_1^*\right) + q_G\left(p_1^*, p_2^*\right) + z_1\left(p_1^*\right), \qquad q_2^* = d_2\left(p_2^*\right) - \omega q_G\left(p_1^*, p_2^*\right) + z_2\left(p_2^*\right).$$

Before explaining the statement of Theorem 2.1, we present the following corollary, which suggests that for high enough values of ω, regardless of other model parameters, the optimal policy is always to block the parallel importer.

Corollary 2.3. *If consumers' relative perception of parallel imports is sufficiently high, i.e., if $\omega \approx 1$, then the manufacturer should block parallel importation.*

The manufacturer controls the parallel importer's order quantity through her prices. However, changing the prices also affects the demand of the authorized channels. Therefore, she should choose prices that balance these effects. For given prices p_1 and p_2, if the parallel importer is allowed to transfer the product, then the change in the manufacturer's total profit will be $(p_1 - \omega p_2 - c(1-\omega))q_G$, which is negative because the importer would enter only if $\omega p_2 - p_1 > c_G$. This means that, while total demand increases according to Proposition 2.2, the manufacturer's profit would always be less in the presence of a parallel importer. Thus, if $\widetilde{p_1}$ and $\widetilde{p_2}$ happen to block the importer (i.e., $\omega\widetilde{p_2} - \widetilde{p_1} - c_G \leq 0$), then the manufacturer simply ignores the parallel importer. This situation can arise in several circumstances. First, if ω is very low, the manufacturer does not need to worry about the gray market because consumers significantly differentiate between the manufacturer and the importer and are not much inclined to buy the product from the gray market. Second, if the importer incurs a high cost (c_G) for transferring the product to market 2, the difference between $\widetilde{p_1}$ and $\widetilde{p_2}$ may not be large enough to cover the costs. Third, if the difference between price sensitivities is small, then $\widetilde{p_1}$ and $\widetilde{p_2}$ will be naturally close and can prevent parallel importation, even if ω is moderately high or c_G is small.

When $\omega\widetilde{p_2} - \widetilde{p_1} - c_G > 0$, the manufacturer has to change her prices and deviate from $\widetilde{p_2}$ and $\widetilde{p_1}$. In this scenario, the optimal policy would be to either block the parallel importer (by setting $\omega p_2 - p_1 - c_G = 0$), or to allow him (by setting $\omega p_2 - p_1 - c_G > 0$). Thus, one can solve the SSG by imposing the constraint $\omega p_2 - p_1 - c_G \geq 0$. The parameter η defined in (3.6) is simply the shadow price of this constraint for $(\widehat{p}_1, \frac{\widehat{p}_1 + c_G}{\omega})$. Theorem 2.1 shows that the optimal blocking price, $\widehat{p_1}$, and its corresponding shadow price, η, are the factors that determine whether the optimal policy is to allow or block the parallel importer. If $\widehat{p_1}$ makes the corresponding shadow price positive, then the constraint will be tight and the manufacturer will block the importer via $\widehat{p_1}$ and $\frac{\widehat{p_1} + c_G}{\omega}$. However, if the shadow price is non-positive, then allowing parallel importation is the optimal policy.

When ω approaches zero, $\frac{c_G}{2\omega(1-\omega)} b_2 + c\frac{b_2}{2\omega}$ will be the dominating term in (3.6), making η positive and hence blocking the parallel importer is again the optimal policy. In this case, the gray market could barely exist as consumers have a very low perception of gray market products. Thus, a small reduction in the price gap between the two markets could make parallel importation no longer profitable. Therefore for low enough values of ω, the manufacturer's optimal policy is to block the importer by slightly altering her prices in both markets.

Similarly, when ω approaches 1, $\frac{c_G}{2\omega(1-\omega)}b_2$ is the dominating term and the manufacturer will block the parallel importer as suggested by Corollary 2.3. In this situation, products in the gray market become perfect substitutes for products in the authorized channel and the competition is highly intense. Therefore, the parallel importer could gain a relatively significant size of market 2 if allowed. Thus, the manufacturer is better off blocking the importer when ω is high enough.

Finally for intermediate values of ω when $\frac{c_G}{2\omega(1-\omega)}b_2 + c\frac{b_2}{2\omega}$ is small enough, the value of η may be negative and the manufacturer's optimal policy would be to allow the parallel importer. In this scenario, blocking the importer is simply too costly as it requires significant departure from optimal prices. The importer is not so weak to be blocked easily and not extremely competitive to pose a significant threat. Therefore for moderate values of ω, the optimal policy for the manufacturer is to let the importer enter market 2.

Theorem 2.1 assumes that the manufacturer is better off staying in both markets. The following corollary provides a necessary condition for when it is better for the manufacturer to exit (not enter) market 1 and only operate in market 2.

Corollary 2.4. *If the optimal policy for the manufacturer is to block the parallel importer by quantity (i.e., leave market 1) for given values of model parameters, then* $N_1 - b_1\left(\omega\widetilde{p_2} - c_G\right) + z_1\left(\omega\widetilde{p_2} - c_G\right) \leq 0$

The interpretation of the necessary condition is that $\omega\widetilde{p_2} - c_G$ must be too high to generate a positive demand (hence a positive quantity) in market 1. If the direction of the inequality is reversed, then the manufacturer can increase her profit by selling the product at price $\omega\widetilde{p_2} - c_G$ in market 1 and still block parallel importation.

Next we look at how the presence of the parallel importer changes the manufacturer's prices.

Proposition 2.5. *If the presence of the parallel importer forces the manufacturer to alter her otherwise optimal prices (i.e., $\omega\widetilde{p_2} - \widetilde{p_1} - c_G > 0$), then the manufacturer always increases her price in market 1 and reduces her price in market 2. That is, $p_1^* > \widetilde{p_1}$ and $p_2^* < \widetilde{p_2}$.*

The presence of the parallel importer forces the manufacturer to reduce her price gap, whether she allows or blocks the importer. When the manufacturer allows the importer to transfer the product, she increases her price in market 1 because the importer generates extra demand in that market. On other hand, because the movement of product to market 2 creates competition, the manufacturer needs to reduce her price in that market. If the manufacturer's decision is to block the importer, she has to choose prices so that $\omega p_2 - p_1 - c_G = 0$. Doing so by increasing p_1 or reducing p_2 alone severely hurts the manufacturer's demand in the authorized channels. Therefore, she reduces the price gap by adjusting p_1 upward and p_2 downward.

Next, we analyze the effect of demand uncertainty on the manufacturer's prices when she faces parallel imports. For this purpose, we define the Deterministic Stackelberg Game (DSG) as the deterministic version of the SSG in which ϵ_1 and ϵ_2 are replaced with their expected values as follows

$$\text{(DSG)} \quad \max_{p_1,p_2} \; \pi^d = \left(p_1^d - c\right)\left(\mathcal{D}_1\left(p_1^d, \mu_1\right) + q_G\right) + \left(p_2^d - c\right)\left(\mathcal{D}_2\left(p_2^d, \mu_2\right) - \omega q_G\right) \tag{2.9}$$

and $q_1^d = N_1 - b_1 p_1^d + \mu_1$ and $q_2^d = N_2 - b_2 p_2^d + \mu_2$. The next proposition compares the prices of the SSG to those of the DSG.

Proposition 2.6. *Given the same average demand, the optimal prices in the stochastic demand case (SSG) are always smaller than the optimal prices when demand is assumed to be deterministic (DSG); i.e.,* $p_1^* < p_1^{*d}$ *and* $p_2^* < p_2^{*d}$.

Prior work has shown a similar result for a single market with additive uncertainty in the absence of parallel importation. For example, Petruzzi and Dada (1999) show that the optimal price of a price-setting newsvendor who serves a single market is always below the optimal price when demand is equal to the average of the stochastic demand. Proposition 2.6 extends this result to the setting in which two markets are connected by parallel importation. We note that this result holds regardless of the policies adopted by the manufacturer under the SSG and DSG scenarios.

2.4 Numerical Experiments

This section and the next present numerical experiments that respond to the motivating questions raised in the introduction. More specifically, in this section we first highlight the value of developing a model that jointly incorporates both competition from parallel importers and demand uncertainty. We then explore the effect of parallel importation on the manufacturer's quantity, price gap, and profit. In the next section, we demonstrate managerial insights that can help address some policy questions of interest to the manufacturer, such as the transition of the manufacturer's optimal policy as model parameters vary, and the value of strategic decision-making in the presence of parallel importation.

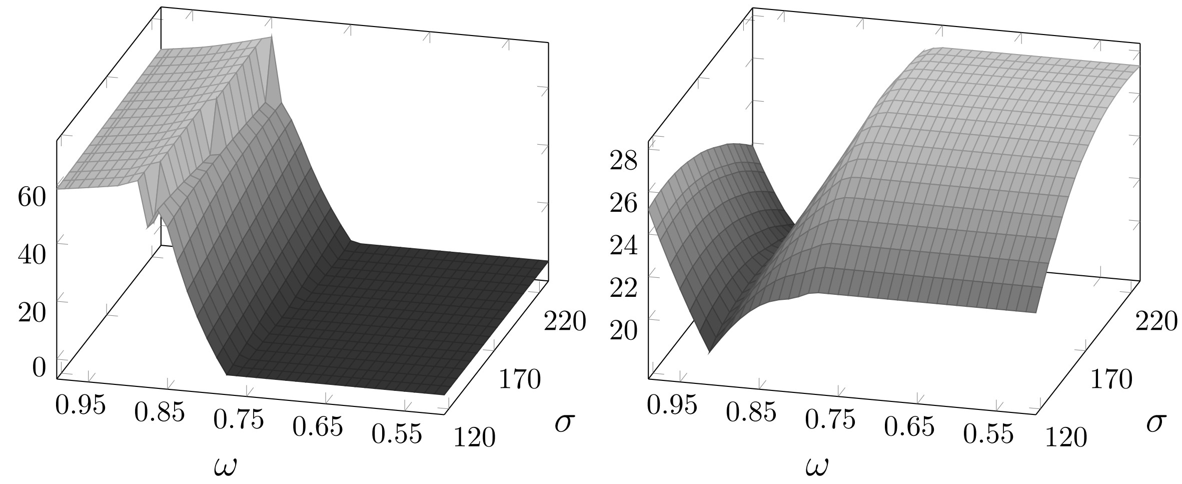
We implemented decisions and evaluated outcomes for more than 250 cases. We based the ranges of the parameter values on the estimated production cost of the iPad and 2010 sales figures and average price in the United States (*Computer World*, 2010; *eMarketer*, 2010), assuming a linear demand-price curve. We varied the manufacturer's cost, c, from \$200 to \$350, and set the parallel importer's transfer cost to $c_G = \$10$. We varied b_1 from 1.25 to 4, varied b_2 from 1 to 2, and varied N_1 and N_2 from 1,000 to 4,000. To account for demand variability, we assumed ϵ_1 and ϵ_2 to have the same distribution, but not necessarily the same parameters. We focused on Uniform and Normal distributions as they are widely used in the literature (e.g., Schweitzer and Cachon, 2000; Yao et al., 2006). Because the behaviors we observed for Normal distribution were not significantly different from those for Uniform distribution, we only present the figures for Uniform distribution.

2.4.1 Value of a Joint Model

In this chapter, we have developed a unified framework that depicts a manufacturer that faces gray market activities and uncertain demand. As pointed out in the literature review section, prior work in marketing and operations management largely dealt with price and quantity decisions in the presence of either a parallel importer or uncertain demand. The joint model presented here considers the simultaneous effects of both parallel importation and demand uncertainty.

Cost of Ignoring Parallel Importation

In the first set of experiments, we determine how much profit the manufacturer would forfeit if she ignores the possibility of parallel importation and treats each market independently. In our experiments, we observe that the magnitude of profit losses can be as high as 70%, depending on consumers' perception and demand uncertainty.



(a) using $\widetilde{p_1}$ and $\widetilde{p_2}$ as a heuristic solution

(b) using p_1^* and p_2^* as a heuristic solution

Figure 2.3: Error percentage of ignoring parallel imports or demand uncertainty. $N_1 = N_2 = 1500, b_1 = 3, b_2 = 2, \sigma_1 = \sigma_2 = \sigma$.

Figure 2.3(a) is an example of our experiments and shows the percentage of the manufacturer's profit loss when she continues to use $\widetilde{p_1}$ and $\widetilde{p_2}$ for different values of ω and $\sigma_1 = \sigma_2 = \sigma$. When parallel imports have a very low reputation, parallel importation is not a threat, and there is no profit loss. As ω increases, however, the profit loss increases. Even for relatively moderate values of ω, the manufacturer can lose between 20% to 30% of her profit if she ignores the parallel importer. For larger values of ω, profit losses exceed 50%.

Cost of Ignoring Demand Uncertainty

We now evaluate the profit loss to the manufacturer when she is aware of the presence of the parallel importer, but ignores demand uncertainty. For this purpose, we first solve the DSG in (2.9) and obtain its optimal solution, $(p_1^{*d}, p_2^{*d}, q_1^{*d}, q_2^{*d})$. We then evaluate the profit of the SSG for the deterministic decision variables. In our test set, the percentage reduction in manufacturer's profit varies between 1% and 30%

depending on the value of ω and the magnitude of demand uncertainty. Figure 2.3(b) illustrates one such example in which ignoring demand uncertainty is detrimental to the manufacturer and could cost her between 18% and 27% of profit. Therefore, it is crucial that manufacturers account for both uncertainty in demand and parallel importation.

2.4.2 Impact of Parallel Importation

Quantities

Proposition 2.2 states that when the parallel importer transfers the product, demand for the manufacturer in market 1 increases (due to orders placed by the parallel importer), while her demand in market 2 decreases (due to some customers switching to the gray market). Therefore, one would expect that under the allow policy and compared to when there are no parallel imports, the manufacturer would store more of the product in market 1 in order to maintain the same service level to her non-parallel importer customers and stock less in market 2, because she will lose the low-end segment of market 2 to the parallel importer. Interestingly, in our numerical experiments we observe that the opposite effect occurs: the manufacturer's quantity in market 1 will be below the quantity level in the absence of parallel imports, and her quantity in market 2 will be more than the quantity before the presence of the parallel importer.

This behavior can be explained as follows. Parallel importation influences the manufacturer's quantities in two ways. First, it increases the demand in market 1 by q_G and decreases the demand in market 2 by ωq_G. Thus, the manufacturer would like to increase q_1 and decrease q_2 accordingly. Second as shown in Proposition 2.5, it forces the manufacturer to increase p_1 and reduce p_2. Because demand is decreasing in price, the demand of the authorized channel in market 1 decreases while the demand of the authorized channel in market 2 increases. The second effect proves to be stronger, and ultimately the manufacturer keeps a lower stockpile in market 1 and a higher quantity in market 2. This tradeoff can be shown analytically when demand is deterministic. Proposition 2.7 formalizes this argument.

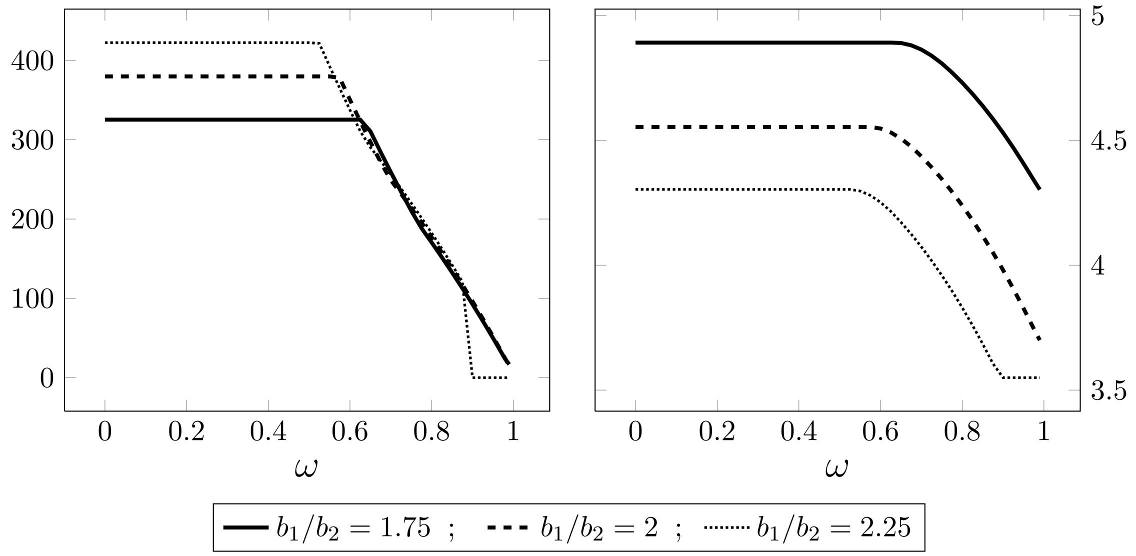
Proposition 2.7. *In the DSG, the optimal quantity in market 1 (market 2) in the presence of parallel importation is smaller (larger) than the optimal quantity when there are no parallel imports, i.e.,* $q_1^{*d} < \widetilde{q_1}^d$ *and* $q_2^{*d} > \widetilde{q_2}^d$*.*

A&Y made the same observation when the manufacturer selects the block policy. We show that the same direction of changes is valid regardless of the policy chosen by the manufacturer. While Proposition 2.7 proves the result for the deterministic demand case, we observed the same behavior across all the experiments when demand was random.

Price Gap and Profits

In this section, we extend the experiments to assess the effect of ω on the manufacturer's profit and price gap between the two markets. Figure 2.4(a) shows the manufacturer's price gap for values of ω. We observe that the price gap is non-increasing in ω. This is hardly surprising as when consumers have high valuation for gray-market products, the competition intensifies and the manufacturer is forced to reduce her price gap in order to reduce the profit margin of the parallel importer.

The reduction in price gap leads to reduction in profit, as we see in Figure 2.4(b). Although the total profit is non-increasing in ω, the profit in each market is not monotone. Figures 2.5(a) and (b) show the profit in market 1 and market 2, respectively. When ω exceeds $\frac{\widetilde{p_1}+c_G}{\widetilde{p_2}}$, the profit in both markets goes down because the manu-

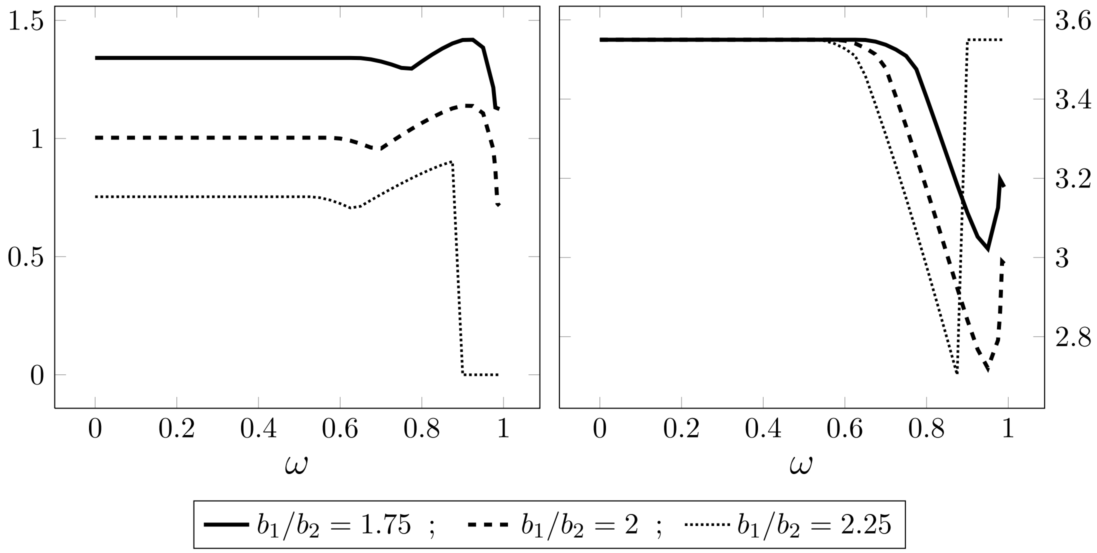


(a) Manufacturer's price gap (b) Manufacturer's total profit (in \$100K)

Figure 2.4: Manufacturer's price gap and total profit ($N_1 = N_2 = 1500$, $\epsilon_1, \epsilon_2 \sim [-200, 200]$)

facturer increases p_1 and reduces p_2 to block the parallel importer. As ω increases further, the manufacturer is better off allowing parallel importation. Thus, the profit in market 1 increases by selling to the parallel importer. However, the profit in market 2 declines because the manufacturer is losing market share. When ω is very high, the revenue from selling to the parallel importer in market 1 no longer outweighs the loss of profit in market 2. At this point, it is better for the manufacturer to block the importer. Therefore, profit in market 2 increases while the profit in market 1 declines.

As we mentioned earlier, one strategy for counteracting gray markets is to raise consumer awareness about the consequences of buying products from gray markets. Figure 2.6 shows how much the manufacturer can increase her profit by investing in programs that encourage consumers to buy the product from the authorized channel and reduce their relative perception of parallel imports (from $\omega = 0.9$). We notice



(a) Manufacturer's profit in market 1 (b) Manufacturer's profit in market 2

Figure 2.5: Manufacturer's profit in each market (in \$100K) ($N_1 = N_2 = 1500$, $\epsilon_1, \epsilon_2 \sim [-200, 200]$)

that a 10% reduction in consumers' perception increases manufacturer profit between 2% and 6%, and that the value of influencing perception is increasing in the price differential. We also notice that the marginal value of reducing perception is decreasing. This is because when perception is sufficiently low, the manufacturer can easily ignore parallel importation or block with a small change in $\widetilde{p_1}$ and $\widetilde{p_2}$. Thus, reducing consumer perception is no longer beneficial.

2.5 Managerial Insights

In this section, we generate managerial insights to inform the debate over policies and strategic decisions that a manufacturer facing the threat of parallel importation would consider. We present most of the results of this section through the lens of two important dimensions of our model: market-specific parameters and product-specific

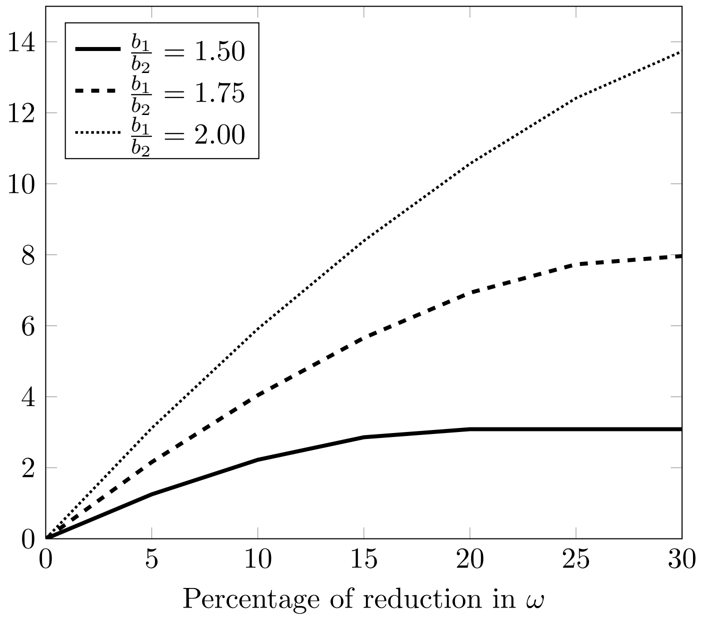


Figure 2.6: Percentage of increase in profit versus percentage of reduction in ω ($N_1 = N_2 = 1500$, $\epsilon_1, \epsilon_2 \sim [-200, 200]$)

parameters. Market-specific parameters describe the features of each market, such as market base, price sensitivity, and demand uncertainty. Product-specific parameters describe the item under consideration.

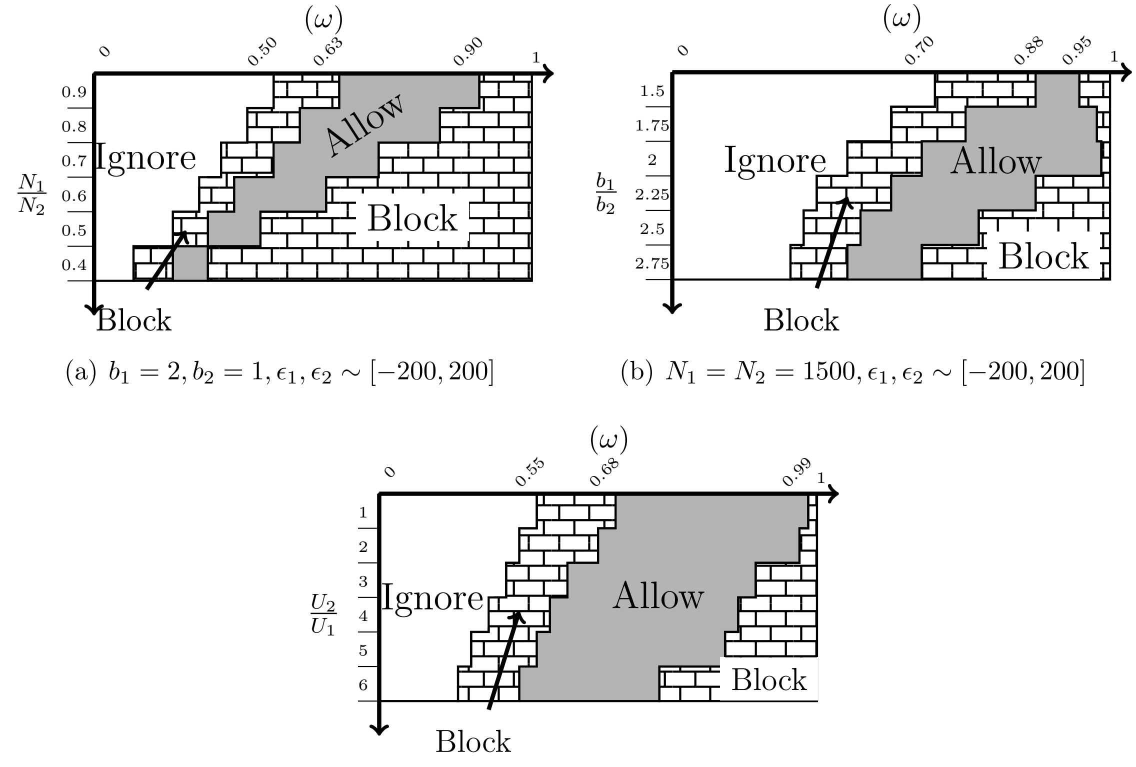
With that understanding, we define *market conditions* as the aggregate effect of relative market-based parameters, such as relative market bases (N_1/N_2), price sensitivities (b_1/b_2), and relative demand uncertainties $\epsilon_1/_{s.t.}\epsilon_2$ in which $/_{s.t.}$ represents the magnitude of ϵ_2 stochastically dominating ϵ_1 in the first order. We say market conditions are *similar* if the parameters of the markets are such that the price gap would naturally be small even if there are no parallel imports (small $\widetilde{p_2} - \widetilde{p_1}$). On the other hand, we say market conditions are *different* if the price gap would naturally be large in the absence of parallel imports (large $\widetilde{p_2} - \widetilde{p_1}$).

The parameter that represents product characteristics in our model is ω. A low value of ω means that parallel imports and authorized-channel products are quite distinct in the eyes of consumers. The higher the ω, the more intense the competition between the manufacturer and the parallel importer. We define *commodity* items as products for which consumers have a relatively high perception of parallel imports and are almost indifferent between the authorized channel and the gray market (i.e., $\omega \approx 1$). At the other end of the spectrum, we define *"fashion"* items as products for which consumers have a relatively low perception of parallel imports ($\omega \ll 1$). The factors that can determine the perception of parallel imports include the maturity of the product, consumers' knowledge of and familiarity with the product, and the risks associated with buying from gray markets.

We begin teasing out the policies for the manufacturer by exploring the implications of various responses to parallel importation and the effects of product characteristic and market conditions on the manufacturer's optimal policy. Then, we compare the strategic pricing policy that our model prescribes to a uniform-pricing policy in which the manufacturer eliminates parallel importation by charging the same price in both markets. Finally, we briefly describe the effect of product characteristics on the parallel importer's order size and profit.

2.5.1 How Do Market Conditions and Product Characteristics Determine Policy?

Though determining the optimal price and quantity decisions for a strategic manufacturer has been the main focus of this chapter, an important high-level question for any manufacturer is: What is the best policy to cope with the gray market? In this section, we examine, through extensive experiments, the manufacturer's optimal policy in response to gray market activities based on various market conditions and product characteristics. Our numerical experiments indicate that neither of the policies completely dominates the others. In fact, each can emerge as the optimal policy for a certain range of parameters. By characterizing the regions of policies, we can illustrate the transition from one policy to another as product characteristics and market conditions change.



(a) $b_1 = 2, b_2 = 1, \epsilon_1, \epsilon_2 \sim [-200, 200]$

(b) $N_1 = N_2 = 1500, \epsilon_1, \epsilon_2 \sim [-200, 200]$

(c) $N_1 = N_2 = 1500, b_1 = 2, b_2 = 1$

Figure 2.7: Optimal policy regions for the manufacturer.

Figure 2.7 depicts the simultaneous effects of product characteristics and different measures of market conditions on the regions characterizing the optimal policy. In all three graphs, the horizontal axis represents the product characteristic, ranging from fashion to commodity from left to right; the vertical axis, denotes various measures of market conditions, ranging from similar to different conditions, from top to bottom. In Figures 2.7(a)-(b), the ratio of market bases and price sensitivities represent market conditions, respectively. In Figure 2.7(c), the ratio U_2/U_1 captures market conditions in which $\epsilon_1 \in [-200, 200]$, $L_2 = -200$, and U_2 varies such that ϵ_2 grows stochastically larger than ϵ_1.

We observe the same pattern in all three graphs. When the product is a fashion item and market conditions are somewhat similar, the manufacturer ignores the parallel importer. As parallel imports gain acceptance from consumers and/or market conditions somewhat differ, the manufacturer's policy is to block the parallel importer by slightly deviating from her otherwise optimal decisions. When there is a higher perception of parallel imports or when market conditions are moderately different, it is no longer beneficial for the manufacturer to block the gray market, as it requires large deviations from her otherwise optimal decisions in each market. In this region, parallel imports are allowed into market 2. Finally, when the product is a commodity or when market conditions are significantly different, the manufacturer goes back to blocking the parallel importer. We note that blocking the gray market can be done by price or quantity. In our experiments, we observed that blocking by quantity only happens when the market conditions are vastly different and the product is a fashion item.

Figure 2.8, which is one of the main managerial insights of this chapter, qualitatively summarizes the effects of product characteristics and market conditions on the manufacturer's optimal policy. Of course, one should consult Theorem 2.1 to determine the regions precisely for given ranges of model parameters. Section 2.8provides a closed-form solution for the profit of each policy when demand is deterministic.

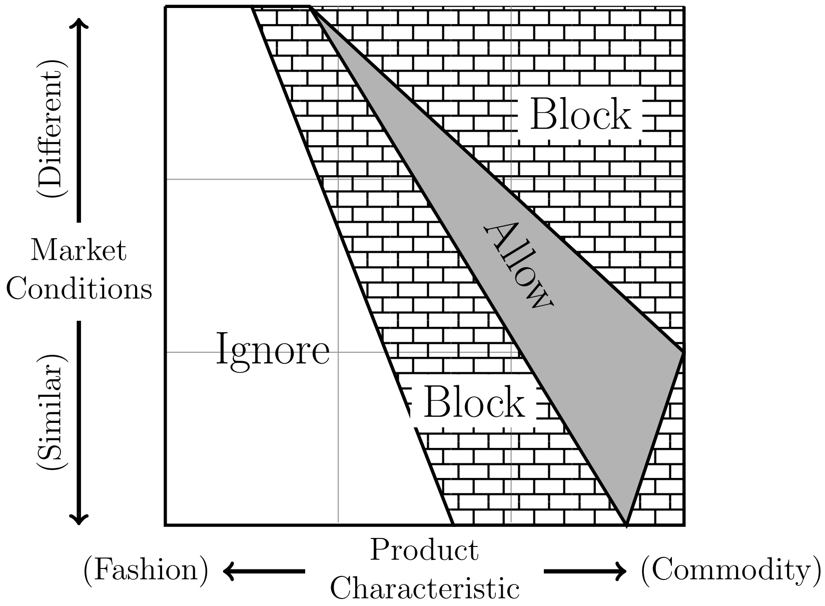


Figure 2.8: Optimal policy as a function of product characteristics and market conditions.

2.5.2 Strategic Versus Uniform Pricing

Implementing the optimal price and quantity decisions prescribed by the SSG requires estimating the value of parameters, such as the relative perception of parallel imports, ω, and the parallel importer's transfer cost, c_G, among others specific to the gray market. In practice, some manufacturers such as TAG Heuer and Christian Dior have adopted a uniform pricing policy and charge the same price for their products across all markets (Antia et al., 2004) to eliminate gray markets entirely. The uniform pricing policy requires less information and facilitates price coordination. Nevertheless this policy bears the risk of fluctuations in exchange rates and the manufacturer parts with the added profit from using the market-specific prices of the SSG.

We conducted experiments to provide insights into the impact of market conditions and product characteristics on the extra profit the manufacturer would earn by using the SSG recommendations. The optimal uniform price, p^u, can be obtained by solving

(3.1), while enforcing $p_1 = p_2$, and we have $\widetilde{p_1} \leq p^u \leq \widetilde{p_2}$.

Figure 2.9 demonstrates a common behavior we observed in these experiments. It illustrates the ratio of the manufacturer's profit under the SSG to her profit under the uniform pricing policy as a function of ω for various market conditions captured with relative price sensitivities, b_1/b_2. The same pattern of behavior will be observed if market conditions are represented by other market parameters.

We observe two important behaviors in this graph. First, despite the benefits of the uniform pricing policy such as easier implementation, there is a large range of ω and market conditions in which the manufacturer's profit will be significantly higher, as high as 25%, if she uses the strategic prices rather than using the uniform price. Therefore, in many situations it is crucial that the manufacturer put effort into market research to obtain the necessary information for strategic prices.

The second observation is that the benefit of strategic pricing is greatest when market conditions are not too similar or too different, namely they are *moderately* different, and the product has not turned to a commodity. For example when $b_1/b_2 = 3.25$ (very different market conditions), the additional profit from strategic pricing is slightly above 5% and is even lower when $b_1/b_2 = 1.25$ (very similar market conditions). The reason is that when market conditions are very similar, the price gap in the absence of parallel imports, $\widetilde{p_2} - \widetilde{p_1}$, is naturally small and the manufacturer blocks the parallel importer with prices for most values of ω. Since p^u is between $\widetilde{p_1}$ and $\widetilde{p_2}$, the manufacturer would not lose too much if she charges p^u in both markets and blocks parallel importation. On the other hand, when market conditions are very different, the manufacturer would not be able to charge a single price that attracts significant portions of both markets simultaneously. In this case, the uniform pricing policy reduces to $p^u = \widetilde{p_2}$, and the manufacturer leaves market

1 for the sake of the higher profit in market 2. For strategic pricing, although the manufacturer has the opportunity to boost her profit by charging markets differently, if the manufacturer wants to stay in both markets, she will have to charge very different prices. The parallel importer can exploit the high price gap and transfer a large amount of the product, reducing the manufacturer's share in market 2. As a result, the small profit from market 1 no longer pays off the loss of profit in market 2 due to intensive parallel importation, especially when ω is high. Therefore, the manufacturer eventually decides to only operate in market 2, which means both the strategic policy and the uniform pricing policy become identical.

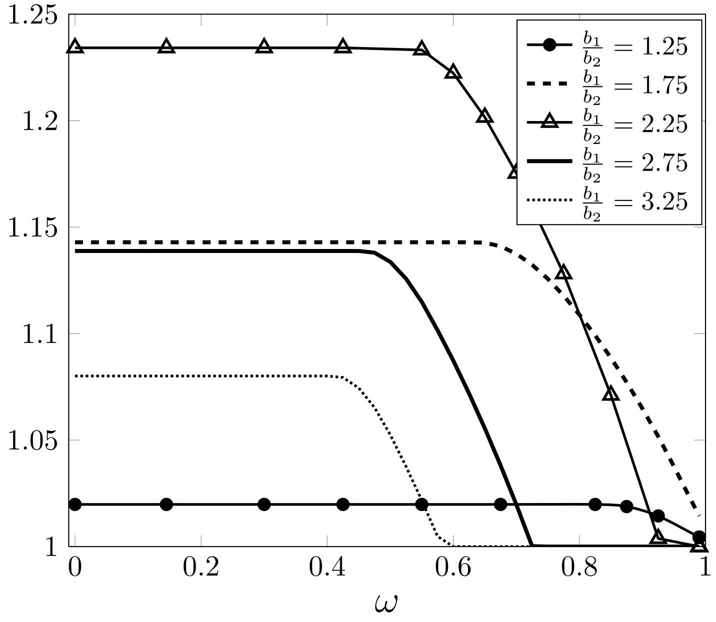


Figure 2.9: Ratio of optimal profit in the SSG to the uniform pricing profit ($N_1 = N_2 = 1500$, $\epsilon_1, \epsilon_2 \sim [-87, 87]$).

In summary, we find that strategic pricing leads to a significant increase in profits for non-commodity products when the two markets are moderately different. In extreme cases, however, when markets are either too similar or too different and when the product is a commodity, a simple uniform pricing policy can be considered a viable alternative to strategic pricing.

2.5.3 Parallel Importer's Problem

We close this section with Figure 2.10, which shows the parallel importer's profit versus consumers' perception. Interestingly, the parallel importer's profit is a unimodal function of consumer's perception. As the product becomes almost a commodity, the parallel importer's profit decreases due to the manufacturer's aggressive pricing as stated in 2.3. The parallel importer's profit is maximized when perception of parallel imports is moderately high. Therefore, even though higher perception strengthens the parallel importer's position in the competition with the manufacturer, a degree of differentiability between the gray market and the authorized channel is actually something the importer needs to survive in the competition and achieve maximum profit.

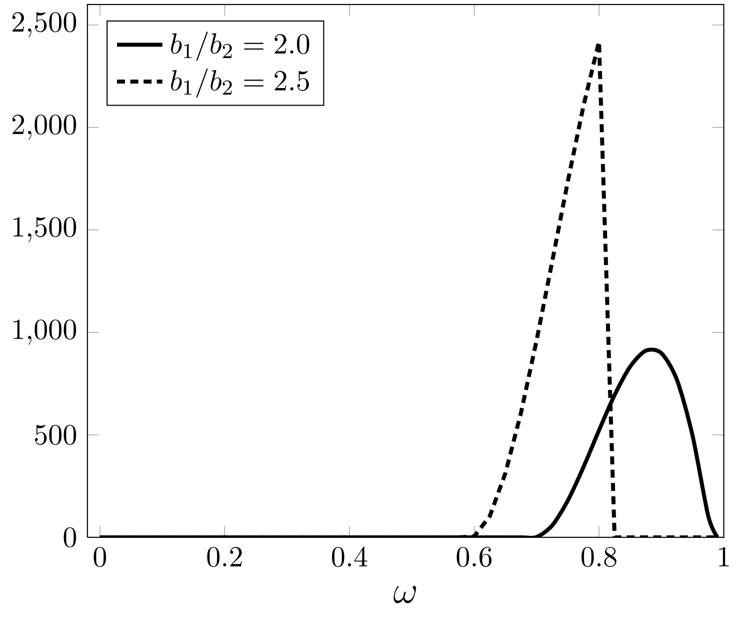


Figure 2.10: Parallel importer's optimal profit. ($N_1 = N_2 = 1500$, $\epsilon_1, \epsilon_2 \sim [-87, 87]$).

2.6 Conclusion

In this chapter, we analyzed the impact of parallel importation on a manufacturer's price and quantity decisions in an uncertain environment, and showed that reducing price gap is more effective in controlling the gray market than reducing product availability. We found that the manufacturer's policy depends heavily on market conditions and product characteristics. If the product is a fashion item, the manufacturer eliminates the parallel importer. For similar market conditions, elimination may be possible without adjusting prices. However, prices need to be adjusted when market conditions are different. The manufacturer also eliminates the importer when the product is a commodity. In that case, she may be forced to leave the less profitable market and only serve the more profitable market. Finally, if the product is in transition from a fashion item to a commodity, the manufacturer allows the importer to operate if the market conditions are moderately different.

We also showed that strategic pricing is significantly more valuable than a uniform pricing policy when the product is not a commodity and market conditions are moderately different. Thus, in these situations it is worth investing in market studies to have a better understanding of market parameters and consumer's perception of gray goods, and set prices strategically. However, if market conditions are too similar or too different and the product is a commodity, uniform pricing is a good alternative to strategic pricing.

Our work has limitations that can be addressed in future research. It would be interesting to analyze the impact of parallel importation on the manufacturer's decisions and policy in a multi-period setting in which the manufacturer and the parallel importer interact repeatedly. We assume that the parallel importer only relies on an estimate of the average demand. One natural extension is to assume that the importer has the means to estimate the parameters of his demand distribution. Finally, we assume that the manufacturer has unlimited capacity. Limited capacity will impact the manufacturer's allocation of quantities to each market, which then changes her prices. Also, because the importer can acts as an agent who transfers the product between markets, he can influence the manufacturer's capacity investment decisions especially when capacity costs are different across the markets.

2.7 Appendix A. Proofs

Proof of Proposition 2.1. Note that $z_1(p) - \int_{L_1}^{z_1(p)} F_1(x) = L_1 + \int_{L_1}^{z_1(p)} (1 - F_1(x))$. Hence,

$$\begin{aligned}
\left.\frac{\partial \pi}{\partial p_1}\right|_{p_1=\widetilde{p_2}} =& N_1 - 2b_1\widetilde{p_2} + cb_1 + L_1 + \int_{L_1}^{z_1(\widetilde{p_2})} (1 - F_1(x)) \\
=& \frac{b_1}{b_2}(-2b_2\widetilde{p_2} + N_2 + cb_2) - \frac{b_1}{b_2}N_2 + N_1 + L_1 + \int_{L_1}^{z_1(\widetilde{p_2})} (1 - F_1(x)) \\
=& \frac{b_1}{b_2}\left(-L_2 - \int_{L_2}^{z_2(\widetilde{p_2})} (1 - F_2(x))\right) - \frac{b_1}{b_2}N_2 + N_1 + L_1 + \int_{L_1}^{z_1(\widetilde{p_2})} (1 - F_1(x)) \\
=& b_1\left(\frac{N_1 + L_1 + \int_{L_1}^{z_1(\widetilde{p_2})} (1 - F_1(x))\,dx}{b_1} - \frac{N_2 + L_2 + \int_{L_2}^{z_2(\widetilde{p_2})} (1 - F_2(x))\,dx}{b_2}\right).
\end{aligned}$$

The third equality is due to (2.2). Since the profit function is strictly quasiconcave in p_1 and p_2, $\widetilde{p_2} > \widetilde{p_1}$ if and only if the expression in the last line is negative. □

Proof of Proposition 2.2. First consider the case of $\frac{\omega p_2 - p_G}{\omega(1-\omega)}b_2 < q_1$. Then $\pi_G = (p_G - p_1 - c_G)\frac{\omega p_2 - p_G}{\omega(1-\omega)}b_2$ is a concave function in p_G. The first order optimality condition and $q_G \geq 0$ give us (2.4), and the feasibility condition $\psi < q_1$. If $\frac{\omega p_2 - p_G}{\omega(1-\omega)}b_2 \geq q_1$,

$\pi_G = (p_G - p_1 - c_G)q_1$ and it is optimal to increase p_G as much as possible. Thus $\frac{\omega p_2 - p_G}{\omega(1-\omega)} b_2 = q_1$, which gives us (2.5). $\square$

Proof of Proposition 2.3. Suppose the manufacturer chooses her prices such that $\omega p_2 - p_1 - c_G > 0$, but she chooses $q_1 \leq \psi$. Then her profit will be

$$\max_{q_1, q_2} \pi \quad = \quad E\Big\{p_1 q_1 + p_2 min\left(q_2, \mathcal{D}_2\left(p_2, \epsilon_2\right) - \omega q_1\right) - c\left(q_1 + q_2\right)\Big\}.$$

After optimizing over q_2, we find that $\frac{\partial \pi}{\partial q_1} = p_1 - \omega p_2 - c(1-\omega) < 0$. Therefore, allowing parallel importation, but limiting its volume is suboptimal. $\square$

Proof of Proposition 2.4. The first part follows because the parallel importer buys q_G from the manufacturer in market 1. The change in the manufacturer's demand in market 2 is $N_2 - \frac{p_2 - p_G}{1-\omega} b_2 - (N_2 - b_2 p_2) = \frac{p_G - \omega p_2}{(1-\omega)} b_2$. Because $\frac{\omega p_2 - p_G}{\omega(1-\omega)} b_2 \leq q_1$, the change of demand is equal to $-\omega q_G$. $\square$

Proof of Theorem 2.1. To prove that $\widehat{p_1}$ is unique, define $h(p_1) = N_1 - 2b_1 p_1 + z_1\left(p_1\right) - \int_{L_1}^{z_1(p_1)} F_1(x) + cb_1$, $K(p_2) = N_2 - 2b_2 p_2 + z_2\left(p_2\right) - \int_{L_2}^{z_2(p_2)} F_2(x) + cb_2$, and $g\left(p_1, p_2\right) = \omega^2 h(p_1) + \omega K(p_2)$. Then $\widehat{p_1}$ solves $g\left(p_1, \frac{p_1 + c_G}{\omega}\right) = 0$. We note that

$$h'(p_1) = z_1'\left(p_1\right)\frac{c}{p_1} - 2b_1 = \frac{1}{c}\frac{\left[1 - F_1\left(z_1\left(p_1\right)\right)\right]}{r_1\left(z_1\left(p_1\right)\right)} - 2b_1, \qquad h''(p_1) = z_1''\left(p_1\right)\frac{c}{p_1} - \frac{c}{p_1^2} z_1'\left(p_1\right)$$

in which $z_1''\left(p_1\right) = \frac{-z_1'(p_1) f_1(z_1(p_1)) r_1(z_1(p_1)) - [1 - F_1(z_1(p_1))] z_1'(p_1) r_1'(z_1(p_1))}{c[r_1(z_1(p_1))]^2}$. Thus,

$$h''(p_1) \quad = \quad \frac{-z_1'\left(p_1\right)\left[1 - F_1\left(z_1\left(p_1\right)\right)\right]^2}{c\left[r_1\left(z_1\left(p_1\right)\right)\right]^2} \times \left(2\left[r_1\left(z_1\left(p_1\right)\right)\right]^2 + r_1'\left(z_1\left(p_1\right)\right)\right) < 0$$

and $h(p_1)$ is concave, which means that $K(p_2)$ and $g\left(p_1, \frac{p_1 + c_G}{\omega}\right)$ are also concave. We find $\widehat{p_1}$ when $\omega\widetilde{p_2} - \widetilde{p_1} - c_G > 0$. Because $\widetilde{p_1} > c$, we have $\frac{c + c_G}{\omega} < \widetilde{p_2}$ and $K\left(\frac{c + c_G}{\omega}\right) > 0$. Therefore, $g\left(c, \frac{c + c_G}{\omega}\right) > 0$. Because $g\left(p_1, \frac{p_1 + c_G}{\omega}\right) < 0$ when p_1 is very large, we conclude that $\widehat{p_1}$ is unique.

Now we prove parts (b) through (d). We consider two cases.

Case 1. $\omega p_2 - p_1 - c_G \leq 0$. In this case, $q_G = 0$ and the SSG can be written as

$$\max_{p_1,p_2,q_1,q_2} \pi \quad = \quad E\Big\{p_1 min\left(q_1, \mathcal{D}_1\left(p_1, \epsilon_1\right)\right) + p_2 min\left(q_2, \mathcal{D}_2\left(p_2, \epsilon_2\right)\right) - c\left(q_1 + q_2\right)\Big\}$$

$$s.t.$$

$$(\gamma) \quad \omega p_2 - p_1 - c_G \leq 0$$

in which $\gamma \geq 0$ is a nonnegative Lagrangian multiplier. For given prices p_1 and p_2, π is concave in q_1 and q_2. Thus $q_i = N_i - b_i p_i + z_i\left(p_i\right)$ for $i = 1, 2$. Replacing the quantities in the profit function, we can write the KKT conditions:

$$\frac{\partial \pi}{\partial p_1} = N_1 - 2b_1 p_1 + z_1\left(p_1\right) - \int_{L_1}^{z_1(p_1)} F_1(x) + cb_1 + \gamma = 0, \tag{2.10}$$

$$\frac{\partial \pi}{\partial p_2} = N_2 - 2b_2 p_2 + z_2\left(p_2\right) - \int_{L_2}^{z_2(p_2)} F_2(x) + cb_2 - \omega\gamma = 0, \tag{2.11}$$

$$\gamma\left(\omega p_2 - p_1 - c_G\right) = 0, \omega p_2 - p_1 - c_G \leq 0, \gamma \geq 0.$$

If $\omega p_2 - p_1 - c_G < 0$, then $\gamma = 0$ and (2.10) and (2.11) reduce to (2.2). On the other hand, if $\omega p_2 - p_1 - c_G = 0$, then (2.10) and (2.11) reduce to solving $g\left(p_1, \frac{p_1 + c_G}{\omega}\right) = 0$. If $\gamma = -N_1 + 2b_1\widehat{p_1} - z_1(\widehat{p_1}) + \int_{L_1}^{z_1(\widehat{p_1})} F_1(x) - cb_1 \leq 0$, then the manufacturer ignores the importer. However, if $\gamma > 0$, then $\left(\widehat{p_1}, \frac{\widehat{p_1} + c_G}{\omega}, \gamma\right)$ is a solution to the KKT conditions. Because π is strictly quasiconcave, if $\gamma > 0$, then $\widehat{p_1} > \widetilde{p_1}$ and $\frac{\widehat{p_1} + c_G}{\omega} < \widetilde{p_2}$, which means that $\omega\widetilde{p_2} - \widetilde{p_1} - c_G > 0$. Thus, if $\gamma > 0$, $\left(\widehat{p_1}, \frac{\widehat{p_1} + c_G}{\omega}, \gamma\right)$ is the only solution to the KKT conditions.

Case 2. $\omega p_2 - p_1 - c_G \geq 0$. In this case because the manufacturer is assumed to not leave market 1, $q_G = \psi$ and the SSG becomes

$$\begin{aligned} \max_{p_1,p_2,q_1,q_2} \pi \quad &= \quad E\Big\{p_1 min\left(q_1, \mathcal{D}_1\left(p_1, \epsilon_1\right) + \psi\right) + p_2 min\left(q_2, \mathcal{D}_2\left(p_2, \epsilon_2\right) - \omega\psi\right) \\ & \quad\quad -c\left(q_1 + q_2\right)\Big\} \\ & \quad s.t. \\ & \quad (\eta) \quad \omega p_2 - p_1 - c_G \geq 0 \end{aligned}$$

in which $\eta \geq 0$. Similar to Case 1, for a given p_1 and p_2 we have, $q_1 = N_1 - b_1 p_1 + z_1\left(p_1\right) + \psi$ and $q_2 = N_2 - b_2 p_2 + z_2\left(p_2\right) - \omega\psi$. The KKT conditions for this case are

$$\begin{aligned} &\frac{\partial \pi}{\partial p_1} = N_1 - 2b_1 p_1 + \frac{2\left(\omega p_2 - p_1\right) - c_G}{2\omega\left(1-\omega\right)} b_2 + c\left(b_1 + \frac{b_2}{2\omega}\right) + z_1\left(p_1\right) \\ &\quad - \int_{L_1}^{z_1(p_1)} F_1\left(x\right) - \eta = 0, \\ &\frac{\partial \pi}{\partial p_2} = N_2 - 2b_2 p_2 - \frac{2\left(\omega p_2 - p_1\right) - c_G}{2\left(1-\omega\right)} b_2 + c\frac{b_2}{2} + z_2\left(p_2\right) \\ &\quad - \int_{L_2}^{z_2(p_2)} F_2\left(x\right) + \omega\eta = 0, \\ &\quad \eta\left(\omega p_2 - p_1 - c_G\right) = 0,\ \omega p_2 - p_1 - c_G \geq 0,\ \eta \geq 0. \end{aligned}$$

Case 2.1. If $\omega p_2 - p_1 - c_G > 0$, then $\eta = 0$ and

$$\begin{aligned} &\frac{\partial \pi}{\partial p_1} = N_1 - 2b_1 p_1 + \frac{2\left(\omega p_2 - p_1\right) - c_G}{2\omega\left(1-\omega\right)} b_2 + c\left(b_1 + \frac{b_2}{2\omega}\right) + z_1\left(p_1\right) \\ &\quad - \int_{L_1}^{z_1(p_1)} F_1\left(x\right) = 0, \end{aligned} \tag{2.12}$$

$$\frac{\partial\pi}{\partial p_2} = N_2 - 2b_2p_2 - \frac{2\left(\omega p_2 - p_1\right) - c_G}{2\left(1-\omega\right)}b_2 + c\frac{b_2}{2} + z_2\left(p_2\right) - \int_{L_2}^{z_2(p_2)} F_2\left(x\right) = 0. \quad (2.13)$$

Case 2.2. If $\omega p_2 - p_1 - c_G = 0$, then

$$\begin{aligned}\frac{\partial\pi}{\partial p_1} &= N_1 - 2b_1p_1 + \frac{c_G}{2\omega\left(1-\omega\right)}b_2 + z_1\left(p_1\right) + c\left(b_1 + \frac{b_2}{2\omega}\right)\\ &- \int_{L_1}^{z_1(p_1)} F_1\left(x\right) - \eta = 0, \qquad (2.14)\\ \frac{\partial\pi}{\partial p_2} &= N_2 - 2b_2\left(\frac{p_1+c_G}{\omega}\right) - \frac{c_G}{2\left(1-\omega\right)}b_2 + z_2\left(\frac{p_1+c_G}{\omega}\right) + c\frac{b_2}{2}\\ &- \int_{L_2}^{z_2\left(\frac{p_1+c_G}{\omega}\right)} F_2\left(x\right) + \omega\eta = 0. \qquad (2.15)\end{aligned}$$

One can see that solving (2.14) and (2.15) is equivalent to solving $g\left(p_1, \frac{p_1+c_G}{\omega}\right) = 0$ the solution to which is $\widehat{p_1}$, similar to Case 1. Define

$$\eta \quad = \quad N_1 - 2b_1\widehat{p_1} + \frac{c_G}{2\omega\left(1-\omega\right)}b_2 + z_1\left(\widehat{p_1}\right) + c\left(b_1 + \frac{b_2}{2\omega}\right) - \int_{L_1}^{z_1(\widehat{p_1})} F_1\left(x\right).$$

If $\eta \leq 0$, the manufacturer should solve (2.12) and (2.13), and allow parallel importation. On the other hand, if $\eta > 0$, then $\left(\widehat{p_1}, \frac{\widehat{p_1}+c_G}{\omega}, \eta\right)$ satisfies the KKT conditions. To show that it is indeed the only solution to the KKT conditions, we show that the profit function for $\psi > 0$ is strictly quasiconcave in p_1 and p_2 (but not jointly). Suppose (2.12) and (2.13) have a feasible solution. Then

$$\begin{aligned} -b_1 - \frac{b_2}{\omega(1-\omega)} \quad &= \quad \frac{1}{(p_1-c)}\Bigg[-N_1 + b_1p_1 - \frac{2\omega p_2 - c_G - c\left(1+\omega\right)}{2\omega(1-\omega)}b_2\\ &+ \int_{L_1}^{z_1(p_1)} F_1\left(x\right) - z_1(p_1)\Bigg]. \qquad (2.16)\end{aligned}$$

First because $p_1 > c$, we have

$$\frac{\partial^2 \pi}{\partial p_1^2} = z_1'\left(p_1\right)\frac{c}{p_1} - 2b_1 - \frac{1}{\omega\left(1-\omega\right)}b_2 < \frac{c}{p_1}\left[z_1'\left(p_1\right) - b_1 - \frac{1}{\omega\left(1-\omega\right)}b_2\right]$$

in which $z_1'\left(p_1\right) = \frac{c}{p_1^2 f_1(z_1)}$. Using (2.16), whenever $\frac{\partial \pi}{\partial p_1} = 0$ we have

$$\begin{aligned}
\frac{\partial^2 \pi}{\partial p_1^2} &< \frac{c}{p_1}\left[\frac{\frac{c}{p_1^2}}{f_1(z_1(p_1))} + \frac{1}{(p_1-c)}\left(-N_1 + b_1 p_1 - \frac{2\omega p_2 - c_G - c(1+\omega)}{2\omega(1-\omega)}b_2\right.\right. \\
&+ \left.\left.\int_{L_1}^{z_1(p_1)} F_1\left(x\right) - z_1(p_1)\right)\right] \\
&= \frac{c}{p_1(p_1-c)}\left[\frac{\frac{c}{p_1^2}(p_1-c)}{f_1(z_1(p_1))} + \left(-N_1 + b_1 p_1 - \frac{2\omega p_2 - c_G - c(1+\omega)}{2\omega(1-\omega)}b_2\right.\right. \\
&+ \left.\left.\int_{L_1}^{z_1(p_1)} F_1\left(x\right) - z_1(p_1)\right)\right] \\
&= \frac{c}{p_1(p_1-c)}\left[\frac{F_1(z_1(p_1))}{r_1(z_1(p_1))} + \int_{L_1}^{z_1(p_1)} F_1\left(x\right) - z_1(p_1)\right. \\
&- \left.\left(N_1 - b_1 p_1 + \frac{2\omega p_2 - c_G - c(1+\omega)}{2\omega(1-\omega)}b_2\right)\right].
\end{aligned}$$

If $K(z_1\left(p_1\right)) = \frac{F_1(z_1(p_1))}{r_1(z_1(p_1))} + \int_{L_1}^{z_1(p_1)} F_1\left(x\right) - z_1\left(p_1\right)$, then we have $K'(z_1\left(p_1\right)) = \frac{-z_1'(p_1)F_1(z_1(p_1))r_1'(z_1(p_1))}{r_1(z_1(p_1))^2} < 0$ because $r_1'(z_1\left(p_1\right)) > 0$ and $z_1'\left(p_1\right) > 0$. Thus $K(z_1\left(p_1\right))$ is decreasing in $z_1\left(p_1\right)$. Given that $z_1(p_1) > L_1$, we get $K(z_1) < k(L_1) = -L_1$. Note that for any p_1 and p_2 that allow parallel importation, the minimum demand in market 1 should be positive; that is, $N_1 - b_1 p_1 + \psi + L_1 > 0$. Because $p_1, p_2 > c$, we have

$$N_1 - b_1 p_1 + L_1 + \frac{2\omega p_2 - c_G - c(1+\omega)}{2\omega(1-\omega)}b_2 > 0.$$

Therefore, $\left.\frac{\partial^2 \pi}{\partial p_1^2}\right|_{\frac{\partial \pi}{\partial p_1}=0} < 0$ and π is quasiconcave in p_1 for any given p_2. Because the minimum demand in market 2, $N_2 - b_2 p_2 - \omega\psi + L_2$, should be positive, we can show in a similar manner that π is quasiconcave in p_2 for any given p_1. Thus, if (p_1, p_2) solve (2.12) and (2.13), then because $\omega p_2 - p_1 - c_G > 0$, we can write

$$N_1 - 2b_1 p_1 + \frac{c_G}{2\omega\left(1-\omega\right)} b_2 + c\left(b_1 + \frac{b_2}{2\omega}\right) + z_1\left(p_1\right) - \int_{L_1}^{z_1(p_1)} F_1\left(x\right) < 0, \tag{2.17}$$

$$N_2 - 2b_2 p_2 - \frac{c_G}{2\left(1-\omega\right)} b_2 + c\frac{b_2}{2} + z_2\left(p_2\right) - \int_{L_2}^{z_2(p_2)} F_2\left(x\right) > 0. \tag{2.18}$$

Therefore, $\widehat{p_1} < p_1$. However, if this inequality holds, we must have

$$N_2 - 2b_2\left(\frac{p_1 + c_G}{\omega}\right) - \frac{c_G}{2\left(1-\omega\right)} b_2 + z_2\left(\frac{p_1 + c_G}{\omega}\right) + c\frac{b_2}{2} - \int_{L_2}^{z_2\left(\frac{p_1+c_G}{\omega}\right)} F_2\left(x\right) < 0.$$

Again because of quasiconcavity, $p_2 < \frac{p_1 + c_G}{\omega}$, which is a contradiction. Therefore, if $\eta > 0$, then (2.12) and (2.13) will not have a feasible solution. To complete the proof, note that if $\eta \leq 0$, then $\gamma > 0$ and the solution of Case 1 is forced to the boundary (block). Also if $\gamma \leq 0$, then $\eta > 0$ and the solution of Case 2 is forced to the boundary (again block). □

Proof of Corollary 2.3. Assume we solve for $\omega \longrightarrow 1$. Note that $\widehat{p_1} < \omega\widetilde{p_2} - c_G$ due to (2.2), and strict quasiconcavity of $\overline{\pi}$. If $c_G \neq 0$, then η will be positive when ω approaches one because $\frac{c_G}{2\omega(1-\omega)} b_2$ will be the dominant term. If $c_G = 0$, then the optimal solution will be to charge $\widehat{p_1}$ in both markets and η will be zero. Because the left hand side of (2.7) is a continuous function of ω, there exists an interval $[\omega_0, 1)$ in which the optimal policy is to block importation. □

Proof of Proposition 2.5. Consider $\widetilde{p_1}$ and $\widetilde{p_2}$ such that $\omega\widetilde{p_2} - \widetilde{p_1} - c_G > 0$. Note that $g\left(\widetilde{p_1}, \widetilde{p_2}\right) = 0$. First, suppose the solution to the SSG is to allow parallel imports.

Then using (2.17) we see that

$$N_1 - 2b_1 p_1^* + c\left(b_1 + \frac{b_2}{2\omega}\right) + z_1\left(p_1^*\right) - \int_{L_1}^{z_1\left(p_1^*\right)} F_1\left(x\right) < 0,$$

which means $p_1^* > \widetilde{p_1}$ because (3.1) is quasiconcave. Similarly, (2.18) gives us

$$N_2 - 2b_2 p_2^* + cb_2 + z_2\left(p_2^*\right) - \int_{L_2}^{z_2\left(p_2^*\right)} F_2\left(x\right) > 0,$$

so $p_2^* < \widetilde{p_2}$. Now assume that the SSG suggests blocking the importer. From (2.2), $\omega\widetilde{p_2} > \widetilde{p_1} + c_G$, and the quasiconcavity of the profit function, we get $g\left(\widetilde{p_1}, \frac{\widetilde{p_1}+c_G}{\omega}\right) > 0$. Therefore, $p_1^* = \widehat{p_1} > \widetilde{p_1}$ must hold. Finally, because $g\left(\widehat{p_1}, \widetilde{p_2}\right) < 0$, $p_2^* = \frac{\widehat{p_1}+c_G}{\omega}$ must be smaller than $\widetilde{p_2}$. □

Proof of Proposition 2.6. Suppose the DSG allows parallel importation. Then p_1^{*d} and p_2^{*d} solve

$$\frac{\partial \pi^d}{\partial p_1} = N_1 + \mu_1 - 2b_1 p_1^{*d} + \frac{2\left(\omega p_2^{*d} - p_1^{*d}\right) - c_G}{2\omega\left(1-\omega\right)} b_2 + c\left(b_1 + \frac{b_2}{2\omega}\right) = 0, \qquad (2.19)$$

$$\frac{\partial \pi^d}{\partial p_2} = N_2 + \mu_2 - 2b_2 p_2^{*d} - \frac{2\left(\omega p_2^{*d} - p_1^{*d}\right) - c_G}{2\left(1-\omega\right)} b_2 + c\frac{b_2}{2} = 0, \qquad (2.20)$$

and $\omega p_2^{*d} - p_1^{*d} - c_G > 0$. Note that

$$z_i(p) - \int_{L_i}^{z_i(p)} F_i\left(x\right) = \mu_i - \int_{z_i(p_i)}^{U_i} \left(x - z_i(p)\right) f_i\left(x\right). \qquad (2.21)$$

If the SSG solution is to block parallel imports, then (2.19) and (2.21) imply that $\eta < 0$ when $\widehat{p_1}$ is replaced with p_1^{*d}. Therefore $p_1^* = \widehat{p_1} < p_1^{*d}$ and $p_2^* = \frac{\widehat{p_1}+c_G}{\omega} < \frac{p_1^{*d}+c_G}{\omega} < p_2^{*d}$. If the SSG allows parallel imports, then $\frac{\partial \pi}{\partial p_i}\left(p_1^{*d}, p_2^{*d}\right) = \mu_i - \int_{z_i\left(p_i^{*d}\right)}^{U_i} \left(x - z_i\left(p_i^{*d}\right)\right) f_i\left(x\right) <$

0 for $i = 1, 2$, which means $g\left(p_1^{*d}, p_2^{*d}\right) < 0$. Now using the quasiconcavity property we have:

1. If $p_2^{*d} \leq p_2^*$, then $g\left(p_1^*, p_2^{*d}\right) \geq g\left(p_1^*, p_2^*\right) = 0$ and $\frac{\partial \pi}{\partial p_2}\left(p_1^*, p_2^{*d}\right) \geq 0$,
   (a) If $p_1^{*d} < p_1^*$, then $g\left(p_1^*, p_2^{*d}\right) < g\left(p_1^{*d}, p_2^{*d}\right) < 0$.
   (b) If $p_1^{*d} > p_1^*$, then $\frac{\partial \pi}{\partial p_2}\left(p_1^{*d}, p_2^{*d}\right) > \frac{\partial \pi}{\partial p_2}\left(p_1^*, p_2^{*d}\right) \geq 0$.
2. If $p_2^{*d} \geq p_2^*$ and $p_1^{*d} \leq p_1^*$, then $\frac{\partial \pi}{\partial p_1}\left(p_1^{*d}, p_2^{*d}\right) \geq \frac{\partial \pi}{\partial p_1}\left(p_1^{*d}, p_2^*\right) \geq \frac{\partial \pi}{\partial p_1}\left(p_1^*, p_2^*\right) = 0$.

All these cases result in a contradiction. Therefore, $p_1^* < p_1^{*d}$ and $p_2^* < p_2^{*d}$.

Now suppose the manufacturer blocks parallel imports in the DSG. Then p_1^{*d} solves

$$N_1 + \mu_1 - 2b_1 p_1^{*d} + \frac{c_G}{2\omega\left(1-\omega\right)} b_2 + c\left(b_1 + \frac{b_2}{2\omega}\right) - \lambda = 0, \tag{2.22}$$

$$N_2 + \mu_2 - 2b_2\left(\frac{p_1^{*d} + c_G}{\omega}\right) - \frac{c_G}{2\left(1-\omega\right)} b_2 + c\frac{b_2}{2} + \omega\lambda = 0, \tag{2.23}$$

and $p_2^{*d} = \frac{p_1^{*d} + c_G}{\omega}$ in which $\lambda \geq 0$ is the shadow price for $\omega p_2 - p_1 - c_G \geq 0$ in the DSG. First, consider the case when the SSG solution is to block parallel imports by $\left(\widehat{p}_1, \frac{\widehat{p}_1 + c_G}{\omega}\right)$. If we replace $\widehat{p}_1$ with p_1^{*d} and use equations (2.21) through (2.23), then

$$\begin{aligned} g\left(p_1^{*d}, \frac{p_1^{*d} + c_G}{\omega}\right) &= -\omega \int_{z_2\left(\frac{p_1^{*d} + c_G}{\omega}\right)}^{U_2} \left(x - z_2\left(\frac{p_1^{*d} + c_G}{\omega}\right)\right) f_2\left(x\right) \\ &- \omega^2 \int_{z_1\left(p_1^{*d}\right)}^{U_1} \left(x - z_1\left(p_1^{*d}\right)\right) f_1\left(x\right) < 0. \end{aligned}$$

Therefore, $\widehat{p}_1 < p_1^{*d}$ and $\widehat{p}_2 < p_2^{*d}$. Now if the SSG solution is to allow parallel imports, then

$$\frac{\partial \pi}{\partial p_1}\left(p_1^{*d}, p_2^{*d}\right) = N_1 - 2b_1 p_1^{*d} + \frac{c_G}{2\omega\left(1-\omega\right)} b_2 + c\left(b_1 + \frac{b_2}{2\omega}\right)$$

$$
\begin{aligned}
& + \quad z_1\left(p_1^{*d}\right) - \int_{L_1}^{z_1\left(p_1^{*d}\right)} F_1(x) < 0 \\
\frac{\partial \pi}{\partial p_2}\left(p_1^{*d}, p_2^{*d}\right) &= -\omega\lambda + z_2\left(\frac{p_1^{*d}+c_G}{\omega}\right) - \int_{L_2}^{z_2\left(\frac{p_1^{*d}+c_G}{\omega}\right)} F_2(x) < 0, \\
g\left(p_1^{*d}, p_2^{*d}\right) &< 0.
\end{aligned}
$$

The inequality in the first line comes from $\eta < 0$ and $\widehat{p}_1 < p_1^{*d}$. This situation is similar to the first part of the proof (when both problems allow parallel imports). Therefore, we conclude that $p_1^* < p_1^{*d}$ and $p_2^* < p_2^{*d}$. □

Proof of Proposition 2.7. When demand is deterministic and there are no parallel imports, the manufacturer sets $\widetilde{p_1}^d = \frac{N_1+b_1c}{2b_1}$, $\widetilde{p_2}^d = \frac{N_2+b_2c}{2b_2}$, $\widetilde{q_1}^d = \frac{N_1-b_1c}{2}$, and $\widetilde{q_2}^d = \frac{N_2-b_2c}{2}$. For the DSG, we have $q_1^{*d} = N_1 - b_1p_1^{*d} + q_G$ and $q_2^{*d} = N_2 - b_2p_2^{*d} - \omega q_G$. Suppose $q_G = 0$. Because $p_1^{*d} > \widetilde{p_1}^d$ and $p_2^{*d} \leq \widetilde{p_2}^d$, we have $q_1^{*d} < \widetilde{q_1}^d$ and $q_2^{*d} \geq \widetilde{q_2}^d$. If $q_G > 0$, then $p_1^{*d} = \frac{\omega[(2-\omega)N_1+N_2]-b_2c_G}{2(b_2+\omega(2-\omega)b_1)} + \frac{c}{2}$ and $p_2^{*d} = \frac{2\omega(1-\omega)b_1N_2+b_2(N_2+\omega N_1)+\omega b_1b_2c_G}{2b_2(b_2+\omega(2-\omega)b_1)} + \frac{c}{2}$, which give us $q_1^{*d} = \frac{N_1-b_1c}{2} - \frac{b_2c}{4\omega} - \frac{b_2c_G}{4\omega(1-\omega)} < \widetilde{q_1}^d$ and $q_2^{*d} = \frac{N_2-b_2c}{2} + \frac{b_2c}{4} + \frac{b_2c_G}{4(1-\omega)} > \widetilde{q_2}^d$.

2.8 Appendix B. Regions for the Optimal Policy for the DSG

Although the regions in Figure 2.7 are derived from numerical experiments, for the DSG they can be obtained analytically by comparing the profit for each policy. The next proposition extends Propositions 5 and 6 of A&Y by laying out the profit functions for different market bases. Note that due to deterministic demand, we can also derive the closed-form expressions for the case of blocking the parallel importer using quantity.

Proposition 2.8. *The manufacturer ignores the parallel importer if* $\frac{\omega(N_2+b_2c)}{2b_2} \leq \frac{N_1+b_1c+2b_1c_G}{2b_1}$*; otherwise, the manufacturer allows the importer, blocks the importer using price, or blocks the importer using quantity. The profit functions* π_a *(allow),* π_{bp} *(block using price), and* π_{bq} *(block using quantity) are provided below*

$$\pi_a = \frac{1}{8\omega(1-\omega)\alpha_2}\Bigg[2\omega(1-\omega)\left[2\omega N_1N_2 + N_2^2 + \omega(2-\omega)N_1^2\right] + \frac{4\omega^2(1-\omega)^2 b_1 N_2^2}{b_2}$$

$$+ b_2(b_2+\omega^2 b_1)c_G^2 + 2\alpha_2 b_2 c(1-\omega)c_G - 4\omega(1-\omega)\alpha_2(N_1+N_2)c$$

$$+ (1-\omega)(2\omega b_1 + (1+\omega)b_2)\alpha_2 c^2 - 4\omega(1-\omega)(b_2N_1 - \omega b_1 N_2)c_G\Bigg],$$

$$\pi_{bp} = \frac{1}{4\alpha_1}\Bigg[\omega N_1(\omega N_1 + 2N_2) + N_2^2 + c^2(\omega b_1 + b_2)^2 - 4b_1b_2c_G^2 + 4c_G(\omega b_1 N_2 - b_2 N_1)$$

$$- 4b_1b_2c(1-\omega)c_G + c\left[2\omega b_1((1-2\omega)N_2 - \omega N_1) - 2b_2(N_2 + (2-\omega)N_1)\right]\Bigg],$$

$$\pi_{bq} = \frac{(N_2 - b_2c)^2}{4b_2},$$

in which $\alpha_1 = \omega^2 b_1 + b_2$ *and* $\alpha_2 = b_2 + \omega(2-\omega)b_1$*. The optimal prices in the order of the profits are*

$$p_{1a}^{*d} = \frac{\omega\left[(2-\omega)N_1 + N_2\right] - b_2c_G}{2\alpha_2} + \frac{c}{2},$$

$$p_{2a}^{*d} = \frac{2\omega(1-\omega)b_1N_2 + b_2(N_2 + \omega N_1) + \omega b_1 b_2 c_G}{2b_2\alpha_2} + \frac{c}{2},$$

$$p_{1bp}^{*d} = \frac{\omega^2 N_1 + \omega N_2 + c\left(\omega^2 b_1 + \omega b_2\right) - 2b_2c_G}{2\alpha_1}, \quad p_{2bp}^{*d} = \frac{\omega N_1 + N_2 + c\alpha_1 + 2\omega b_1 c_G}{2\alpha_1},$$

$$p_{1bq}^{*d} \geq \max\left\{\frac{N_1}{b_1}, \omega p_2^{*d} - c_G\right\}, \qquad p_{2bq}^{*d} = \frac{N_2 + b_2c}{2b_2}.$$

We omit the details of obtaining the optimal profits and prices. The manufacturer ignores the parallel importer if $\omega\widetilde{p_2}^d \leq \widetilde{p_1}^d + c_G$, which means $\frac{\omega(N_2+b_2c)}{2b_2} \leq \frac{N_1+b_1c+2b_1c_G}{2b_1}$. Otherwise, she has to choose her policy by comparing the profit functions for allowing, or blocking using price or quantity. Note that abandoning market 1 is equivalent to choosing a price in market 1 that is large enough to make the demand zero and block the importer. That is why for the policy of blocking using quantity p_1^{*d} must be larger than $\frac{N_1}{b_1}$ and $\omega p_2^{*d} - c_G$. □

Chapter 3

Beyond Price Mechanisms: How Much Can Service Help Manage the Competition from Gray Markets?

3.1 Introduction

In Chapter 2, we focused on how a manufacturer should adjust her price and quantity decisions when she faces parallel importation in an uncertain environment. We showed that the manufacturer should reduce her price gap to cope with parallel importation. Although price techniques such as reducing the price gap, price matching, and uniform pricing can lower the pressure from gray markets, there is a limit on how much manufacturers can compromise on price. Changing the price can confuse consumers, especially previous buyers who bought the product at a higher price, damage manufacturer reputation, and lead to lower profit margins, hence exposing manufacturers to risks. As a result, it is essential that manufacturers use non-price mechanisms to control the diversion of their products into gray markets. One such mechanism we consider in this chapter is providing service, which plays an important role in boosting demand. Service refers to all efforts that a manufacturer exerts to increase product demand such as advertising, product illustration, in-store promotions, providing information to customers before and during the sale process, warranty, and after-sales support. The contribution of service to total profits has grown substantially in many industries. A recent study by Deloitte Research reports that after sales service contributes between 19% to 47% to manufacturer revenue across various industries (Kumar and Sailesh 2011). Providing service before, during, and after the sale enhances customer satisfaction significantly and encourages customers to return for shopping in the future, generating more revenue for manufacturers as a result.

The ubiquity and proliferation of gray markets has motivated companies to pay more attention to the role of service provision in persuading customers to buy products from authorized channels. Recently Mercedes–Benz reported that the percentage of Benzes sold in Thailand that are supplied by the gray market has risen from 13% in 2008 to 51% in 2011. In response to this rapid growth of the gray market, Mercedes-Benz cut the prices of seven models by between 2–5% and offered more leasing alternatives to induce demand. In addition, the company announced that it would no longer honor warranty to gray market vehicles to protect its brand image, unless owners of such vehicles pay a one-time fee and register the vehicle with official Mercedes-Benz dealers[*] (*Bangkok Post* 2011a and 2011b). Prior to Mercedes-Benz, BMW Thailand had decided to deny warranty and service to gray vehicles (*Bangkok*

[*] The company also added that the fee will be waived for gray market vehicles that had visited an authorized Mercedes-Benz service center in Thailand prior to August 30, 2011.

Post 2010).

Several years before Mercedes-Benz and BMW decided to change their service policies, Hyundai was facing a similar challenge in the Philippines. After the relationship between Hyundai and its authorized distributor in the Philippines discontinued, a large volume of Starex vans were imported from Korea by parallel importers, and many customers came to Hyundai with service requests. After several rounds of lowering prices which was immediately thwarted by the importers slashing their prices, Hyundai decided to offer a three-year 100,000-kilometer warranty, a service that the importers could not match (*Philstar.com*, 2003).

The goal of this chapter is to analyze the effectiveness of investing in and providing service in helping companies to contend against parallel importation. We expand the modeling framework in Chapter 2 to incorporate service competition between the manufacturer and the parallel importer. For tractability, however, we assume demand is deterministic. The manufacturer determines her price and the amount of service she offers in each market. If the prices are sufficiently different, the parallel importer chooses his quantity, resell price, and the level of service he offers to his customers to maximize his total profits. We address the following questions:

1. How should a manufacturer change her price and service decisions in the presence of the gray market? How do price and service decisions determine the manufacturer's policy against the gray market?
2. How valuable is service in counteracting the competition from the gray market?
3. What is the impact of consumers' perception of parallel imports relative to products sold by the authorized channel on the manufacturer's decisions?
4. How does parallel importer's free-riding change the manufacturer's decisions?

5. Should the manufacturer deny service to customers who buy parallel imports completely (as did Hyundai and BMW), or should she consider the option of selling service to such customers (as did Mercedes-Benz)?

This chapter is organized as follows. Section 3.2 reviews the relevant literature. In Section 3.3, we first describe the problem setting. Then, we formulate the Stackelberg game model, characterize the optimal decisions of the players, and describe the impact of service and parallel importation on decisions. In Section 3.4, we highlight the value of service for the manufacturer in achieving higher profits and controlling gray market activities. In Section 3.5, we explore the effect of free-riding on the manufacturer. Section 3.6 focuses on the manufacturer's service policy towards customers who buy the product from the parallel importer. We conclude the chapter in Section 3.7 with a summary of this chapter and future research directions.

3.2 Literature Review

The topic of this chapter is related to the literature on price and non-price competition and the literature on gray markets. The role of service as a non-price mechanism has been studied in previous research. Iyer (1998) analyzes coordination in a distribution channel when two retailers compete on price and service, and shows that the manufacturer might need to offer menu-based contracts, instead of a uniform contract, to induce retail differentiation. Tsay and Agrawal (2000) consider a manufacturer who sells a product to two retailers who compete on price and service. They show that the intensity of competition and the degree of cooperation between the retailers affect policy, total sales, and profitability, and also propose wholesale pricing mechanisms to coordinate the channel. Perdikaki et al. (2011) address the timing of service investments under demand uncertainty and competition and analyze the effects of demand variability, intensity of competition, and the service cost differential on optimal timing. Lu et al. (2011) study the importance of manufacturer service in interactions between two competing manufacturers and their common retailer under Stackelberg and vertical Nash supply chain, and find that consumers receive higher service level under vertical Nash. Other papers that have looked at non-price decisions include Jeuland and Shugan (1983), Perry and Porter (1990), Winter (1993), Banker et al. (1998), and Moorthy (1998). None of these papers consider potential gray market activities and their impact on price and service decisions.

Despite the crucial role of service in coping with the competition from gray markets, this mechanism has received little attention in the literature. Dutta et al. (1994) study the optimal policy towards retailers selling across their territories (bootlegging) and show the optimal policy is to tolerate some level of bootlegging. Although service is a decision variable in their paper, they use a transaction cost approach and focus on the enforcement policy and the deployment of the exclusive territory distribution system. Coughlan and Soberman (1998) consider two competing manufacturers who sell products through retailers. The retailers compete on price and service and there are two types of customers with different sensitivity to price and service. The authors assume that the products are sold to the gray market at the marginal cost so the gray market does not facilitate market expansion. They show that gray markets can still increase the profits for the manufacturers. However, when price-insensitive customers are highly service sensitive, the manufacturers will prevent gray markets and keep both price-sensitive and service-sensitive customers in the authorized channel. Chen (2002) uses a service variable in a simple duopoly model and suggests that authorized distributors should compete on services to reduce the negative effects of unfair competition from gray markets. Chen (2008) analyzes the effect of demand function on the profit of a decentralized manufacturer who faces parallel importation. The author shows that if increase in parallel imports sales (which is an exogenous variable) leads to lower profits for the manufacturer, the profit loss will be higher when the authorized retailer decides the service level.

The research questions and focus of this work differentiate it from the above-mentioned papers in the following directions: (1) we investigate the price and service decisions of a vertically-integrated manufacturer who faces a profit-maximizing parallel importer who competes with the manufacturer both on price and service. The manufacturer's decisions determine her policy against the importer; whether she should ignore, block or allow parallel importation; (2) we compare the price and service decisions of the manufacturer with the scenario in which there is no parallel importation and with the scenario in which there is a parallel importer but the manufacturer does not offer any service. These comparisons show the impact of parallel importation on the manufacturer's decisions as well as the impact of service investment on the manufacturer and the parallel importer; (3) we highlight the value of service by numerically exploring how much leverage the manufacturer will get if she provides service when she encounters parallel importation; and (4) we address the question of whether companies should deny service to buyers of gray goods completely, or they should consider charging such customers a fee for benefiting from manufacturer service.

3.3 Model and Analysis

Consider the model setting in Chapter 2. The manufacturer provides service s_i in market i. We assume that demand for the product in market $i = 1, 2$ is deterministic, linear, decreasing in price, and increasing in service in the form of $d_i = N_i - b_i p_i + \theta_i s_i$ in which $\theta_i > 0$ represent demand sensitivity to change in price and service, respectively. Table 3.1 lists the new notations used in this chapter.

The cost of providing service s_i is fixed and equal to $\lambda_i \frac{s_i^2}{2}$. This quadratic cost function is commonly used in the literature (Tsay and Agrawal 2000, Iyer 1998, Banker 1998, Moorthy 1988, Mussa and Rosen 1978) and suggests that the marginal cost of providing service is increasing. Service cost will be quadratic if service has a significant store-level inventory component. For other types of service, managers usually invest in the lowest-hanging fruit so that further increments in the service level become progressively more costly (Tsay and Agrawal, 2000). Although we assume a quadratic service cost function, our results continue to hold if $d_i = N_i - b_i p_i + \theta_i \sqrt[n]{f(s_i)}$ and the cost of service is $\lambda_i \frac{\sqrt[n]{f(s_i)^2}}{2}$ for any (increasing) function $f(s_i) > 0$ and $n > 1$. Therefore, our model can capture many cases of linear or concave service cost and demand function.

Table 3.1: Notations for price and service

Service Parameters	
θ_i	demand sensitivity to change in manufacturer's service
λ_i	manufacturer's service cost parameter
θ_G	demand sensitivity to change in parallel importer's service
λ_G	Parallel importer's service cost parameter
Manufacturer's optimal prices and services	
p_i^m, s_i^m	when there are no parallel imports
p_i^a, s_i^a	when parallel importation is allowed
p_i^b, s_i^b	when parallel importation is blocked
$\widehat{p_i}^a$	when parallel importation is allowed and no service is offered
$\widehat{p_i}^b$	when parallel importation is blocked and no service is offered
$\overline{p}_i$, $\overline{s}$, $\overline{p}_s$	when gray market customers are charged a fee for service
Parallel importer's decisions	
s_G	service

Assumption 3.1. *Customers who buy the product from the parallel importer do not*

benefit from the manufacturer's service.

In practice, many manufacturer services are exclusively offered to customers who buy the product from authorized distributors. For example, the vast majority of companies deny after-sales support and warranty, or do not offer promotions to customers who have bought their products from gray markets. In addition to Mercendes-Benz and Hyundai's warranty policies described earlier, companies such as Nikon (*Nikon Canada*, 2012), Pentax (*Pentax Canada*, 2012), Sigma (*Sigmaphoto.com*, 2012), and Ticino (*Ticino USA*, 2012) are a few examples of companies that deny warranty and after-sales support if the product is purchased from gray markets. Nevertheless, it is difficult to offer services such as advertising and in-store demonstration and assistance exclusively and these services inevitably increase the demand both for the authorized channel and the gray market channel, a phenomenon known as free riding. In Section 3.5, we relax Assumption 3.1 and numerically examine the effect of free-riding on the manufacturer's decisions.

The sequence of events in the price and service Stackelberg game is as follows:

1. The manufacturer announces the price of the product and the level of service in each market.

2. The parallel importer observes the price differential and decides whether he wants to transfer the product from the low-price market to the high-price market.

3. If the parallel importer decides to transfer the product, he chooses his order quantity, resell price in the high-price market, and service investment. When the importer enters the high-price market, the market is segmented and profits are earned.

3.3.1 No Parallel Importation

When there is no parallel importation, the manufacturer maximizes her total profit by solving

$$\max_{p_1,p_2,s_1,s_2} \sum_{i=1}^{2} \left[(p_i - c)\left(N_i - b_i p_i + \theta_i s_i\right) - \lambda_i \frac{s_i^2}{2} \right] \tag{3.1}$$

We make the following assumptions about the parameters of the model:

Assumption 3.2. *$N_i > b_i c$ and $\lambda_i > \theta_i^2/2b_i$ for $i = 1, 2$.*

The first inequality ensures that the market bases are large enough to offset the production cost. The second inequality imposes a lower bound on the service cost parameters such that providing service is sufficiently expensive. If this inequality does not hold, the manufacturer can achieve unbounded profits by offering abundant service, which is clearly unrealistic. This inequality also ensures the concavity of the profit function. The optimal solution to (3.1) is

$$p_i^m = \frac{\lambda_i N_i + c\left(\lambda b_i - \theta_i^2\right)}{2\lambda_i b_i - \theta_i^2}, \qquad s_i^m = \frac{\left(N_i - b_i c\right)\theta_i}{2\lambda_i b_i - \theta_i^2}, \qquad i = 1, 2. \tag{3.2}$$

3.3.2 Parallel Importer's Entry

After the manufacturer announces her prices and service levels, the parallel importer considers buying the product in the low-price market and reselling it in the high-price market. Let us assume, without loss of generality, that $p_2^m > p_1^m$, so that the direction of transfer will be from market 1 to market 2. The parallel importer transfers q_G units of the product to market 2 at the per unit cost c_G, and sells the units at price p_G. He also provides service s_G to his customers at cost $\lambda_G \frac{s_G^2}{2}$.

When the parallel importer resells the product in market 2, consumers have three choices; they can buy the product from the manufacturer, buy from the parallel importer, or do not buy the product at all. The first step of analysis is to model market segmentation and consumer purchase decision. We adopt and expand the market segmentation approach in A&Y. We assume that the parallel importer faces demand $d_G = N_2 - \frac{b_2}{\omega} p_G + \theta_G s_G$ in market 2. The parameter $0 < \omega < 1$ denotes consumers' lower perception of parallel imports relative to products sold by the manufacturer; if the manufacturer and the parallel importer sell the product at the same price and neither offers service, all customers prefer to buy the product from the manufacturer due to the manufacturer's higher reputation and the peace of mind that consumers get when they buy from the authorized channel. θ_G represents the sensitivity of the demand for parallel imports to changes in the parallel importer's service. We assume $\theta_G < \theta_2$ so that the manufacturer can attract more customers when she increases her service by one unit than does the parallel importer when he offers one more unit of service. This assumption, however, does not change our analytical results. In summary, d_G implies that if the parallel importer wants to achieve the same demand as the manufacturer, he has to sell the product below the manufacturer's price and/or offer more service than the manufacturer.

In a manner similar to Chapter 2, we model consumer purchase decision by interpreting d_2 and d_G as outcomes of consumer surplus functions. Specifically, if consumers' net surplus of buying from the manufacturer is $\psi + \frac{\theta_2 s_2}{N_2} - \frac{b_2 p_2}{N_2}$ where $0 \leq \psi \leq 1$ represents the heterogeneity in consumer taste, the total demand for the manufacturer will be $d_2 = N_2 - b_2 p_2 + \theta_2 s_2$ if we assume that the utility of not buying the product is zero. Similarly, we define $\omega \left(\psi + \frac{\theta_G s_G}{N_2} \right) - \frac{b_2 p_G}{N_2}$ as consumers' net surplus of buying the product from the parallel importer, so that the total demand for parallel imports is $d_G = N_2 - \frac{b_2}{\omega} p_G + \theta_G s_G$.

Using these two surplus functions, we can obtain the size of the three segments of market 2. If ψ_1 represents the taste of the consumer who is indifferent between buying from the manufacturer and buying from the parallel importer, then

$$\psi_1 + \frac{\theta_2 s_2}{N_2} - \frac{b_2 p_2}{N_2} = \omega\left(\psi_1 + \frac{\theta_G s_G}{N_2}\right) - \frac{b_2 p_G}{N_2}$$

Similarly, if ψ_2 represents the taste of the consumer who is indifferent between buying from the gray market and not buying the product, then

$$\omega\left(\psi_2 + \frac{\theta_G s_G}{N_2}\right) - \frac{b_2 p_G}{N_2} = 0$$

Therefore, the manufacturer's demand is $N_2\left(1-\psi_1\right) = N_2 - \frac{b_2(p_2-p_G)-\theta_2 s_2+\omega\theta_G s_G}{1-\omega}$ and the parallel importer's demand is $N_2\left(\psi_1-\psi_2\right) = \frac{b_2(\omega p_2-p_G)-\omega\theta_2 s_2+\omega\theta_G s_G}{\omega(1-\omega)}$. The parallel importer solves the following problem to maximize his profit

$$\max_{p_G} \quad \pi_G = (p_G - p_1 - c_G)\left(\frac{b_2\left(\omega p_2 - p_G\right) - \omega\theta_2 s_2 + \omega\theta_G s_G}{\omega\left(1-\omega\right)}\right) - \lambda_G \frac{s_G^2}{2} \tag{3.3}$$

The first order conditions give

$$\begin{aligned} p_G &= \frac{\lambda_G(1-\omega)[b_2(\omega p_2 + p_1 + c_G) - \omega\theta_2 s_2] - \omega\theta_G^2(p_1+c_G)}{2(1-\omega)\lambda_G b_2 - \omega\theta_G^2} \\ q_G &= max\left\{0, \frac{\lambda_G b_2[b_2\left(\omega p_2 - p_1 - c_G\right) - \omega\theta_2 s_2]}{\omega[2(1-\omega)\lambda_G b_2 - \omega\theta_G^2]}\right\} \qquad s_G = max\left\{0, \frac{\omega\theta_G}{\lambda_G b_2} q_G\right\} \end{aligned} \tag{3.4}$$

which is a valid solution if

$$2(1-\omega)\lambda_G b_2 - \omega\theta_G^2 > 0 \tag{3.5}$$

We can see that this condition is violated when ω is sufficiently close to 1 and the manufacturer is better off blocking parallel importation. The parallel importer's optimal decisions indicate that the gray market will emerge if

$$b_2 \left(\omega p_2 - p_1 - c_G\right) > \omega \theta_2 s_2 \tag{3.6}$$

In A&Y, the gray market emerges if $b_2 \left(\omega p_2 - p_1 - c_G\right) > 0$. Although the importer exploits the price gap to transfer the product, because consumers have a lower perception of parallel imports, $\omega p_2 - p_1$ needs to be sufficiently large to payoff the importer's cost. Clearly (3.6) is a more stringent condition and requires that $\omega p_2 - p_1$ be even larger to cover not only the transfer cost but also the market share the importer loses to the manufacturer due to her service provision.

The next proposition explains the impact of parallel importation on manufacturer's sales, which also appears in A&Y. All proofs are provided in Section 3.8.

Proposition 3.1. *When the parallel importer transfers the product from market 1 to market 2, the manufacturer's sales in market 1 increase by q_G and her sales in market 2 decrease by ωq_G.*

In light of Proposition 3.1, we formulate the Stackelberg game as

$$\begin{aligned} \max_{p_1, p_2, s_1, s_2} \pi \; &= \left(p_1 - c\right)\left(N_1 - b_1 p_1 + \theta_1 s_1 + q_G\right) \\ &+ \left(p_2 - c\right)\left(N_2 - b_2 p_2 + \theta_2 s_2 - \omega q_G\right) - \sum_{i=1}^{2} \lambda_i \frac{s_i^2}{2} \end{aligned} \tag{3.7}$$

The manufacturer's optimal price and service *decisions* result in four *policies* against the parallel importer:

1. **Ignore** the parallel importer and continue to use $\left(p_1^m, s_1^m, p_2^m, s_2^m\right)$

2. **Allow** parallel importation by choosing prices and services $\left(p_1^a, s_1^a, p_2^a, s_2^a\right)$ such

that $b_2\left(\omega p_2^a - p_1^a - c_G\right) > \omega\theta_2 s_2^a$.

3. **Block** parallel importation. The manufacturer can eliminate parallel importation in two ways:

   a) Stay in both markets and use prices and services $\left(p_1^b, s_1^b, p_2^b, s_2^b\right)$ such that $b_2\left(\omega p_2^b - p_1^b - c_G\right) = \omega\theta_2 s_2^b$.

   b) Leave market 1 and only operate in market 2 with price p_2^m and service s_2^m.

The next proposition characterizes the optimal decisions and policy of the manufacturer.

Proposition 3.2. *Suppose the manufacturer does not leave market 1. Then her optimal policy is to ignore parallel importation if* $b_2\left(\omega p_2^m - p_1^m - c_G\right) \leq \omega\theta_2 s_2^m$. *Otherwise, the manufacturer has to change her decisions. In specific, she should allow parallel importation if*

$$q_G^* = \frac{\lambda_G b_2}{\lambda_G b_2\left(2-\omega\right) - \omega\theta_G^2}\left[\frac{\omega N_2 - b_2(c+c_G)}{\omega} - \frac{\lambda_1 b_2 s_1^a}{\omega\theta_1} - \frac{\lambda_2 b_2 s_2^a}{\theta_2}\right] > 0 \tag{3.8}$$

where

$$s_1^a = \frac{\omega\left(N_1 - b_1 c\right)\left[(2-\omega)\lambda_G b_2 - \omega\theta_G^2\right] + \lambda_G b_2[\omega N_2 - b_2 c - b_2 c_G]}{\omega\left(2\lambda_1 b_1 - \theta_1^2\right)\left[(2-\omega)\lambda_G b_2 - \omega\theta_G^2\right] + 2\lambda_1\lambda_G b_2^2}\theta_1 \tag{3.9}$$

$$s_2^a = \frac{\left(N_2 - b_2 c\right)\left[(2-\omega)\lambda_G b_2 - \omega\theta_G^2\right] - \lambda_G b_2(\omega N_2 - b_2 c - b_2 c_G)}{\left(2\lambda_2 b_2 - \theta_2^2\right)\left[(2-\omega)\lambda_G b_2 - \omega\theta_G^2\right] - 2\lambda_2\lambda_G b_2^2\omega}\theta_2 \tag{3.10}$$

The optimal prices that allow parallel importation are

$$\begin{aligned} p_1^a &= c + \lambda_1\frac{s_1^a}{\theta_1}, \\ p_2^a &= c + \frac{1}{\left(2-\omega\right)\lambda_G b_2 - \omega\theta_G^2}\left[\lambda_G b_2\frac{\lambda_1 s_1^a}{\theta_1} + \left(2\lambda_G b_2\left(1-\omega\right) - \omega\theta_G^2\right)\frac{\lambda_2 s_2^a}{\theta_2}\right] \end{aligned} \tag{3.11}$$

If $q_G^* \leq 0$, then the manufacturer blocks parallel importation by setting the following prices and services:

$$
\begin{aligned}
p_1^b &= \frac{\omega^2 \lambda_2 b_2 \left[\lambda_1 N_1 + \left(\lambda_1 b_1 - \theta_1^2\right) c\right] - \lambda_1 b_2 \left(2\lambda_2 b_2 - \theta_2^2\right) c_G}{\gamma} \\
&+ \frac{\omega \lambda_1 \left[\left(\lambda_2 b_2 - \theta_2^2\right) N_2 + \lambda_2 b_2^2 c\right]}{\gamma} \\
p_2^b &= \frac{\omega^2 \theta_2^2 \left(2\lambda_1 b_1 - \theta_1^2\right) N_2 + \omega b_2 \left(\lambda_2 b_2 - \theta_2^2\right) \left[\lambda_1 N_1 + (\lambda_1 b_1 - \theta_1^2) c + \left(2\lambda_1 b_1 - \theta_1^2\right) c_G\right]}{b_2 \gamma} \\
&+ \frac{\lambda_1 b_2^2 \left[\lambda_2 N_2 + \left(\lambda_2 b_2 - \theta_2^2\right) c\right]}{b_2 \gamma} \\
s_1^b &= \frac{\theta_1}{\gamma} \left[\omega^2 \lambda_2 b_2 \left(N_1 - b_1 c\right) + \omega \left[\left(\lambda_2 b_2 - \theta_2^2\right) N_2 + \lambda_2 b_2^2 c\right] - b_2 \left(2\lambda_2 b_2 - \theta_2^2\right) \left(c + c_G\right)\right] \\
s_2^b &= \frac{\theta_2}{\gamma} \Bigg[-\omega^2 N_2 \left(2\lambda_1 b_1 - \theta_1^2\right) + \omega b_2 \left[\lambda_1 N_1 + (\lambda_1 b_1 - \theta_1^2) c + \left(2\lambda_1 b_1 - \theta_1^2\right) c_G\right] \\
&- \lambda_1 b_2 (N_2 - b_2 c)\Bigg]
\end{aligned}
\tag{3.12}
$$

where $\gamma = b_2 \left[\lambda_1 \left(2\lambda_2 b_2 - \theta_2^2\right) + \lambda_2 \omega^2 \left(2\lambda_1 b_1 - \theta_1^2\right)\right]$.

The manufacturer does not need to change her price and service decisions if they automatically eliminate parallel importation. This situation arises when ω is very low, i.e., consumers are reluctant to buy the product from the gray market, or when the cost of transferring the product is high. When $(p_1^m, s_1^m, p_2^m, s_2^m)$ no longer eliminates the gray market, the manufacturer has to deviate from her decisions. Although the deviation causes loss of profit for the manufacturer, she may be better off allowing gray market activities and sharing market 2 with the parallel importer.

In Proposition 3.2, it is assumed that the manufacturer is better off operating in both markets. The next corollary provides a necessary condition for when it is better to leave market 1.

Corollary 3.1. *If the optimal policy for the manufacturer is to eliminate parallel importation by leaving market 1, then* $N_1 - b_1 p_1' + \theta_1 s_1' \leq 0$ *must hold where* $p_1' = \omega p_2^m - \frac{\omega \theta_2 s_2^m}{b_2} - c_G$ *and* $s_1' = \frac{(p_1' - c)\theta_1}{\lambda_1}$.

If the optimal policy is to only operate in market 2, then the value of p_2^m and s_2^m must be such that p_1' and s_1' do not generate any demand in market 1. If $N_1 - b_1 p_1' + \theta_1 s_1' > 0$, then the manufacturer can increase her profit by setting (p_1', s_1') in market 1 and setting (p_2^m, s_2^m) in market 2 while blocking the parallel importer.

Proposition 3.3. *The following inequalities are necessary conditions for the parallel importer's entry to the competition:*

i) $\omega N_2 - b_2(c + c_G) > 0$

ii) $\theta_1 < \sqrt{2\lambda_1 b_1 - \frac{2\lambda_1 b_2 (N_1 - b_1 c)}{\omega N_2 - b_2 c - b_2 c_G}}$

iii) $\theta_2 < \sqrt{min\Big\{\lambda_2 b_2, 2\lambda_2 b_2 \left(\frac{2(1-\omega)\lambda_G b_2 - \omega \theta_G^2}{(2-\omega)\lambda_G b_2 - \omega \theta_G^2}\right)\Big\}}$.

The minimum cost that the parallel importer incurs for selling the product is $c + c_G$ when the manufacturer sells the product at cost in market 1 and the parallel importer sells the product at cost in market 2. The first necessary condition requires that market 2 be large enough to payoff the minimum cost; otherwise, transferring the product will not be profitable for the importer.

The second and third necessary conditions are more stringent than the second part of Assumption 3.2 and highlight the value of service in coping with gray markets. In order for the parallel importer to survive in the competition, the range of consumer sensitivity to manufacturer service in both markets must be smaller than the ranges defined by Assumption 3.2. We note that when ω grows sufficiently large, the effective threshold for θ_2 becomes $\frac{2(1-\omega)\lambda_G b_2-\omega\theta_G^2}{(2-\omega)\lambda_G b_2-\omega\theta_G^2}$, which is decreasing in ω and negative when $\omega = 1$. As ω increases and consumers become almost indifferent between the manufacturer and the parallel importer, the role of service becomes more crucial; service helps the manufacturer differentiate herself from the parallel importer. Therefore, in addition to the price mechanism and the first mover advantage, service brings the manufacturer additional leverage in managing gray market activities, even if the parallel importer is capable of competing with the manufacturer on service. Finally, we note that rearranging conditions (b) and (c) results in lower bounds on service cost parameters λ_i which are greater than $\theta_i^2/2b_i$. Thus, the parallel importer will be able to resell the product if it is more costly for the manufacturer to offer service than when there are no parallel imports.

The next proposition explains how the presence of gray markets and investing in service impact the manufacturer's decisions. First, we define $\widehat{p_1}, \widehat{p_2}$ to be the optimal prices of the Stackelberg game when service is not offered, which solve

$$\begin{aligned} \max_{p_1,p_2} \pi &= (p_1 - c)(N_1 - b_1 p_1 + q_G) + (p_2 - c)(N_2 - b_2 p_2 - \omega q_G) \\ q_G &= max\left\{0, \frac{b_2(\omega p_2 - p_1 - c_G)}{2\omega(1-\omega)}\right\} \end{aligned} \tag{3.13}$$

Proposition 3.4. *(a) Whether the manufacturer allows or blocks parallel importation, she increases the price in market 1, reduces the price in market 2, and increases the service in both markets, i.e.,* $p_1^a, p_1^b > p_1^m$, $p_2^a, p_2^b < p_2^m$, $s_1^a, s_1^b > s_1^m$, *and* $s_2^a, s_2^b > s_2^m$. *(b) The optimal prices when parallel importation is allowed (blocked) are larger than the optimal prices when parallel importation is allowed (blocked) and no service is* *offered, i.e.,* $p_i^a > \widehat{p_i}^a$ *and* $p_i^b > \widehat{p_i}^b$.

Part (a) describes the impact of the gray market on manufacturer's price and service decisions when p_i^m and s_i^m can no longer eliminate the gray market. The manufacturer increases the price in market 1 and reduces the price in market 2. The direction of change in prices is same as the direction observed in previous research when a manufacturer only uses the price mechanism. Second, the manufacturer increases her service *both* in market 1 and market 2. The manufacturer increases the service in market 2 to counteract the competition. Moreover, she increases her service level in market 1. The reason is that increasing p_1 curbs the competition from the parallel importer, but at the same time hurts the customers of the authorized channel. Consequently, the manufacturer offers more service in market 1 to compensate for the effect of raising p_1.

Part (b) explains how introducing service impacts the optimal prices in the presence of gray markets. The prices will be higher when service is offered if the optimal policy is the same when service is offered and when service is not offered (both scenarios allow or block parallel imports). If, however, the optimal policies are different, then the statement need not hold. For example, suppose $N_1 = N_2 = 10,000$, $b_1 = 22$, $b_2 = 10$, $\theta_1 = 2$, $\theta_2 = 3$, $\lambda_1 = \lambda_2 = 10$, $c = 100$, $c_G = 5$, and $\theta_G = 0$. Then $\widehat{p_1}^a = 301.931$, $\widehat{p_2}^a = 517.451$, and $\widehat{q}_G > 0$, whereas $q_G^* = 0$ and $p_1^b = 297.221 < \widehat{p_1}^a$.

For the rest of the analysis, we focus on the more interesting case of $q_G^* > 0$, because in this scenario the parallel importer transfers the product and initiates a price and service competition with the manufacturer. The next proposition describes the impact of variations in consumers' sensitivity to service when the manufacturer tolerates parallel importation.

Proposition 3.5. *(a)* p_1^a, s_1^a, *and* s_2^a *are increasing in* θ_G, *whereas* p_2^a *is decreasing* *in θ_G.*

(b) The price gap $p_2^a - p_1^a$ is increasing convex in θ_2. It is increasing convex in θ_1 if $\theta_G^2 > (1-\omega)\lambda_G b_2/\omega$ and it is decreasing concave otherwise.

(c) If service parameters are symmetric; i.e., $\lambda_1 = \lambda_2 = \lambda$ and $\theta_1 = \theta_2 = \theta$, then $p_2^a - p_1^a$ and $s_2^a - s_1^a$ are increasing and convex in θ. Moreover, if $\theta_G = 0$, then the price gap will be larger than the price gap when no service is offered, i.e., $p_2^a - p_1^a > \widehat{p_2}^a - \widehat{p_1}^a$.

(d) s_1^a and s_2^a are increasing in ω.

Part (a) of Proposition 3.5 shows that when consumer sensitivity to parallel importer service increases, the manufacturer reduces her price gap and offers more service in market 2 to protect her market share. She also increases her service in market 1 to compensate for the increase in p_1^a.

Part (b) explains an interesting observation. The price gap is increasing and convex in service sensitivity in market 2, because service in market 2 reduces the parallel importer's market share directly and allows the manufacturer to increase p_2^a. On the other hand, the price gap can increase or decrease with sensitivity to service in market 1, depending on consumer sensitivity to parallel importer service, θ_G. One implication of part (b) is that even though service allows the manufacturer to charge higher prices according to part (c) of Proposition 3.4, the price gap may actually be lower than the price gap when service is not offered.

Part (c) shows that if the markets have symmetric service valuations and cost parameters, then the price gap and the service gap in the allow region will be increasing and convex in θ, and if θ_G, the price gap will be always higher than when no services are offered. The condition $\theta_G = 0$ is equivalent to the parallel importer not offering service. When service valuations are symmetric, the effect of higher service valuation on counteracting the importer in market 2, which leads to higher p_2^a, is stronger than the increase in p_1^a. In addition, the service gap also increases with θ. Therefore, parts (a) and (b) indicate that asymmetry between consumer sensitivity to manufacturer service in market 1 and market 2 has an impact on price gap.

Finally, part (d) shows that both service levels are increasing in ω. When the relative perception of parallel imports grows, consumers become almost indifferent between the vendors and the parallel importer becomes a strong competitor. This situation arises for example when the product is a commodity that has passed the maturity phase of its life cycle and consumers are quite familiar with it. For such products, in addition to reducing the price gap, the manufacturer offers more service in market 2 to differentiate herself from the parallel importer and protect her market share. She also offers more service in market 1 to counterbalance the increase in p_1. This result underlines the important role of service in dominating the competition with gray markets.

We can see from (3.3) and (3.4) that the optimal profit of the parallel importer is $\pi_G^* = \frac{\lambda_G[b_2\left(\omega p_2^a - p_1^a - c_G\right) - \omega\theta_2 s_2^a]^2}{2\omega\left(2\lambda_G b_2(1-\omega) - \omega\theta_G^2\right)}$. The denominator of π_G^* decreases with θ_G, and based on Proposition 3.5, the numerator also decreases with θ_G. In our extensive numerical experiments, we observed an interesting behavior. Although for *given* manufacturer's decisions, p_1, p_2, and s_2, the parallel importer is better off providing service to his customers if $2(1-\omega)\lambda_G b_2 - \omega\theta_G^2 > 0$, π_G^* is monotone decreasing in θ_G, meaning that the parallel importer's profit actually goes down in equilibrium when he competes with the manufacturer on service. Also, using the Envelope Theorem we see that the manufacturer's optimal profit, π^*, is decreasing in θ_G. Therefore, the equilibrium of the Stackelberg game is a Prisoner's Dilemma: when the parallel importer offers service, the profits are reduced compared to when the parallel importer avoids service competition. This interesting observation is a consequence of the players selfish actions. The parallel importer invests in service to achieve higher profit. The manufacturer anticipates the service competition from the parallel importer so she reduces her price gap and provides more service in market 2. The net effect of the manufacturer's decision is lower profit for the parallel importer. The manufacturer also loses profit, because she has to compromise further on her price discrimination and invest more in service.

Figure 3.1 is an example of our experiments in which π_G^* is plotted against θ_G for $\omega = 0.7$ and $\omega = 0.85$. The parameters of this experiment are $N_1 = N_2 = 10,000, b_1 = 24, b_2 = 10, c = 100$, and $c_G = 5$. These values of N_1, N_2, b_1, b_2, c, and c_G are used in other experiments as well. These parameters are chosen such that market 1 consumers are more price sensitive, whereas market 2 consumers are willing to pay a higher price. Service parameters are $\theta_1 = 2$, $\theta_2 = 4$, $\lambda_1 = 10$, $\lambda_2 = 20$, and $\lambda_G = 30$. We have chosen service parameters such that consumers in market 2 value service more than consumers in market 1, and the cost of providing service is higher in market 2. All the parameters together represent market 2 as a more developed country in which consumers demand more service, but the cost of service is also higher due to higher wages.

3.4 Value of Service

Having looked at the impact of service provision and parallel importation on the manufacturer's decisions, we now explore the value of service in managing gray market activities. We measure the additional profit the manufacturer gains by providing service to her customers when she is facing the gray market. For this purpose, we use a numerical experiment with two sets of service parameters: (1) $\theta_1 = 4, \theta_2 = 5$, $\lambda_1 = 10$, $\lambda_2 = 24$; and (2) $\theta_1 = 4, \theta_2 = 6$, $\lambda_1 = 10$, $\lambda_2 = 17$. With these values, the

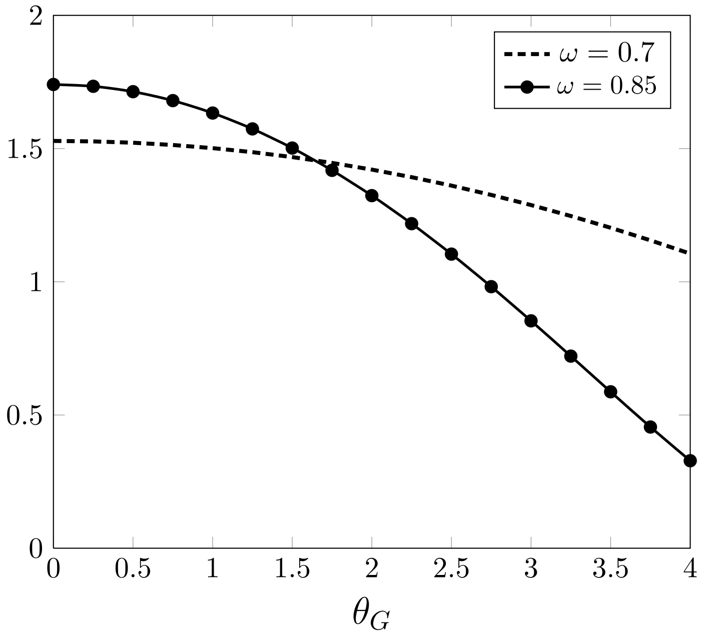


Figure 3.1: Effect of service competition: parallel importer's profit (in 1,000)

manufacturer will be able to increase her profit by 5% and 10%, respectively, if she invests in service when there is no parallel importation. The goal is to see how much service increases the manufacturer's profit *in the presence* of parallel importation.

Figure 3.2 shows the ratio of the manufacturer's optimal profit when she offers service to control parallel importation to her optimal profit when she only uses the price mechanism to cope with the gray market. We observe that the additional profits are constant and equal to 5% and 10% when ω is very low, because parallel importation can be ignored both with and without service. As ω increases, the ratio of profits increases with ω, indicating that service significantly helps the manufacturer boost her profits when the competition intensifies. Note that the manufacturer's profit is decreasing in ω both when she offers service and when she does not offer service. However, the ratio of profits is increasing in ω since service helps the manufacturer reduce the loss of profit to the parallel importer.

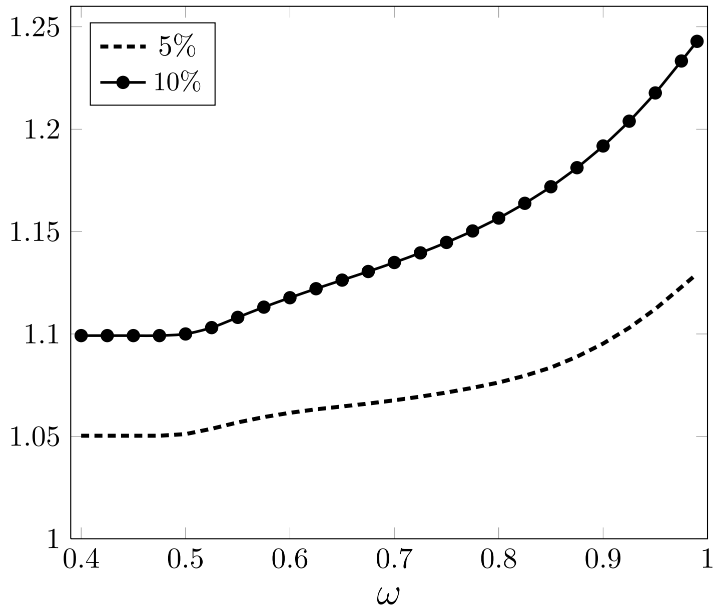
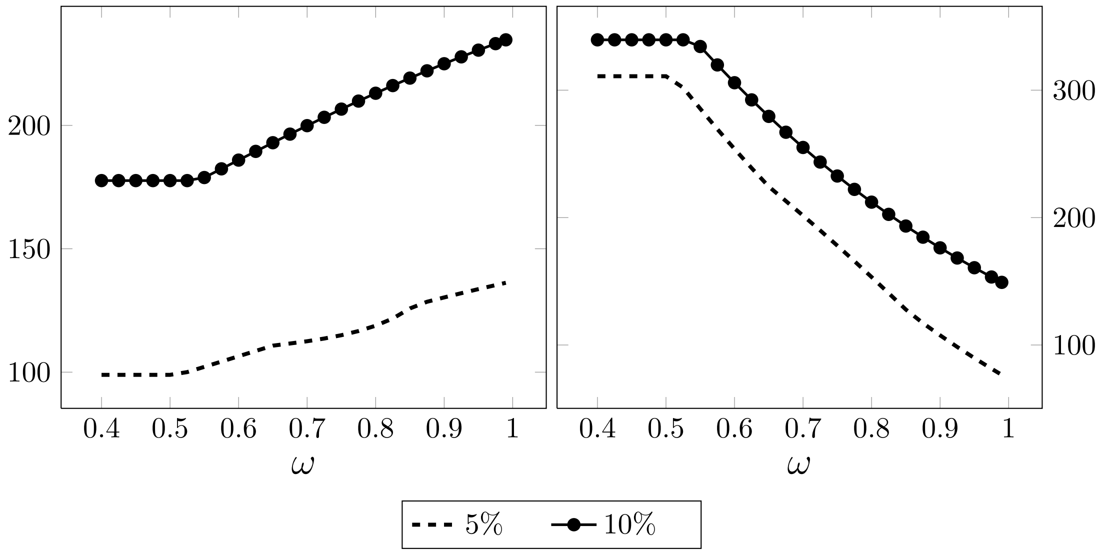


Figure 3.2: Ratio of profit when service is offered to profit when service is not offered

Figure 3.2 clearly shows that a little service can go a long way. Even if service increases the manufacturer's profit by only 5% in the absence of gray markets, the manufacturer can increase her profit by as much as 12% parallel importation exists. Likewise, a 10% increase in profit in the absence of gray markets can bring as much as 25% higher profits to the manufacturer in the presence of parallel importation.

Figure 3.3 shows the level of service offered in market 2 as well as the price gap for the same parameters as in Figure 3.2. When ω grows and the parallel importer becomes a strong competitor, the manufacturer reduces the price gap by increasing the price in market 1 and reducing the price in market 2. At the same time, she offers more service in market 2 to attract more consumers and counteract the parallel importer's price competition.



(a) service in market 2

(b) price gap

Figure 3.3: Manufacturer's service in market 2 and price gap

3.5 Effect of Free-riding

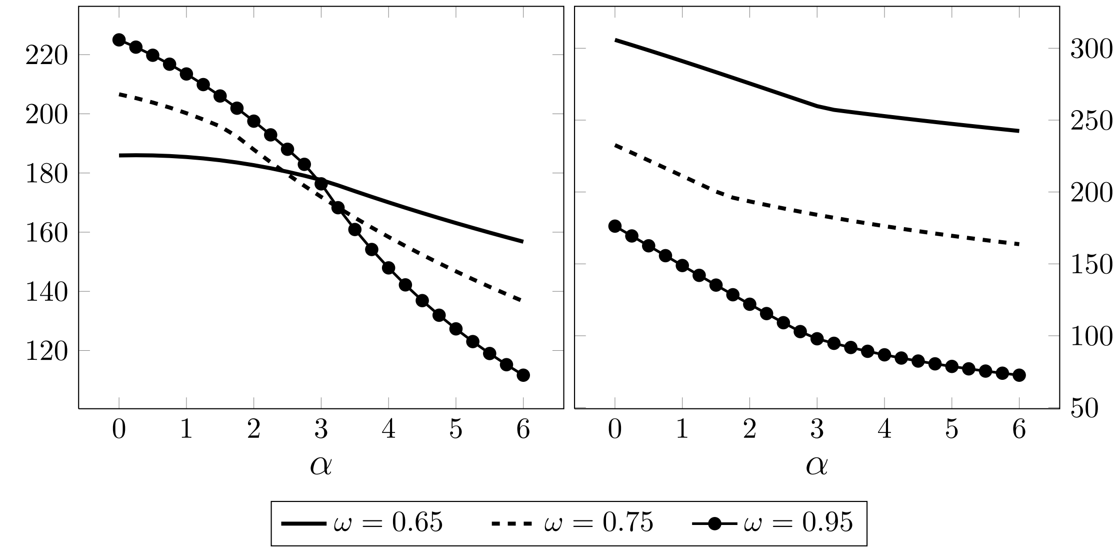
Our analysis so far has been based on the assumption that services offered by the manufacturer do not increase the demand for parallel imports. In reality, gray marketers free-ride on services provided by manufacturers or their authorized distributors. While price differential is the primary driver of gray market activities, free-riding is another reason gray markets emerge. For example, in the fashion industry the cost of marketing fragrances is frequently over 30% of the selling price, whereas transportation costs are usually below 10% (Gallini and Hollins, 2000). For such products, free-riding helps gray marketers keep the total cost of selling the product low and achieve a higher profit margin.

In this section, we relax Assumption 3.1 and assume that some of the services provided by the manufacturer have positive externality and increase the demand for parallel imports. We assume one unit of service increases the demand for parallel imports by $\alpha > 0$. Because there always exist services that can be exclusively offered to customers of authorized channels, we assume that $\alpha < \theta_2$. Thus, the new demand function for parallel imports is $N_2 - \frac{b_2}{\omega} p_G + \alpha s_2$. Revising the market segmentation analysis in Section 3.3, the parallel importer's quantity and price in (3.4) change to

$$\begin{aligned} p_G &= \frac{p_1 + c_G + \omega p_2}{2} - \frac{\omega \left(\theta_2 - \alpha\right) s_2}{2b_2} \\ q_G &= max \left\{ 0, \frac{b_2 \left(\omega p_2 - p_1 - c_G\right) - \omega \left(\theta_2 - \alpha\right) s_2}{2\omega \left(1 - \omega\right)} \right\} \end{aligned} \tag{3.14}$$

Figures 3.4 shows the effect of free-riding on the manufacturer's service investment in market 2 and her price gap. Figure 3.4(a) corroborates the argument made by opponents of gray markets that free-riding of gray marketers discourages authorized channels from investing in service (e.g., Gallini and Hollins, 2000). We also observe that unlike Figure 3.3(a), service in market 2 is not monotone in ω when there is free-riding. When free-riding is low, the manufacturer continues to provide more service as the gray market develops. However, when the parallel importer can free-ride on a larger portion of services, the manufacturer retreats from her strategy and reduces her investment in service as ω increases, because providing more service in this situation helps the gray market grow larger. Figure 3.4(b) shows that free-riding compels the manufacturer to reduce her price gap further to protect her market share. Therefore, free-riding undermines the effectiveness of the service mechanism and intensifies the price competition.

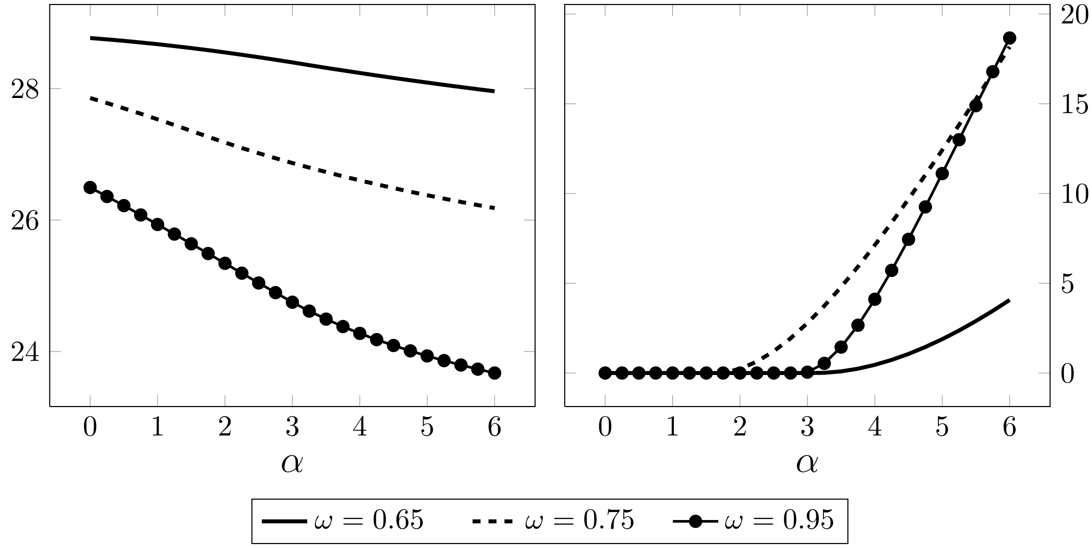
Figure 3.5 shows the effect of free-riding on profits. The manufacturer loses profit because she invests less in service and has to compromise on her prices. Free-riding, however, enables the parallel importer to charge a higher price and transfer a larger quantity, hence higher profits. We observe that the parallel importer's profit is in-



(a) service in market 2

(b) price gap

Figure 3.4: Effect of free-riding on the manufacturer's service in market 2 and price gap ($\theta_1 = 4, \theta_2 = 6$, $\lambda_1 = 10$, $\lambda_2 = 17$)



(a) manufacturer's profit (in 100,000)

(b) parallel importer's profit (in 1,000)

Figure 3.5: Effect of free-riding on profits (same parameters as Figure 3.4)

creasing convex in the level of free-riding. Even a free-riding level of $\alpha/\theta_2 = 50\%$, helps the parallel importer establish the gray market.

3.6 To Sell or Not to Sell Service

In this section, we focus on the manufacturer's service policy towards parallel imports. It seems that companies have taken two rather extreme positions on this issue. While companies such as Hyundai, BMW, and Nikon completely deny warranty and after-sales service to customers who purchase parallel imports, companies such as Toyota Lexus, Porsche, and Nintendo (*Bangkok Post*, 2010; *Nintendo Co.*, 2012) accept service requests from customers who have acquired their products from gray markets. There are a few companies, however, that have adopted an intermediate policy: they charge owners of parallel imports for services that are provided for free to customers of authorized channels. We mentioned earlier that Mercedes-Benz services imported vehicles in Thailand if the owner pays a one-time fee. Also, Porsche and Toyota Lexus reserve the right to charge gray-market car owners more for some services.

These examples have motivated us to look at the policy of offering service to gray market customers for a fee. When should the manufacturer in our problem consider selling service to customers who purchase parallel imports? What should be the service fee and the level of service offered to these customers? To answer these questions, we use an stylized extension of our model. For tractability, we drop service variables s_1 and s_2, and only focus on the amount of service and the service fee for parallel imports. Let $\overline{p}_1$, $\overline{p}_2$, $\overline{p}_s$, and $\overline{s}$, be the optimal price and service decisions when the manufacturer offers service s at price p_s to owners of parallel imports.

We need to model the number of customers who buy the product from the gray market and later pay the manufacturer for service. Because the volume of parallel imports is q_G, we define the demand for manufacturer service as $q_G - \nu_1 p_s + \nu_2 s$ with ν_1 and ν_2 representing the responsiveness of parallel import owners to changes in the level and price of the service offered. This service demand function, which has been used in the extended warranty literature (e.g., Li et al., 2011), assumes that the parallel import demand is not influenced by the service demand. This assumption holds in scenarios in which the manufacturer announces its service policy for parallel imports after some quantity of the product is resold in the gray market. For example, Mercedes-Benz, Hyundai, and BMW decided their policy for servicing gray products after the gray market had been established. We assume $\lambda \nu_1 > \nu_2^2$ so that no more than q_G customers will buy service in equilibrium. The Stackelberg game for selling service can be formulated as:

$$\begin{aligned} \max_{p_1, p_2, p_s, s} \quad & (p_1 - c)(N_1 - b_1 p_1 + q_G) + (p_2 - c)(N_2 - b_2 p_2 - \omega q_G) \\ & + p_s(q_G - \nu_1 p_s + \nu_2 s) - \lambda_2 \frac{s^2}{2} \end{aligned} \tag{3.15}$$

Remember that $\widehat{p}_1^a$ and $\widehat{p}_2^a$ denote the optimal prices in the absence of any services, and define $\widehat{q}_G = max \left\{ 0, \frac{\omega \widehat{p}_2^a - \widehat{p}_1^a - c_G}{2\omega(1-\omega)} b_2 \right\}$. The next proposition determines when, how much, and at what price service should be offered to gray market customers.

Proposition 3.6. *Suppose the manufacturer's optimal policy when she does not sell service to gray market customers is to allow parallel importation, i.e., $\widehat{q}_G > 0$. Then, the manufacturer can earn higher profits by allowing more parallel imports and selling service to gray market customers if and only if the following prices, level of service,* *and service fee are positive:*

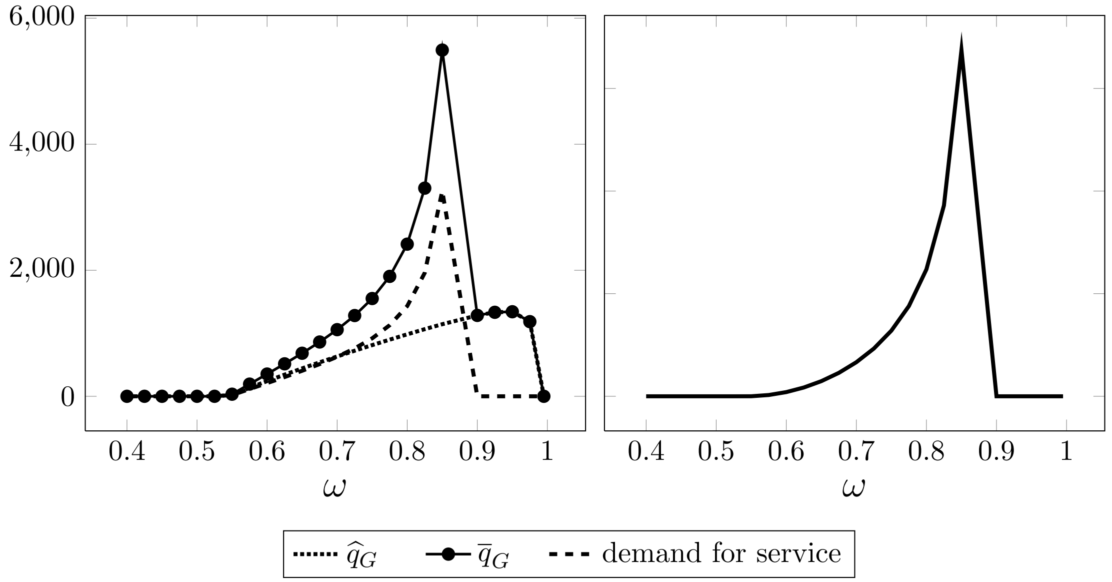
$$\begin{aligned}
\overline{p}_1 &= \widehat{p}_1^a - \frac{\lambda_2 b_2}{2\left(2\lambda_2\nu_1 - \nu_2^2\right)\left[b_2 + \omega(2-\omega)b_1\right]}\overline{q}_G, \\
\overline{p}_2 &= \widehat{p}_2^a + \frac{\lambda_2 b_1 \omega}{2\left(2\lambda_2\nu_1 - \nu_2^2\right)\left[b_2 + \omega(2-\omega)b_1\right]}\overline{q}_G, \\
\overline{s} &= \frac{\nu_2}{2\lambda_2\nu_1 - \nu_2^2}\overline{q}_G, \\
\overline{p}_s &= \frac{\lambda_2}{2\lambda_2\nu_1 - \nu_2^2}\overline{q}_G.
\end{aligned} \tag{3.16}$$

In this case, the volume of the gray market will be

$$\overline{q}_G = \frac{4\omega(1-\omega)\left(2\lambda_2\nu_1 - \nu_2^2\right)\left[b_2 + \omega(2-\omega)b_1\right]}{4\omega(1-\omega)\left(2\lambda_2\nu_1 - \nu_2^2\right)\left[b_2 + \omega(2-\omega)b_1\right] - \lambda_2 b_2\left(\omega^2 b_1 + b_2\right)}\widehat{q}_G \tag{3.17}$$

This proposition explains that if it is better to allow parallel importation, then the manufacturer may be able to achieve higher profit by selling service to gray market customers. When selling service is beneficial to the manufacturer, she increases her price gap and tolerates more gray market activity to take advantage of the opportunity to sell service to gray market customers.

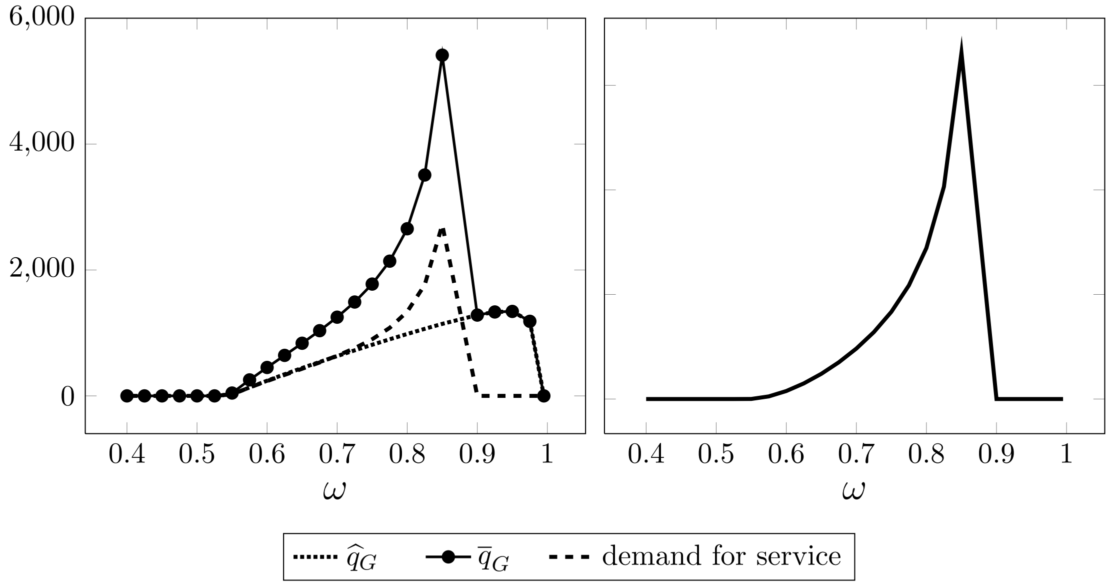
Figure 3.6 shows the effect of selling service to gray market customers on the size of the gray market and the manufacturer's profit. When ω is between 0.55 and 0.9, the manufacturer is better off selling service to gray market customers. The size of the gray market and the number of service buyers increase until $\omega = 0.85$ and decline when $0.85 < \omega \leq 0.9$. When ω exceeds 0.9, additional profit from selling service does not payoff the sales lost to the parallel importer. Thus, the manufacturer stops selling service and $\overline{q}_G = \widehat{q}_G$. Figure 3.6(b) indicates that the manufacturer can increase her profit by up to 6% if she considers selling service to buyers of parallel imports.



(a) $\overline{q}_G$ and demand for service (b) increase in manufacturer profit (%)

Figure 3.6: Effect of selling service ($\lambda_2 = 17, \nu_1 = 12, \nu_2 = 8$)

In this experiment, we assumed that the values of ν_1 and ν_2 are fixed. One may argue that consumers' relative perception of parallel imports, ω, influences their responsiveness to the level of service the manufacturer offers as well as the service fee charged. When consumers have a high perception of parallel imports, they are indifferent between buying from the manufacturer and buying from the parallel importer. Therefore, it is reasonable to expect that in this situation consumers have a low willingness to pay for service, unless the service fee is low or the level of service is high, meaning that ν_1 increase with ω and ν_2 decrease with ω. On the other hand, when ω is low (but not too low), consumers are more willing to pay for service to compensate for their lack of peace of mind when they buy the product from the gray market. We repeated our experiment with the same parameters, except that we let $\nu_1 = 12\omega$ and $\nu_2 = 8(1 - \omega)$. Figure 3.7 shows that varying ν_1 and ν_2 has minimal impact on the outcomes and does not change the nature of the observations.



(a) $\overline{q}_G$ and demand for service (b) increase in manufacturer profit (%)

Figure 3.7: Effect of selling service when $\nu_1 = 12\omega$ and $\nu_2 = 8(1-\omega)$.

3.7 Conclusion

In this chapter, we developed a game-theoretic model to analyze price and service competition between a manufacturer who operates in two markets and a parallel importer who transfers the product to the high-price market for resale. We characterized the optimal decisions of the manufacturer, which determine whether she should ignore, allow, or block parallel importation. We observed that the Stackelberg game equilibrium is a Prisoner's Dilemma. The parallel importer offers service to increase her profit, but both players will be better off if the parallel importer does not compete with the manufacturer on service.

We showed that parallel importation forces the manufacturer to reduce the price gap and provide more service both in the high-price market and in the low-price market. Even though the manufacturer always earns higher profits when she provides service, her price gap may be higher or lower than when she does not provide any service, depending on sensitivity of the markets to manufacturer service and parallel importer service.

We explored the value of service in coping with the competition from gray markets through the necessary entry conditions for the parallel importer and numerical experiments. We found that even a little service can be of great help for the manufacturer. In addition, when consumers become indifferent between buying from the manufacturer and from the parallel importer, as in the case of commodity products, service can help the manufacturer differentiate herself from the gray market.

We also explored the effect of the parallel importer's free-riding behavior on the outcome of the competition. We observed that the manufacturer's reaction to free-riding is to reduce the amount of service in the high-price market and reduce the price gap.

Finally, we addressed the manufacturer's policy for parallel imports and considered the possibility of selling service to customers who buy imported products. We found that the manufacturer may be able to increase her total profit by charging these customers a service.

This research can be extended in several directions. In this work, we focused a vertically integrated manufacturer. An interesting extension is to consider a decentralized supply chain in which the manufacturer sells the product through one or competing retailers in each market and the parallel importer obtains the product from the retailers. The manufacturer can offer service to customers directly or delegate the service decision to the retailers using different contractual agreements. Incorporating demand uncertainty is another potential direction for future research.

3.8 Appendix A. Proofs

Proof of Proposition 3.1. When the manufacturer allows parallel importation, her sales in market 1 go up by selling q_G units to the importer. On the other hand, the manufacturer's sales in market 2 changes by $N_2 - \frac{b_2(p_2-p_G)-\theta_2 s_2}{1-\omega} - (N_2 - b_2 p_2 + \theta_2 s_2) = -\frac{b_2(\omega p_2 - p_G)-\omega\theta_2 s_2}{1-\omega} = -\omega q_G$. □

Proof of Proposition 3.2. We have $\frac{d}{dq_G}\pi = p_1 - \omega p_2 - c(1-\omega)$, which is negative when the parallel importer enters, because $q_G^* > 0$ necessitates $\omega p_2 - p_1 > 0$. Therefore, the manufacturer's profit is decreasing in q_G, and the optimal decisions are (p_i^m, s_i^m) is they eliminate parallel importation. It is easy to show that when $q_G > 0$, π is concave in strictly concave in the price and service decisions. The first order optimality conditions result in (3.8)–(3.11). If $q_G^* < 0$, then the solution to the game is obtained by solving (3.1) subject to $b_2(\omega p_2 - p_1 - c_G) = \omega\theta_2 s_2$, which gives (3.12). □

Proof of Proposition 3.3. The gray marketer enters the competition only if $q_G^* > 0$ and manufacturer's prices and services that allow the importer are non-negative. First $q_G^* > 0$ necessitates $\omega N_2 - b_2 c - b_2 c_G > 0$. To derive the second and third conditions, we first note that $q_G^* > 0$ requires $\omega p_2^a > p_1^a$. From equations (3.11) we have

$$\omega p_2^a - p_1^a \;=\; \frac{2(1-\omega)\lambda_G b_2 - \omega\theta_G^2}{(2-\omega)\lambda_G b_2 - \omega\theta_G^2}\left(\frac{\omega\lambda_2 s_2^a}{\theta_2} - \frac{\lambda_1 s_1^a}{\theta_1}\right) > 0 \Longrightarrow \frac{\omega\lambda_2 s_2^a}{\theta_2} - \frac{\lambda_1 s_1^a}{\theta_1} > 0$$

which gives us

$$0 < q_G^* < \frac{\lambda_G b_2}{\lambda_G b_2(2-\omega) - \omega\theta_G^2}\left[\frac{\omega N_2 - b_2(c+c_G)}{\omega} - \frac{2\lambda_1 b_2 s_1^a}{\omega\theta_1}\right]$$

$$\Longrightarrow \theta_1^2 < 2\lambda_1 b_1 - \frac{2\lambda_1 b_2(N_1 - b_1 c)}{\omega N_2 - b_2 c - b_2 c_G}.$$

For the third condition, first define $R(\theta_2) = b_2\left(\omega p_2^m - p_1^m - c_G\right) - \omega\theta_2 s_2^m$. Then $\frac{d}{d\theta_2}R(\theta_2) = \frac{-2\lambda_2 b_2\theta_2(N_2 - cb_2)\omega}{(2\lambda b_2 - \theta_2^2)^2} < 0$ and $R\left(\sqrt{\lambda_2 b_2}, \omega\right) = -b_2\left(\omega c - p_1^m - c_G\right) < 0$. Therefore, if $\theta_2 \geq \sqrt{\lambda_2 b_2}$, then the optimal prices and services in the absence of parallel importation automatically eliminate the gray market. The second bound on θ_2 ensures that s_2^a is positive and bounded. □

Proof of Proposition 3.4. For part (a), suppose the optimal policy is to allow importation. To prove $p_2^a < p_2^m$, we note that q_G^* is increasing in b_1 whereas $\Delta p_2 = p_2^a - p_2^m$ is decreasing in b_1. Let $\widehat{b}_1$ be the value of b_1 for which $\Delta p_2 = 0$ and $\bar{b}_1$ be the value of b_1 for which $q_G^* = 0$. Then after some algebra, we can show that

$$q_G^*\left(\widehat{b}_1\right) = -\frac{2\lambda_G b_2\left(\lambda_2 b_2 - \theta_2^2\right)\left[\lambda_2\left(c\left(1-\omega\right) + c_G\right)b_2^2 + \frac{1}{2}\theta_2^2\left(\omega N_2 - b_2\left(c + c_G\right)\right)\right]}{\omega\left(2\lambda_2 b_2 - \theta_2^2\right)\left[\left(2\lambda_2 b_2 - \theta_2^2\right)\left[\left(2-\omega\right)\lambda_G b_2 - \omega\theta_G^2\right] - 2\lambda_2\lambda_G b_2^2\omega\right]},$$

$$\Delta p_2\left(\bar{b}_1\right) = \frac{\omega}{b_2}q_G^*\left(\widehat{b}_1\right).$$

Because of Assumption 1 and Proposition 3.3, $q_G^*\left(\widehat{b}_1\right)$ and $\Delta p_2\left(\bar{b}_1\right)$ are negative. This means that the smallest b_1 that makes q_G^* positive makes Δp_2 negative. Therefore $p_2^a < p_2^m$ holds for all solutions that allow parallel importation. Regarding service in market 2, we have

$$s_2^a - s_2^m = \frac{\lambda_G b_2\left[2\lambda_2 b_2^2(b_2 c_G + b_2 c\left(1-\omega\right)) + \theta_2^2\left(\omega N_2 - b_2 c - b_2 c_G\right)\right]}{\left[\left(2\lambda_2 b_2 - \theta_2^2\right)\left[\left(2-\omega\right)\lambda_G b_2 - \omega\theta_G^2\right] - 2\lambda_2\lambda_G b_2^2\omega\right]\left(2\lambda_2 b_2 - \theta_2^2\right)}\theta_2 > 0$$

For service in market 1,

$$s_1^a - s_1^m = \frac{\lambda_G b_2\left[\left(2\lambda_1 b_1 - \theta_1^2\right)\left(\omega N_2 - b_2(c + c_G)\right) - 2\lambda_1 b_2\left(N_1 - b_1 c\right)\right]}{\left[\omega\left(2\lambda_1 b_1 - \theta_1^2\right)\left[(2-\omega)\lambda_G b_2 - \omega\theta_G^2\right] + 2\lambda_1\lambda_G b_2^2\right]\left(2\lambda b_1 - \theta_1^2\right)}\theta_1$$

If $s_1^a - s_1^m \leq 0$, then we will have

$$\theta_1^2 \geq 2\lambda_1 b_1 - \frac{2\lambda_1 b_2 \left(N_1 - b_1 c\right)}{\omega N_2 - b_2 c - b_2 c_G} \tag{3.18}$$

which contradicts with the necessary condition for $q_G^* > 0$. Thus, $s_1^a > s_1^m$. Finally, because $p_1^a = c + \lambda_1 \frac{s_1^a}{\theta_1}$ and $p_1^m = c + \lambda_1 \frac{s_1^m}{\theta_1}$, $s_1^a > s_1^m$ results in $p_1^a > p_1^m$.

Now suppose the optimal policy is to block importation. Then there exists a positive Lagrangian multiplier $\mu > 0$ such that $\left(p_1^b, s_1^b, p_2^b, s_2^b\right)$ and μ satisfy the following optimality conditions

$$N_1 - 2b_1 p_1 + cb_1 + \theta_1 s_1 + b_2 \mu = 0 \tag{3.19}$$

$$N_2 - 2b_2 p_2 + cb_2 + \theta_2 s_2 - \omega b_2 \mu = 0 \tag{3.20}$$

$$(p_1 - c)\theta_1 - \lambda_1 s_1 = 0 \tag{3.21}$$

$$(p_2 - c)\theta_2 - \lambda_2 s_2 + \omega \theta_2 \mu = 0. \tag{3.22}$$

Because $(p_1^m, s_1^m, p_2^m, s_2^m)$ satisfy the above equations with $\mu = 0$, by obtaining s_1 in the third equation and replacing it in the first equation and using Assumption 3.1, we conclude that $p_1^b > p_1^m$ which means $s_1^b > s_1^m$. Next, we replace find s_2 in the fourth equation and replace it in the second equation and use Assumption 3.1 and $\theta_2 < \sqrt{\lambda_2 b_2}$ to conclude that $p_2^b < p_2^m$. Finally, if we equate the solution to μ in the second and fourth equations, we get $s_2 = \frac{N - b_2 p_2}{\lambda_2 b_2 - \theta_2^2}\theta_2$. Because $s_2^m = \frac{N - b_2 p_2^m}{\lambda_2 b_2 - \theta_2^2}\theta_2$ and we showed that $p_2^b < p_2^m$, s_2^b is larger than s_2^m.

Next, we prove part (c). For the allow policy, the result follows because p_1^a and p_2^a are convex and increasing in θ_1 and θ_2 and reduce to $\widehat{p_1}^a$ and $\widehat{p_2}^a$ when $\theta_1 = \theta_2 = 0$. For the block policy, if s_1 and s_2 are exogenously determined, then the optimality equations reduce to those for $\widehat{p_1}^b$ and $\widehat{p_2}^b$, which are increasing in N_1 and N_2. Therefore, when service levels are also optimized, p_1^b and p_2^b must be larger than $\widehat{p_1}^b$ and $\widehat{p_2}^b$. □

Proof of Proposition 3.5. Part (a) follows from taking the derivative of p_1^a, p_2^a, s_1^a, and s_2^a, and using Proposition 3.3. Part (b) follows from

$$p_2^a - p_1^a = \lambda_2 \frac{2\lambda_G b_2 (1-\omega) - \omega\theta_G^2}{[(2-\omega)\lambda_G b_2 - \omega\theta_G^2]\,\theta_2} s_2^a + \lambda_1 \frac{\omega\theta_G^2 - \lambda_G b_2 (1-\omega)}{[(2-\omega)\lambda_G b_2 - \omega\theta_G^2]\,\theta_1} s_1^a \tag{3.23}$$

and that s_1^a/θ_1 and s_2^a/θ_2 are increasing convex in θ_1 and θ_2. For part (c), define

$$K_1 = \frac{2\omega[(2-\omega)\lambda_G b_2 - \omega\theta_G^2] s_1^a}{\omega\left(2\lambda b_1 - \theta^2\right)[(2-\omega)\lambda_G b_2 - \omega\theta_G^2] + 2\lambda\lambda_G b_2^2} \tag{3.24}$$

$$K_2 = \frac{2[(2-\omega)\lambda_G b_2 - \omega\theta_G^2] s_2^a}{\left(2\lambda b_2 - \theta^2\right)[(2-\omega)\,\lambda_G b_2 - \omega\theta_G^2] - 2\lambda\lambda_G b_2^2\omega}. \tag{3.25}$$

Then $K_2 > K_1$ because $s_2^a \geq s_1^a$ and it is easy to show that the ratio multiplied by s_2^a is larger than the ratio multiplied by s_1^a due to $\omega < 1$. Now we have

$$\frac{d}{d\theta}\left(s_2^a - s_1^a\right) = \frac{s_2^a - s_1^a}{\theta} + (K_2 - K_1)\theta > 0$$

$$\begin{aligned}\frac{d^2}{d\theta^2}\left(s_2^a - s_1^a\right) &= 3(K_2 - K_1) \\ &\quad + 4\theta^2[(2-\omega)\lambda_G b_2 - \omega\theta_G^2]\left[\frac{K_2}{\left(2\lambda b_2 - \theta^2\right)[(2-\omega)\,\lambda_G b_2 - \omega\theta_G^2] - 2\lambda\lambda_G b_2^2\omega}\right. \\ &\quad \left. - \frac{\omega K_1}{\omega\left(2\lambda b_1 - \theta^2\right)[(2-\omega)\lambda_G b_2 - \omega\theta_G^2] + 2\lambda\lambda_G b_2^2}\right] > 0\end{aligned}$$

For $p_2^a - p_1^a$ we have

$$\frac{d}{d\theta}\left(p_2^a - p_1^a\right) = \left(\frac{\lambda}{[(2-\omega)\lambda_G b_2 - \omega\theta_G^2]}\right)\frac{d}{d\theta}\left([2\lambda_G b_2(1-\omega) - \omega\theta_G^2]\frac{s_2^a}{\theta}\right.$$

$$+\left[\omega\theta_G^2-\lambda_G b_2(1-\omega)\right]\frac{s_1^a}{\theta}\Bigg)$$

$$>\frac{\lambda\lambda_G b_2(1-\omega)K_1}{[(2-\omega)\lambda_G b_2-\omega\theta_G^2]}>0$$

and $\frac{d^2}{d\theta^2}\left(p_2^a-p_1^a\right)$ is equal to

$$\frac{\lambda}{[(2-\omega)\lambda_G b_2-\omega\theta_G^2]}\left[\frac{K_2(2\lambda_G b_2(1-\omega)-\omega\theta_G^2)-K_1((1-\omega)\lambda_G b_2-\omega\theta_G^2)}{\theta}\right]$$

$$+4\theta\Bigg[\frac{K_2(2\lambda_G b_2(1-\omega)-\omega\theta_G^2)}{(2\lambda b_2-\theta^2)\left[(2-\omega)\,\lambda_G b_2-\omega\theta_G^2\right]-2\lambda\lambda_G b_2^2\omega}$$

$$-\frac{\omega K_1((1-\omega)\lambda_G b_2-\omega\theta_G^2)}{\omega\left(2\lambda b_1-\theta^2\right)\left[(2-\omega)\lambda_G b_2-\omega\theta_G^2\right]+2\lambda\lambda_G b_2^2}\Bigg]>0$$

Since $p_2^a-p_1^a=\widehat{p_2}^a-\widehat{p_1}^a$ when $\theta=0$ and $\theta_G=0$, we conclude that the price gap is larger when service is offered.

For part (d), we have

$$\frac{d}{d\omega}s_1^a=\frac{-\lambda_G b_2\Phi}{\left[\omega\left(2\lambda b_1-\theta^2\right)\left[(2-\omega)\lambda_G b_2-\omega\theta_G^2\right]+2\lambda\lambda_G b_2^2\right]^2}\theta_1>0$$

$$\frac{d}{d\omega}s_2^a=\frac{(2\lambda_2 b_2-\theta_2^2)(c+c_G)\theta_G^2+\lambda_G\left[(2N_2-b_2c)\theta_2^2+(4\lambda_2 b_2-\theta_2^2)b_2c_G\right]}{\left[(2\lambda b_2-\theta^2)\left[(2-\omega)\,\lambda_G b_2-\omega\theta_G^2\right]-2\lambda\lambda_G b_2^2\omega\right]^2}(\lambda_G b_2^2\theta_2)>0$$

where

$$\Phi=-\omega(\lambda_G b_2+\theta_G^2)\left[(2\lambda_1 b_1-\theta_1^2)(\omega N_2-b_2(c+c_G))-2\lambda_1 b_2(N_1-b_1c)\right]$$

$$+b_2\left[(\lambda_G b_2+\theta_G^2)\omega-2\lambda_G b_2\right]\left[(2\lambda_1 b_1-\theta_1^2)c_G+2\lambda_1 N_1-c\theta_1^2\right]-2\lambda_1\lambda_G b_2^2 N_2<0$$

Φ is negative due to the necessary conditions for θ_1 to allow parallel importation, (3.5), and $2\lambda_i N_i-c\theta_i^2>0$. □

Bibliography

2020health. 2011. Parallel importing and exporting of pharmaceuticals severely limits the options in designing an effective UK drug pricing scheme. (November 15), `http://2020health.wordpress.com/2011/11/15/parallel-importing-and-exporting-of-pharmaceuticals-severely-limits-the-options-in-designing-an-effective-uk-drug-pricing-scheme`

Adler, P., A. Mandelbaum, V. Nguyen, E. Schwerer. 1992. From project to process management in engineering: strategies for improving development cycle time. *Sloan School of Management,* MIT.

AGMA, KPMG. 2008. Effective channel management is critical in combating the gray-market and increasing technology companies bottom line. `http://www.agmaglobal.org/press_events/KPMG%20AGMA%20Gray%20Market%20Whitepaper%20FINAL%20PRESS%20RELEASE.pdf`

Ahmadi, R., S. Carr., S. Dasu. 2012. Gray markets, demand uncertainty, and excess inventory. *Production Operations Management.* Forthcoming.

Ahmadi, R., T. A. Roemer, R. H. Wang. 2001. Structuring product development processes. *Eur. J. Oper. Res.* **130** 539–558.

Ahmadi, R., H. Matsuo. 2000. A mini-line approach to pull production. *Eur. J. Oper. Res.* **125** 340–358.

Ahmadi, R., B. R. Yang. 2000. Parallel imports: Challanges from unauthorized distribution channels. *Marketing Science.* **19**(3) 279–294.

Altug, M., G. van Ryzin. 2011. Supply chain efficiency and contracting in the presence of gray market. *Working paper.*

Antia, K. D., S. Dutta, M. Bergen. 2004. Competing with gray markets. *Sloan Management Review.* **46**(1) 63-69.

Asundi, J., S. Sarkar. 2005. Staffing software maintenance and support projects. *Proceedings of 38th Hawaii International Conference on System Sciences.* Hawaii, 1–8.

Autrey, R. L., F. Bova., D. Soberman. 2012. Organizational structure and gray markets. *Working paper.*

Banerji, S. 1990. A theory of gray markets: The case of the personal computer industry. Ph.D. thesis, Newswestern University.

Bangkok Post. 2010. BMW Thailand to No Longer Service Gray-Market Vehicles. Available at `http://wardsauto.com/news-amp-analysis/bmw-thailand-no-longer-service-gray-market-vehicles`

Bangkok Post. 2011a. Benz cutting prices to foil grey market. Available at `http://www.bangkokpost.com/auto/autoscoop/255693/benz-cutting-prices-to-foil-grey-market`

Bangkok Post. 2011b. Mercedes-Benz to defend market share from direct imports. Available at `http://www.bangkokpost.com/business/marketing/256095/mercedes-benz-to-defend-market-share-from-direct-imports`

Banker, R. D., I. Khosla, K. K. Sinha. 1998. Quality and competition? *Management Sci.* **44**(9) 1179–1192.

Black, K. 2011. *Business Statistics: For Contemporary Decision Making.* Wiley, Hoboken, NJ.

Brooks, F. P. 1975. *The Mythical Man-Month: Essays on Software Engineering.* Addison–Wesley, Reading, MA.

Browning, T. R., R. V. Ramasesh. 2007. A survey of activity network-based process models for managing product development projects. *Prod. Oper. Management.* **16**(2) 217–240.

Bucklin, L. P. 1993. Modeling the international gray market for public policy decisions. *International Journal of Research in Marketing.* **10** 387-405.

Carrascosa, M., S. D. Eppinger, D. E. Whitney. 1998. Using the design structure matrix to estimate product development time. *ASME Design Automation Conference.* Atlanta, GA.

Chan, L. M. A., Z. J. M. Shen, D. Simchi-Levi, J. Swann. 2004. Coordination of pricing and inventory decisions: A survey and classification. *Handbook of Quantitative Supply Chain Analysis: Modeling in the E-Business Era.* D. Simchi-Levi, S. D. Wu, Z. J. M. Shen. Kluwer Academic Publishers. 335-392.

Chen, H. 2009. Gray marketing: Does it hurt the manufacturers? *Atlantic Economic Journal.* **37**(1) 23–35.

Chen, H. 2002. Gray marketing and unfair competition *Atlantic Economic J.* **30**(2) 196–204.

Computer World. 2010. Apple makes $208 on each $499 iPad. (January 29). http://www.computerworld.com/s/article/9150045/Apple_makes_208_on_each_499_iPad

Coughlan, A. T., D. A. Soberman. 1998. When is the best ship a leaky one? Segmentation, competition, and gray markets. INSEAD, Working Paper 98/60/MKT.

Cusumano, M. A. 1997. How MicroSoft makes large teams work like small teams. *Sloan Management Review.* **39**(1) 9–20.

Cusumano, M., A. MacCormack, C. F. Kemerer, B. Crandall. 2003. Software development worldwide: The state of the practice. *IEEE Software,* **20** 28–34.

Dawande, M., M. Johar, S. Kumar, V. S. Mookerjee. 2008. A comparison of pair versus solo programming under different objectives: An analytical approach. *Inf. Syst. Res.* **19**(1) 71–92.

Dutta, S., J. B. Heide, M. Bergen. 1999. Vertical territorial restrictions and public policy: Theories and industry evidence. *Journal of Marketing.* **63** 121–134.

Dutta, S., M. Bergen, G. John. 1994. The governance of exclusive territories when dealers can bootleg. *Marketing Science.* **13**(1) 83–99.

Electronic Tax Administration Advisory Committee. 2011. Annual Report to the Congeress. Accessed on September 26, 2011 at http://www.irs.gov/pub/irs-pdf/p3415.pdf

eMarketer. 2010. iPad sales to more than double next year. http://www.emarketer.com/Article.aspx?R=1008098

Feng, Q., V. S. Mookerjee, S. P. Sethi. 2006. Optimal policies for the sizing and timing of software maintenance projects. *Eur. J. Oper. Res.* **173** 1047–1066.

Gallini, N. T., A. Hollis. 1999. A contractual approach to the gray market *International Review of Law and Economics.* **19**(1) 1–21.

Ganslandt, G., K. E. Maskus. 2004. Parallel imports and the pricing of pharmaceutical products: Evidence from the European Union. *Journal of Health Economics.* **23**(5) 1035-1057.

Garey, M. R., D. S. Johnson. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman, San Francisco.

Graves S. C., H. C. Meal, D. Stefek, A. H. Zeghmi. 1983. Scheduling of re-entrant flow shops . *J. of Oper. Management.* **3**(4) 197–207.

Hax A. C., D Candea. 1984. *Production and Inventory Management.* Prentice Hall, Englewood Cliffs, NJ.

Hos, C. J., K. G. Shin. 1997. Allocation of periodic task modules with precedence and deadline constraints in distributed realtime systems. *IEEE Transactions on Computers.* **46**(12) 1338–1356.

Hu, M., M. Pavlin., M. Shi. 2011. When gray markets have silver linings: All-unit discounts, gray markets and channel management. *Working paper.*

Iyer, G. 1998. Coordinating channels under price and nonprice competition. *Marketing Sci.* **17**(4) 338–355.

Jeuland, A. P., S. M. Shugan. 1983. Managing channel profits *Marketing Sci.* **2** 239–272.

Ji, Y., V. S. Mookerjee, S. P. Sethi. 2005. Optimal software development: A control theoretic approach. *Inf. Syst. Res.* **16**(3) 292–306.

Joglekar, N. R., A. A. Yassine, S. D. Eppinger, D. E. Whitney. 2001. Performance of coupled product development activities with a deadline. *Management Sci.* **47**(12) 1605–1620.

Kanavos, P., P. Holmes. 2005. *Pharmaceutical Parallel Trade in the U.K.* The Institute for the Study of Civil Society. London, U.K.

Kekre, S., N. Secomandi, E. Sönmez., K. West. 2009. Balancing risk and efficiency at a major commercial bank. *Manufacturing Service Oper. Management.* **11** 160–173.

Krishnan, H., J. Shao., T. McCormick. 2011. Impact of gray markets on a decentralized supply chain. *Working paper.*

Krishnan, V., S. D. Eppinger, D. E. Whitney. 1997. A model-based framework to overlap product development activities. *Management Sci.* **43**(4) 437–451.

Kulkarni, V. G., S. Kumar, V. S. Mookerjee, S. P. Sethi. 2009. Optimal allocation of effort to software maintenance: A queueing theory approach. *Prod. Oper. Management.* **18**(5) 506–515.

Kumar, C. N. S., S. A. B. Sailesh. 2011. Leveraging after-sales service to gain competitive advantage. Available at `http://www.infosys.com/supply-chain/white-papers/Documents/leveraging-after-sales.pdf`

Kumar, S., Y. Ji, S. P. Sethi, D. H. Yeh. 2006. Dynamic optimization of software enhancement effort. *Proceedings of 16th Workshop on Information Technologies and Systems (WITS).* Milwaukee, WI, 133–138.

Li, K., S. Mallik., D. Chhajed. 2012. Design of extended warranties in supply chains under additive demand. *Production & Operations Management.* Forthcoming.

Li, C., K. E. Maskus. 2006. The impact of parallel imports on investments in cost-reducing research and development. *Journal of International Economics.* **68** 443-455.

Loch, C. H., C. Terwiesch. 1998. Communication and uncertainty in concurrent engineering. *Management Science.* **44**(8) 1032–1048.

Lu, J., Y. Tsao., C. Charoensiriwath. 2011. Competition under manufacturer service and retaile price. *Economic Modelling.* **28** 1256–1264.

Maskus, K. E. 2000. Parallel imports. *World Economy.* **23**(9) 1269-1284.

Matsushima, N., T. Matsumura. 2010. Profit-enhancing parallel imports. *Open Economics Review.* **21**(3) 433–447.

Moorthy, K. S. 1998. Product and price competition in a duopoly. *Marketing Sci.* **7** 141–168.

Mussa, M., S. Rosen. 1978. Monopoly and product quality *J. Econom. Theory.* **18** 301–317.

Myers, M. B. 1999. Incidents of gray market activity among U.S. exporters: Occurrences, characteristics, and consequences. *Journal of International Business Studies.* **30**(1) 105-126.

New York Times. 2008. After China ships out iPhones, smugglers make it a return trip. `http://www.nytimes.com/2008/02/18/business/worldbusiness/18iphone.html`

Nikon Canada. 2012. Product service & repair. Available at `http://en.nikon.ca/Service-And-Support/Service-And-Repair.page`

Nintendo Co.. 2012. US & Canada Warranty. Available at `http://www.nintendo.com/consumer/manuals/warrantytext_eng.jsp`

PCWorld. 2011. Apple's IPhone 4S Selling for $2,000 in China. `http://www.pcworld.com/article/241999/apples_iphone_4s_selling_for_2000_in_china.html`

Pentax Canada. 2012. Warranty information. Available at `http://pentaxcanada.ca/en/support/warranty_info.php`

Perdikaki, O., D. Kostamis, J. Swaminathan. 2011. Timing of price and service level decisions under competition and demand uncertainty *Working paper.*

Perry, M. K., R. H. Porter. 1990. Can reasale price maintenance and franchise fees correct sub-optimal levels of retail service. *International J. Indust. Organization.* **8** 114–141.

Petruzzi N. C., M. Dada. 1999. Pricing and the Newsvendor Problem: A review with extensions. *Operations Research.* **47**(2) 183–194.

Phistar.com. 2003. Beating the gray. Available at `http://www.philstar.com/Article.aspx?articleId=195992`

Richardson, M. 2002. An elementary proposition concerning parallel imports. *Journal of International Economics.* **56**(1) 233–245.

Roemer, T., R. Ahmadi. 2004. Concurrent crashing and overlapping in product development. *Oper. Res.* **52**(4) 606–622.

Roemer, T., R. Ahmadi, R. Wang. 2000. Time-cost trade-offs in overlapped product development. *Oper. Res.* **48**(6) 858–865.

Schonfeld, M. 2010. Fighting grey goods with trademark law. *Managing Intellectual Property.* `http://www.pancommunications.com/_docs/burnslevinson20100601.pdf`

Schweitzer, M. E., G. P. Cachon. 2000. Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Sci.* **46**(3) 404–420.

Shaw, M., P. Clements. 2006. The golden age of software architecture. *IEEE Software.* **23**(2) 31–39.

Shulman, J. D. 2012. Product diversion to a direct competitor. *Working paper.*

Sigmaphoto.com. 2012. Warranty registration. Available at `http://warrantystatus.sigmaphoto.com/`

Smith, R. P., S. D. Eppinger. 1997. A predictive model of sequential iteration in engineering design. *Management Sci.* **43**(8) 1104–1120.

Software Magazine. The 2010 Software 500. Accessed on February 2, 2011 at `http://www.softwaremag.com/editors-desk/2010-software-500-another-good-year-for-outsourcers/`

Su X., S. K. Mukhopadhyay. 2011. Controlling power retailer's gray activities through contract design. *Production Operations Management.* Forthcoming.

Tavares, L. V. 1998. *Advanced Models for Project Management.* Kluwer Academic Publishers, Norwell, MA.

Ticino USA. 2012. Warranty service information. Available at `http://www.ticinowatches.com/service.html`

Tsay, A. A., N. Agrawal. 2000. Channel dynamics under price and service competition? *Manufacturing & Service Operations Management.* **2**(4) 372–391.

Wall Street Journal. 2011. Would-Be iPhone customers still facing weeks-long waits. `http://blogs.wsj.com/digits/2011/11/17/would-be-iphone-customers-still-facing-weeks-long-waits`

Winter, R. 1993. Vertical control and price versus non-price competition. *Quarterly J. Econom* **108**(1) 61–67.

Xiao, Y., U. Palekar., Y. Liu. 2011. Shades of gray – The impact of gray markets on authorzied distribution channels. *Quantitative Marketing and Economics.* **9** 155–178.

Xu, M., Y. Chen., X. Xu. 2010. The effect of demand uncertainty in a price-setting newsvendor model. *European Journal of Operational Research.* **207**(2) 946–957.

Yao, L., Y. Chen, H. Yan. 2006. The newsvendor problem with pricing: Extensions. *International Journal of Management Science and Engineering Management.* **1**(1) 3–16.