Carnegie Mellon
Tepper
SCHOOL OF BUSINESS

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
INDUSTRIAL ADMINISTRATION
(MANAGEMENT OF MANUFACTURING & AUTOMATION)

Titled

"IMPACT OF INFORMATION ON OPERATIONS MANAGEMENT IN EMERGING BUSINESSES"

Presented by
Ying Xu

Accepted by

Alan Scheller-Wolf	7/20/15
Co-Chair: Prof. Alan Scheller-Wolf	**Date**
Katia Sycara	7/20/15
Co-Chair: Prof. Katia Sycara	**Date**

Approved by The Dean

Robert M. Dammon	5/31/16
Dean Robert M. Dammon	**Date**

Impact of Information on Operations Management in Emerging Businesses

by

Ying Xu

Submitted to Tepper School of Business, Carnegie Mellon University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Management of Manufacturing and Automation

Dissertation Committee:
Professor Alan Scheller-Wolf (Co-Chair)
Professor Katia Sycara (Co-Chair)
Professor Soo-Haeng Cho
Professor Mor Harchol-Balter

May 27, 2016

Acknowledgments

First and foremost, I would like to give my sincerest gratitude to my two advisors, Prof. Alan Scheller-Wolf and Prof. Katia Sycara, for their continuous support of my Ph.D study and research.

Alan has always been the greatest source of my research motivation: Every time I got lost or frustrated with my research, after talking to him I would be fully loaded with confidence and encouragement again. His knowledge, intelligence, patience and support have helped me overcome a number of crisis situations during my Ph.D research. Without him this thesis would not have been completed. I sincerely hope that one day I would become an advisor as good as him.

I also would like to thank Katia for all the training she has provided on my research habits and skills. From the very beginning of my Ph.D research, she guided me how to express my research ideas, how to explore research opportunities through brainstorming questions, how to argue for my research opinions during discussions, and how to write reports and papers step by step. In particular, I appreciate her great effort in building and maintaining a vital group research atmosphere at her lab, which has prepared me for teamwork and collaborative research, from which I believe I would benefit for my whole life.

Alan and Katia together have built a strong foundation for me to finish the thesis, as well as my future research and career. I could not have imagined having better or more helpful advisors and mentors than them.

Besides my advisors, I would like to thank the following professors for their tremendous efforts and support. Prof. Soo-Haeng Cho has set up an excellent example for me to be a rigorous and professional scholar. Prof. Sridhar Tayur has opened up my mind in connecting theoretic research with real world practices with his brilliance and insight. My sincere thanks also go to Prof. Mor Harchol-Balter for arousing my interest in queueing research, for providing insightful advice on my research and career, and, most importantly, for inflaming me with her passion and desire to serve and support students. My dissertation have also received invaluable comments and suggestions from Prof. Javier Peña and Prof. Rhonda Righter (University of California, Berkeley).

I am sincerely obliged to the faculty members and labmates in the Advanced Agent-Robotics Technology Lab at Carnegie Mellon University: Dr. Tinglong Dai (Johns Hopkins University), Dr. Ronghuo Zheng (to be in the University of Texas at Austin), Dr. Nilanjan Chakraborty

(Stony Brook University) and Andreas Weit. I would not forget all the inspiring and stimulating discussions, all the sleepless nights we were working together before deadlines, and all the fun we have had in the lab.

During my six years at Carnegie Mellon University, I have also been blessed with a group of friendly and cheerful fellow students: Helen Zhou, Jessie Wang, Jieming Liu, Xiao Liu, Yang Yang, Xin Fang, Xin Wang, Ying Zhang, Vince Slaugh, Sherwin Doroudi, and Leela Nageswaran. I could not finish my Ph.D without their accompany and support. Special thanks go to Lawrence Rapp, who has always be patient and helpful in fulfilling all my requests.

In addition, I am grateful to my master advisor Prof. Ajay Joneja (Hong Kong University of Science and Technology) for enlightening me the first glance of research. I also thank Dr. Liang Lu, for his unconditional support, remarkable tolerance and supreme sacrifice. Last but not the least, I would like to thank my parents Linfu Xu and Jinfeng Yang for giving birth to me at the first place and constantly support me pursuing my dream throughout my life.

Dedicated to my parents

Contents

Chapter 1

Introduction

My research is dedicated to understand the role of information in managing emerging operations problems, particularly in the domains of socially responsible management and service operations.

In the first chapter I study how to combat child labor in global supply chains, specifically investigating how supply chain transparency influences the fight against child labor. There are nearly 200 million children engaged in child labor in the world, many in developing countries that are part of the supply base of global manufacturing networks. I model such a situation: A multinational firm in a developed nation selling a product made by a supplier in a developing country. I analyze the firm's strategies to control its supplier's use of child labor, and examine factors that affect the firm's incentives to use these strategies. I then investigate the potential effects of new legislation – the California Transparency in Supply Chain Act (the "Act") – that requires disclosure of corporate efforts to combat child labor. I find that supply chain transparency could backfire and inadvertently induce additional child labor: The Act, by serving as a commitment device, may enable the firm to credibly commit to conducting *no* internal inspections and hence encourage the supplier to use child labor. I also study several measures which can potentially mitigate such adverse effect (e.g., consumer boycotts, third-party organizations' support for firms' inspections) or eliminate it completely (e.g., a zero-tolerance policy).

In the second chapter, I propose a static service differentiation policy for a single-server queueing model of a service system. The policy randomly assigns homogeneous customers different service rates – independent of system state – while keeping the mean service time unchanged. Although conventional wisdom held that such differentiation would increase service time variability and thus

increase waiting time, I show the contrary: Such differentiation may reduce total waiting time because it creates information that enables the implementation of service-rate-based scheduling, which mitigates the increased service time variance. I provide conditions under which such a static service differentiation reduces waiting, and further derive closed-form expressions for the optimal differentiation policy. I also illustrate the policy in the context of quality-based service domains, in which customers value service time but dislike waiting. Numerically I find that providing differentiated service can improve system performance by 5% without the investment in any additional capacity.

The third chapter relates to a client who outsources service requests to two service vendors (servers). I uniquely consider the scenario in which during the service process the two service vendors can share their system state information as well as their capacity in their own interests, which I call "*strategic collaboration.*" I conduct both theoretical and numerical analysis to investigate a collaboration mechanism and the impact of the strategic collaboration. Through equilibrium analysis, I use a Markov Perfect Equilibrium (MPE) to characterize the collaboration equilibrium, and find that, counter intuitively, in equilibrium a server might be less likely to transfer jobs when there are more jobs waiting in its system, or when there are fewer jobs waiting for the other server. Further I conduct simulations to numerically examine the impact of strategic collaboration on dual-outsourcing. My numerical study shows that even though strategic collaboration reduces capacity investment, it could still reduce waiting by increasing capacity utilization. Specifically, strategic collaboration reduces waiting time more when the client implements capacity competition between servers.

Chapter 2

Combating Child Labor: Incentives and Information Transparency in Global Supply Chains

2.1 Introduction

International Labor Organization (ILO) defines child labor as "work that is mentally, physically, socially or morally dangerous and harmful to children; and interferes with their schooling." In 2012 there were 168 million child laborers worldwide, accounting for around 11% of the entire child population (ILO 2013). According to ILO (ILO 2007), child labor is deemed as a severe human rights violation that should be eliminated:

"All child labour, and especially the worst forms, should be eliminated. It not only undermines the roots of human nature and rights but also threatens future social and economic progress worldwide. Trade, competitiveness and economic efficiency should not be a pretext for this abuse."

Alarmingly, the decline of child labor has slowed down (ILO 2010). The progress to end child labor is challenged by the prevalence of global outsourcing. Economic research suggests that firms' global search for cheap labor has boosted the demand for child labor (Iram and Fatima 2008, Acaroglu and Dagdemir 2010). There are numerous examples of products produced by child labor, including cotton from Uzbekistan, cocoa from Ivory Coast, carpets and garments from India and

Pakistan, electronic products and toys from China, and so on (U.S. Department of Labor 2014). Child labor enables local suppliers to keep production costs down, and the surplus from such low costs is then passed on to multinational firms along the global value chain (Fyfe and Jankanish 1997, Locke 2003). Therefore, many multinational firms lack financial motivation to control their local suppliers' use of child labor.

Even when some firms have incentives to address child labor in their supply chains, they are hindered by the lack of direct control over their suppliers' use of child labor. Thus, in a decentralized supply chain, firms often adopt two indirect approaches to tackle the issue: internal inspections and contractual relationships (Kolk and van Tulder 2002a). Each approach has its own drawbacks. First, internal inspections on child labor are costly and imperfect. Monitoring of labor conditions requires investigation of production sites, which is often challenging or "virtually impossible" (International Finance Corporation 2002). For example, in carpet production in India, 175,000 looms were estimated to be in the Utta Paradesh carpet belt alone, and most were located in small workshops and even local households. IKEA, one of major retailers of the Indian carpets, concluded that "no one could monitor such a fragmented production process" (Bartlett et al. 2006). Kolk and van Tulder (2002b) also mention the complexity of inspecting suppliers in the garment industry because sourcing networks may involve thousands of factories spread across multiple countries. Second, it is likewise costly for a firm to influence its supplier's practice through contractual relationships. A firm may deter its supplier from hiring child labor by threatening to terminate the contract, but this may not be effective unless the firm leaves considerable profits to the supplier, which in turn increase the firm's outsourcing cost. For example, Obeetee, an Indian carpet manufacturer, increased wages significantly as an incentive to loom owners, while informing them (in writing) that if found employing child labor, they would lose their business and be blacklisted from doing any future business with the company (International Finance Corporation 2002). Similarly, Bayer CropScience put 5% of its procurement price as a bonus for crop farmers who did not use child labor (Subramanian 2013).[1]

In light of these challenges, third-party organizations have stepped in. First, to impose pressure

[1]In order to urge Bayer Group to eliminate child labor used in the company's crop production in India, a group of European NGOs cosigned an open letter in 2003, which states: "In order to prevent your suppliers from using child labor, the prices paid to their products need be high enough so that employment of adults is profitable to the suppliers" (Subramanian 2013).

on global firms, these organizations have developed programs to monitor child labor practices at local production sites, and launched consumer education campaigns to call for consumer boycott of products involved with child labor. Due to extensive media reach coupled with advances in information technology, firms that sell such products would likely incur a reputation loss in both consumer and financial markets (Smith 2003). Second, to ameliorate difficulties in internal inspections, non-governmental organizations started to provide support to firms' internal monitoring. They have undertaken a variety of strategies to facilitate firms' abilities to monitor suppliers' child labor practices in a more cost-effective way (U.S. Department of Labor 2000). For example, they encourage collaborations with industry, employer, and worker organizations (e.g., Child Labor Elimination Group in the agriculture industry of India, the Atlanta Agreement in the global footwear industry), set up a data base system to record information gathered from various auditing programs (e.g., the ILO's International Program on the Elimination of Child Labor), and consolidate village-located football production sites into large-scale stitching centers (Lund-Thomsen P. 2008).

Other third-party organizations have advocated legislative measures to promote firms' efforts in regulating their suppliers. One recent achievement is the passing of a bill on supply chain transparency – the California Transparency in Supply Chains Act (Senate Bill 657) (hereinafter the "Act"). The Act requires manufacturers and retailers in California with annual revenue of $100 million or more to publicly disclose to which degree the firms are engaged in combating forced labor (including child labor) in their supply chains. In particular, those firms are required to disclose their inspection policies on whether they have relevant internal auditing on their suppliers. Although the primary purpose of the Act is to inform consumers of firms' efforts, their suppliers are also informed.[2] Therefore, the Act, serving as a commitment device, enables a firm to credibly commit its inspection policy to its supplier. However, this may have unintended consequences.

To date, there has been little research on evaluating the impact of these initiatives on child labor in global supply chains. This paper aims to provide insights into how these initiatives affect firms' strategies and incentives to control their suppliers' use of child labor. To this end, we develop a game theoretic (principal-agent) model based on a two-tier supply chain, in which a

[2]Prior to the enactment of the Act, it may have been difficult for suppliers to access information related to inspection policy, or even if suppliers are informed of inspection polices, it would have been difficult for them to verify its truthfulness and the information disclosed might not have been credible because such disclosures were not subject to supervision under the Act. In contrast, after the enactment of the Act in California, websites such as `https://www.knowthechain.org/` record the information disclosed by firms.

multinational manufacturer ('she') outsources her production to a local supplier ('he') through a wholesale-price contract. The supplier has an option to use child labor in place of legitimate labor at a lower cost. However, child labor, if exposed to the public by third party organizations, would incur a goodwill loss to the manufacturer. The manufacturer may carry out costly (but imperfect) internal inspections to monitor her supplier's child labor practice. When the manufacturer detects a violation, she requires the supplier to take corrective actions by removing child laborers and compensating them. When the supplier's use of child labor has been missed by the manufacturer's inspection but is later found by third parties, the manufacturer may terminate the contract with her supplier. The structure of the game is impacted by the enactment of the Act: before the introduction of the Act, the supplier is unable to verify the manufacturer's inspection strategy. In contrast, after the Act, the supplier is informed of the manufacturer's inspection strategy before deciding whether to employ child labor, since the Act requires the manufacturer to disclose its inspection strategy truthfully. Therefore, we formulate two sequential games – before or after the Act – and characterize the choice of each party in each game by using subgame-perfect equilibrium.

We summarize our main findings for three major stakeholders.

1. Global Manufacturers: A global manufacturer may use three different strategies to control the use of child labor in its supply chain. First, when internal inspections are affordable, the manufacturer is able to reduce the incidence of child labor by undertaking inspections to detect and remove child labor hired by the supplier. Second, when internal inspections are costly, the manufacturer can deter the supplier's child labor employment by offering a sufficiently high wholesale price, which will guarantee the supplier a high profit margin and thus cause him a high potential loss when he loses a contract with the manufacturer. Third, the manufacturer could simultaneously use internal inspections and a high wholesale price to deter the employment of child labor. This option would be adopted only after the enactment of the Act, because the manufacturer then can credibly commit to undertaking inspections even if she expects no child labor to be employed by the supplier under the high wholesale price.

2. Third-Party Organizations: Third-party organizations should be cautious that their support for manufacturers' inspections may not necessarily reduce child labor. When inspections are costly, global manufacturers faced with high public pressure would choose to combat child labor by offering high wholesale prices, which could incentivize suppliers not to employ child labor in the first place.

However, when inspections become less costly with the support of third-party organizations, these manufacturers may rely entirely on internal inspections (instead of offering high wholesale prices) to reduce the incidence of child labor. This would introduce more child labor being used on average because inspections are imperfect and may not always find child labor employed by suppliers. However, after the Act is introduced, this unintended consequence occurs less often because some manufacturers would adopt the combined strategy of using both internal inspections and high wholesale prices (which is more effective than the strategy of relying entirely on imperfect internal inspections).

3. Policy Makers: Policy makers should understand that the Act, serving as a commitment device, has both pros and cons. As discussed above, supply chain transparency achieved by the Act can help reduce child labor by enabling manufacturers to use the strategy of simultaneously using internal inspections and high wholesale prices. However, the Act may introduce more child labor by inducing some manufacturers to commit to exerting lower inspection effort even when inspection costs are low. Such commitment enables those manufacturers to convince their suppliers that there will be a lower level of inspection, thereby cutting their wholesale prices; whereas, in the absence of the Act, such commitment could not be made credibly. We further find that this adverse effect of the Act could be mitigated by several measures. One counter-intuitive measure is to allow manufacturers themselves to choose the amount of compensation paid to child labor detected during internal inspections. Interestingly, this adverse effect does not occur when firms adopt a zero-tolerance policy (i.e., terminate contracts whenever child labor is found).

2.2 Literature Review

There are three related research streams: (1) economics of child labor, (2) socially responsible supply chain management, and (3) quality management.

(1) Economics of Child Labor Economic research on child labor mainly examines the issue from two aspects: supply and demand of child labor. Research on the supply side analyzes factors that affect parents' incentives to send their children to work. These factors include agriculture output (e.g., accidental crop loss) by Beegle et al. (2006), crop price under trade liberalization by Edmonds and Pavcnik (2005), and household holding of lands by Basu et al. (2010). On the other

hand, other papers, including ours, focus on the demand side and study how to induce employers to hire less child labor (Basu and Van 1998, Basu and Zarghamee 2009, Davies 2005). Most economic papers implicitly assume that pressure imposed on local employers directly come from concerned consumers and organizations. We instead study the case in which such pressure is transferred from consumers to local employers through global supply chains. This perspective captures the current trend that child labor is increasingly involved in the products procured by multinational firms through global outsourcing, and that end-consumers have influence on multinational firms but less so on local employers of child labor.

(2) Socially Responsible Supply Chain Management This stream of work aims to achieve an overall socially responsible goal through coordinating various supply chain members. Analytical work in this area is emerging, and several papers are parallel to ours. Babich and Tang (2012) study mechanisms for dealing with product adulteration such as deferred payment, inspection or combination of both. Guo et al. (2014) analyze a manufacturer's outsourcing choice between responsible and risky suppliers when consumers are socially conscious. Chen and Lee (2014) focus on screening responsible suppliers through delayed-payment contracts. Kim (2014) investigates the relationship between a regulator's inspection activities and a production firm's voluntary disclosure of self-noncompliance.

Similar to our paper, Plambeck and Taylor (2014) study how a buyer can motivate its supplier to exert more effort to comply with labor and environmental standards. Our paper differs in several aspects, as we focus on the specific context of child labor. First, while much of analysis in Plambeck and Taylor (2014) assumes a fixed wholesale price, as documented in §1, the wholesale price has a significant influence on the incentive of a supplier hiring child labor. Thus, our model considers a global firm which can influence a supplier's decision on child labor by determining a wholesale price as well as a level of internal inspection. Our analysis hinges crucially on the interplay between these two levers. Second, based on several case studies on child labor policies of global firms, our model considers various penalty schemes for a supplier who employs child labor: corrective action, contract termination, or combination of both. Moreover, a global firm has some influence over the magnitude of the supplier's potential loss (through endogenous choice of the wholesale price and the compensation paid to a child worker). Different from our penalty schemes, Plambeck and Taylor (2014) consider a penalty scheme in which both firms will earn zero profit if noncompliance is found

from the internal audit, while a supplier incurs a fixed cost if noncompliance is found after passing the audit. Third, motivated by the Act, we examine the impact of supply chain transparency on a global firm's joint decision on inspection level and wholesale price, and its consequence on child labor. Similar in spirit (albeit not motivated by the Act), Plambeck and Taylor (2014) study the impact of a buyer's pre-commitment to auditing effort under a fixed wholesale price. Although they state that if the supplier were unable to hide, the ability to pre-commit would cause the buyer to do more auditing, we find that this is not always true when wholesale price is determined endogenously.

(3) Quality Management Although there are some similarities between the control of quality and that of child labor, the two differ in the nature of defection and its difficulty of inspections. Defective products could be detected by manufacturers alone through sampling approaches and inspection technology, and if not, they could often be found later by consumers after they are sold in the market. In contrast, a product made by child labor is not necessarily defective in product functionality, and detection of child labor is more complicated because neither manufacturers nor consumers can learn it from inspecting products. It may also involve a third stakeholder such as non-governmental organizations and the media. Manufacturers may need these organizations' support, training, and information for their internal inspections, and consumers may well get some information from these organizations. Public policy such as the Act also plays a role in the availability of credible information. Our paper puts special emphasis on studying the roles of third-party organizations and information transparency in the control of child labor, which is absent in most quality control papers.

Our paper is further distinct from most quality papers in studying the joint effects of contract, inspections, and information. Hwang et al. (2006) study how to control suppliers' product quality through a combination of inspections and contracts, but they do not consider the role of information in inspections. There are some papers that examine the role of information in quality investment under a fixed wholesale price (e.g., observability of information in Hsieh and Liu 2010, information noise in third-party certification in Chen and Deng 2013). Some papers study the joint impact of contract, inspections and information on product quality, but their focus is different from ours. Balachandran and Radhakrishnan (2005) consider the setting in which a product comprises components made by a buyer and a supplier, and examine how the information of the buyer's quality

effort shapes the design of penalties imposed onto the supplier during internal and external failures. Baiman et al. (2000) study the impact of the supplier's knowledge of the buyer's inspection effort on product quality. They find that when the supplier cannot verify the buyer's inspection results, the buyer may conduct no inspections, but still claim that products are defective and return all products. This would undermine the supplier's incentive to improve quality. Once information about the buyer's inspection effort is known to the supplier, the supplier would make more effort to improve quality, and therefore product quality would always be improved. In contrast, in our paper, when the supplier cannot verify the buyer's inspection effort, the buyer may still use inspections as a non-credible threat to deter the supplier from employing child labor; furthermore, even if information about the buyer's effort is known to the supplier (i.e., after the Act), it may induce more child labor, causing inferior product quality in the socially responsible sense (although the real product quality may remain the same).

2.3 The Model

We consider a decentralized supply chain in which a risk-neutral manufacturer ('she') outsources her production to a risk-neutral supplier ('he') via a wholesale-price contract. The outsourced production quantity is fixed, and is normalized to one. We assume that the supplier needs one unit of labor to produce this product. Let d denote the supplier's decision: $d = 1$ means that the supplier employs a child laborer, and $d = 0$ means that the supplier employs an adult laborer. For ease of exposition, we consider the case where the supplier chooses a pure strategy of either employing a child laborer or an adult laborer. In the online supplement, we analyze the case where the supplier may choose a mixed strategy as well as the case where the supplier may hire only a portion of workforce with child labor; and show that the two cases produce similar equilibrium outcomes and the key insights we obtain for a binary d continue to hold for these cases. We denote by s_H and s_L the labor cost for the supplier to hire an adult and a child, respectively, where $s_H > s_L$.[3] Without loss of generality, all other production costs are normalized to 0. The product is sold to the market at a fixed retail price v (> 0).

[3]We assume that the productivity of an adult laborer is the same as that of a child laborer. Although one might think that the former is higher than the latter, the latter could be higher than the former especially when small hands are useful for jobs (e.g., cotton picking and hand-knitting). Our model and analysis can be easily extended to the case when there is a fixed ratio of productivity between adult and child laborers.

The manufacturer decides the wholesale price w (> 0) and the amount of effort θ ($\in [0,1]$) made to inspect the supplier's employment of child labor. The effort θ determines the probability that the supplier's employment of child labor, if it exists, is detected by the manufacturer during the manufacturer's internal inspections. We thus refer to θ as the level of "internal" inspections. For tractable analysis, following Hwang et al. (2006) and Babich and Tang (2012), we adopt a binary inspection level: $\theta \in \{\theta_L, \theta_H\}$, where θ_H (θ_L) denote high (low) inspection level, and θ_L is set to 0 (i.e., no inspections).[4] The corresponding inspection costs are $I(\theta_H) = I > 0$ and $I(\theta_L = 0) = 0$, respectively. When the supplier's employment of child labor is discovered by the manufacturer, the supplier is mandated to perform a corrective action by paying the child worker a monetary compensation of m (> 0) and rehiring an adult to complete production. For simplicity, we assume that the detection occurs at the beginning of production, and that wages are paid at the end of production; thus when a child worker is found, s/he is removed with no wage but the monetary compensation m, and the adult worker is rehired at the adult wage s_H.[5]

Child labor is also subject to monitoring by third parties (e.g., non-profit organizations such as UNICEF and ILO, or the media). Let e ($\in (0,1)$) denote the probability that the supplier's use of child labor, if it exists, will be detected through such "external" inspections. If the manufacturer is found to use child labor in the outsourced production of her product, she will suffer from goodwill cost g (> 0), which includes short-term sales loss and long-term damage in reputation. In this case we assume that the manufacturer will discontinue her contract with the supplier, and the supplier will incur opportunity costs of losing future business.

Before the Act is enacted (i.e., in the pre-Act scenario), the sequence of decisions and events in our model is as follows (see Figure 2.1 for illustration):

(S1) The manufacturer offers a wholesale price w to the supplier. The supplier accepts the contract if he can earn a higher expected profit than his reservation profit (normalized to zero).

(S2) If the supplier accepts the contract, he makes a hiring decision on d, and carries out production. Simultaneously, the manufacturer chooses her inspection level θ.

[4] In practice, several companies such as IDEX, Caterpillar, and Danaher state that they do not verify their supply chains or audit suppliers to evaluate risks of human trafficking and slavery (see the online supplement). Similarly, there are a number of companies that have remained silent on the Act (Business & Human Rights Resource Center 2014). One may interpret that these companies commit to no or low-level inspection efforts.

[5] Alternatively, we may assume that a child worker receives cs_L (where $c \in [0,1]$) before s/he is found. This does not affect our results.

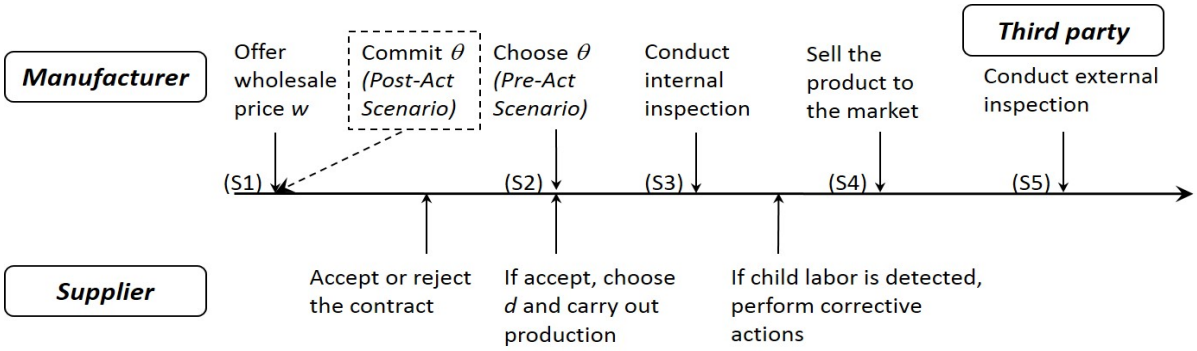

Figure 2.1: The Sequence of Decisions and Events

(S3) The manufacturer conducts internal inspections with the level θ. If child labor is found, then the supplier is required to pay compensation m to the child worker, and to rehire an adult worker at the cost of s_H.

(S4) Once the production is over, the manufacturer pays w to the the supplier, and sells the product to the market at the retail price v.

(S5) External inspections are conducted with the level e. If the child labor is discovered during these inspections, the manufacturer incurs the goodwill cost g, and she terminates her contract with the supplier.

After the Act is enacted (i.e., in the post-Act scenario), the supplier has information about the manufacturer's internal inspection level before he accepts the contract. Thus the sequence above is revised as follows:

(S1') The manufacturer offers a wholesale price w and commits to her inspection level θ. The supplier decides whether to accept the contract or not.

(S2') If the supplier accepts the contract, he makes a hiring decision on d, and carries out production.

(S3')-(S5'): The same as (S3)-(S5), respectively.

The revised sequence is also illustrated in Figure 2.1 with a different timing for the manufacturer's inspection decision shown in the dashed text box. Note that even before the Act is enacted, the manufacturer may disclose her inspection level θ, but her inspection level is not credible to the supplier. Thus, the manufacturer cannot commit to θ in the pre-Act scenario, while she can do so in the post-Act scenario.

The manufacturer's expected profit U is given as:

$$U(w, \theta, d) = v - w - I(\theta) - egd(1 - \theta), \tag{2.1}$$

where $egd(1-\theta)$ represents the expected goodwill cost. Recall that the manufacturer incurs the goodwill cost g when: (i) child labor is employed by the supplier (i.e., $d = 1$); (ii) it passes the manufacturer's internal inspections with probability $(1 - \theta)$; and (iii) it is detected and reported to the public by third parties with probability e. A similar assumption is made by Plambeck and Taylor (2014) and Chen and Lee (2014). For ease of exposition, we define $d_E \equiv d(1 - \theta)$ as the

expected amount of child labor that is *used* in outsourced production (and thus can be potentially exposed to the public). In our subsequent analysis, d_E is used as a measure of the severity of the child labor issue.

Let $\Delta(\theta)$ denote the supplier's expected labor cost saving from hiring child labor when the manufacturer has chosen her internal inspection level θ. Then we can express $\Delta(\theta)$ as:

$$\Delta(\theta) \equiv s_H - \{(1-\theta) s_L + \theta (s_H + m)\} = (s_H - s_L) - (s_H - s_L + m)\theta. \tag{2.2}$$

Using $\Delta(\theta)$, we can express the supplier's expected profit Π as

$$\Pi(w, \theta, d) = (1 - \gamma e d (1-\theta))(w - s_H + d\Delta(\theta)). \tag{2.3}$$

In (2.3), $w - s_H + d\Delta(\theta)$ represents the supplier's profit when there is no risk of child labor being detected by third parties. The term $\gamma e d (1-\theta)(w - s_H + d\Delta(\theta))$ represents the supplier's expected opportunity cost due to external inspections on the use of child labor, in which γ (> 0) is a constant that represents a (discounted) ratio of potential future orders to the current order (of size 1),[6] and $e(1-\theta)$ represents the probability that the manufacturer will discontinue her contract with the supplier due to child labor discovered during external inspections after passing internal inspections.

The manufacturer chooses her wholesale price w and inspection level θ to maximize $U(w, \theta, d)$. Then the supplier chooses his child labor employment level d to maximize $\Pi(w, \theta, d)$. In the pre-Act scenario the supplier knows only w, whereas he knows both w and θ in the post-Act scenario. We make the following assumptions to rule out some unrealistic or uninteresting cases: (A1) $\gamma e \leq 1$ and (A2) $\theta_H \leq \dfrac{s_H - s_L}{s_H - s_L + m}$. Assumption (A1) requires that either the level of external inspections e or the ratio of potential future orders to the current order γ is sufficiently low. Without (A1), the supplier would incur such a high expected opportunity cost of employing child labor that he would never hire child labor, even when no internal inspection is undertaken (i.e., $\Pi(w, \theta = \theta_L, d = 1) < 0$ for any value of w). Assumption (A2) ensures that the supplier could

[6]The parameter γ may also represent a portion of future business cut by the manufacturer due to external inspections on the use of child labor. For example, Samsung cut 30% of its business with a Chinese supplier after evidence of child laborer was found at its factory (BBC 2014).

save his labor cost in expectation by hiring child labor even when the manufacturer has chosen to undertake internal inspections (i.e., $\Delta(\theta_H) \geq 0$). This assumption rules out an uninteresting case in which the manufacturer's internal inspections can stop the supplier from hiring child labor for any wholesale price w.

Before we proceed to our analysis, we discuss a few issues of our model. First, while we consider a fixed retail price in our base model, in §6.1, we consider the case when the manufacturer commands a higher retailer price when choosing to conduct internal inspections in the post-Act scenario than when choosing no inspections. This represents a situation in which consumers are aware of socially-responsible manufacturers and have higher valuation on the products made by such manufacturers. Second, we assume that the supplier is required to pay compensation m to child laborers detected during internal inspections. The amount of compensation m reflects living and education costs which may differ by countries. The provision of such stipends is important so as to avoid children from moving from one workplace to another, and it is often required by industry agreements (e.g., the Atlanta Agreement in the footwear industry and the agreement among Bangladesh Garment Manufacturers and Exporters Association, ILO and UNICEF). However, it is also plausible that a manufacturer has flexibility in choosing the amount of compensation m, so we analyze this case in §6.2. Finally, in §6.3, we consider different penalty schemes of a manufacturer who requires only corrective actions from the supplier ("soft policy") or always terminates the contract permanently ("zero-tolerance policy") whenever child labor is found in her supply chain.

In §2.4 and §2.5.1, we derive firms' equilibrium decisions in the pre-Act and post-Act scenarios, respectively. In §2.5.2, we then compare firms' equilibrium strategies under these two scenarios to examine the impact of the Act on each party's strategies. All proofs are presented in Appendix A.

2.4 Equilibrium Analysis in the Pre-Act Scenario

Before the Act is enacted, the manufacturer cannot credibly commit to her decision on θ. Thus, in a subgame for a given wholesale price w, the manufacturer and the supplier simultaneously determine the inspection level θ and the child labor decision d, respectively, anticipating the best response of the other party to his/her decision. Let $(\theta^{pre}(w), d^{pre}(w))$ denote the equilibrium in the simultaneous subgame for a given w. Throughout the paper we let the superscript "pre"

indicate results in the *pre*-Act scenario. Anticipating $(\theta^{pre}(w), d^{pre}(w))$, the manufacturer solves the following program at the contract stage (S1) to choose the wholesale price w that maximizes her expected profit U:

$$\max_{w} \quad U\left(w, \theta^{pre}(w), d^{pre}(w)\right) \tag{2.4}$$

$$s.t. \quad \Pi\left(w, \theta^{pre}(w), d^{pre}(w)\right) = w - s_H + \Delta\left(\theta^{pre}(w)\right) d^{pre}(w) \geq 0 \tag{2.5}$$

$$\theta^{pre}(w) \quad = \underset{\theta \in \{0, \theta_H\}}{\arg\max}\, U\left(w, \theta, d^{pre}(w)\right) = \underset{\theta \in \{0, \theta_H\}}{\arg\max}\, v - w - I\left(\theta\right) - e\left(1-\theta\right) g d^{pre}(w) \tag{2.6}$$

$$d^{pre}(w) = \underset{d \in \{0,1\}}{\arg\max}\, \Pi\left(w, \theta^{pre}(w), d\right) = \underset{d \in \{0,1\}}{\arg\max}\ \{1 - \gamma e\left(1 - \theta^{pre}(w)\right) d\}\{w - s_H + \Delta\theta^{pre}(w) d\}. \tag{2.7}$$

Constraint (2.5) ensures that the supplier earns non-negative expected profits to accept the contract. Constraints (2.6) and (2.7) ensure that the manufacturer's inspection decision $\theta^{pre}(w)$ and the supplier's employment decision $d^{pre}(w)$ are the best response to each other's equilibrium strategy for any given w. We first find equilibrium $(\theta^{pre}(w), d^{pre}(w))$ in a subgame for a given w from (2.6) and (2.7), and then substitute them into (2.4) and (2.5) to find a subgame-perfect equilibrium w^{pre}.

2.4.1 Equilibrium in a Subgame of a Fixed w in the Pre-Act Scenario

For any fixed w, we first find the supplier's best response function $d^{pre}\left(\theta, w\right)$ to the manufacturer's decision on θ, and the manufacturer's best response function $\theta^{pre}\left(d, w\right)$ to the supplier's decision on d. Then we derive equilibrium $\left(\theta^{pre}\left(w\right), d^{pre}(w)\right)$ in a subgame that satisfies $\theta^{pre}(w) = \theta^{pre}(d^{pre}(w), w)$ and $d^{pre}(w) = d^{pre}(\theta^{pre}(w), w)$.

The supplier determines his best response $d^{pre}\left(\theta, w\right)$ for any given (θ, w) by evaluating the difference in his expected profit between hiring child labor and not, which is given as:

$$\Pi\left(w, \theta, d=1\right) - \Pi\left(w, \theta, d=0\right) = \Delta\left(\theta\right) - \gamma e\left(1-\theta\right)\left(w - s_H + \Delta\left(\theta\right)\right).$$

The supplier faces a trade-off between the expected labor cost saving from hiring child labor, $\Delta\left(\theta\right)$, and the potential opportunity cost due to external inspections, $\gamma e\left(1-\theta\right)\left(w - s_H + \Delta\left(\theta\right)\right)$. The supplier will not employ child labor if and only if the potential opportunity cost is no less than the cost saving; i.e., $d^{pre}\left(\theta, w\right) = 0$ if and only if $\gamma e\left(1-\theta\right)\left(w - s_H + \Delta\left(\theta\right)\right) \geq \Delta\left(\theta\right)$, which is simplified

to $w \geq s_H + \{1/(\gamma e(1-\theta)) - 1\}\Delta(\theta)$; otherwise, $d^{pre}(\theta, w) = 1$.[7] This suggests that in order to incentivize the supplier not to hire child labor, the manufacturer should pay the supplier a premium of at least $\{1/(\gamma e(1-\theta)) - 1\}\Delta(\theta)$ over the supplier's labor cost s_H. Since the premium increases the supplier's marginal profit, it increases the supplier's potential opportunity cost of losing the contract when hiring child labor, and thus reduces the supplier's incentive to hire child labor. Here we assume that the required premium is lower when the manufacturer chooses a high inspection level θ_H; i.e.,

$$\{1/(\gamma e(1-\theta_H)) - 1\}\Delta(\theta_H) < \{1/(\gamma e(1-\theta_L)) - 1\}\Delta(\theta_L), \tag{2.8}$$

which holds under (A3): $\theta_H \geq 1 - \dfrac{m}{(s_H - s_L + m)\gamma e}$. Assumption (A3) ensures that the supplier's expected profit from hiring child labor is lower when the manufacturer conducts inspections than when no inspections are undertaken. Otherwise the supplier would be more likely to hire child labor when the manufacturer chooses a higher inspection level, but this is unrealistic and thus not considered. Consequently, we obtain the supplier's best response as follows:

$$\begin{aligned}
&d^{pre}(\theta, w) = 1 \ \forall \theta \in \{\theta_L, \theta_H\} \text{ for } w \in \left[0, s_H + \tfrac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H)\right);\\
&d^{pre}(\theta_L, w) = 1 \text{ and } d^{pre}(\theta_H, w) = 0 \text{ for } w \in \left[s_H + \tfrac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H), s_H + \tfrac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L)\right);\\
&d^{pre}(\theta, w) = 0 \ \forall \theta \in \{\theta_L, \theta_H\} \text{ for } w \in \left[s_H + \tfrac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L), +\infty\right).
\end{aligned} \tag{2.9}$$

From above, we can observe that the manufacturer's internal inspections reduce the supplier's child labor employment when the wholesale price is neither too high nor too low (i.e., $s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H) \leq w < s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L)$). When the wholesale price is very low (i.e., $w < s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H)$), the supplier always employs child labor to lower his labor cost, no matter if the manufacturer conducts inspections or not. Conversely, when the wholesale price is very high (i.e., $w > s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L)$), high expected opportunity costs from contract termination always deter the supplier from hiring child labor.

We next derive the manufacturer's best response function $\theta^{pre}(d, w)$ to the supplier's decision on d for any given w. The manufacturer evaluates the difference in her expected profit between

[7] Without loss of generality, we assume that the supplier, if indifferent between hiring child labor and not, chooses not to hire child labor. This equilibrium is obtained if the manufacturer could simply offer an infinitesimal amount more than $s_H + \{1/(\gamma e(1-\theta)) - 1\}\Delta(\theta)$ to induce the supplier not to employ child labor.

choosing θ_H and choosing $\theta_L\,(=0)$, which is given as: $U\,(w,\theta_H,d)-U\,(w,\theta_L,d)\;=\;de\theta_H g-I$. Thus, for any given w, we obtain the manufacturer's best response function $\theta^{pre}\,(d,w)$ as follows:

$$\theta^{pre}\,(d=0,w)=\theta_L;\;\theta^{pre}\,(d=1,w)=\begin{cases}\theta_H & \text{if } I<e\theta_H g;\\ \theta_L & \text{if } I\geq e\theta_H g.\end{cases} \tag{2.10}$$

This can be interpreted as follows. Clearly, if the manufacturer knows that the supplier does not hire any child labor, then she will not conduct internal inspections (i.e., $\theta^{pre}\,(d=0,w)=\theta_L=0$). However, if the manufacturer knows that the supplier has hired child labor, she will conduct internal inspections (i.e., $\theta^{pre}\,(d=1,w)=\theta_H$) only when the inspection cost I is lower than the amount of expected goodwill loss reduced by inspections, $e\theta_H g$.[8]

By examining $d^{pre}\,(\theta,w)$ and $\theta^{pre}\,(d,w)$, we obtain the following fixed point $(\theta^{pre}\,(w)\,,d^{pre}\,(w))$ that satisfies $\theta^{pre}\,(w)=\theta^{pre}\,(d^{pre}\,(w)\,,w)$ and $d^{pre}\,(w)=d^{pre}\,(\theta^{pre}\,(w)\,,w)$:

$$(\theta^{pre}\,(w)\,,d^{pre}\,(w))=\begin{cases}(\theta_L,1) & \text{if } I\geq e\theta_H g \text{ and } w<s_H+\frac{\Delta(\theta_L)}{\gamma e}-\Delta\,(\theta_L)\,;\\ (\theta_L,0) & \text{if } w\geq s_H+\frac{\Delta(\theta_L)}{\gamma e}-\Delta\,(\theta_L)\,;\\ (\theta_H,1) & \text{if } I<e\theta_H g \text{ and } w<s_H+\frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)}-\Delta\,(\theta_H)\,.\end{cases} \tag{2.11}$$

2.4.2 Subgame-Perfect Equilibrium in the Pre-Act Scenario

In this section we study the manufacturer's decision on the wholesale price. The manufacturer determines a wholesale price w that maximizes her expected profit by solving the following program:

$$\max_{w}\;U\,(w,\theta^{pre}(w),d^{pre}(w))\,;\;\;s.t.\;(2.5). \tag{2.12}$$

The following proposition presents the subgame-perfect equilibrium outcome $(w^{pre},\theta^{pre},d^{pre})$:

[8]Without loss of generality, we assume that the manufacturer, if indifferent between conducting inspections and not, conducts no inspections. This does not affect the subgame-perfect equilibrium presented in Proposition 1 in §4.2 because the manufacturer always prefers to offer a wholesale price that induces $(\theta_L,1)$ rather than $(\theta_H,1)$.

Proposition 1 *The subgame-perfect equilibrium in the pre-Act scenario is:*

$$(w^{pre}, \theta^{pre}, d^{pre}) = \begin{cases} (s_H + (1/(\gamma e) - 1)\,\Delta(\theta_L), \theta_L, 0) & \text{if } \xi_3^{pre} \leq I < \xi_1^{pre} \text{ or } \{I \geq \xi_1^{pre} \text{ and } g \geq \xi_2^{pre}\}; \\ (s_H - \Delta(\theta_H), \theta_H, 1) & \text{if } I < \min\{\xi_1^{pre}, \xi_3^{pre}\}; \\ (s_H - \Delta(\theta_L), \theta_L, 1) & \text{otherwise,} \end{cases} \tag{2.13}$$

where $\xi_1^{pre} \equiv e\theta_H g$, $\xi_2^{pre} \equiv \Delta(\theta_L)/\gamma e^2$, *and* $\xi_3^{pre} \equiv (1/(\gamma e) - 1)\,\Delta(\theta_L) + \Delta(\theta_H) - ge(1 - \theta_H)$.

Proposition 1 presents three possible subgame-perfect equilibrium outcomes. The first outcome in (2.13) indicates that, to incentivize the supplier not to employ child labor (i.e., $d^{pre} = 0$), the manufacturer should adopt no inspections in equilibrium (i.e., $\theta^{pre} = \theta_L$), and offer the supplier the wholesale price $s_H + (1/(\gamma e) - 1)\,\Delta(\theta_L)$. This price is higher than his labor cost s_H, and as a result the supplier earns positive surplus. As indicated by our analysis in §2.4.1, such a price premium deters the supplier from hiring child labor (so that $d_E = d(1 - \theta) = 0$) by imposing high opportunity costs onto the supplier. In the other two equilibrium outcomes given in (2.13), however, no price premium is paid. In both outcomes, the manufacturer pays the wholesale price that covers only the supplier's labor cost of child labor (i.e., $w^{pre} = s_H - \Delta(\theta^{pre})$), and thus the supplier cannot but employ child labor (i.e., $d^{pre} = 1$) and obtain zero surplus. Whether or not the manufacturer undertakes internal inspections separates these two outcomes. When the manufacturer conducts inspections with θ_H, the second outcome in (2.13), $(w^{pre}, \theta^{pre}, d^{pre}) = (s_H - \Delta(\theta_H), \theta_H, 1)$, emerges in equilibrium. In this case, the child labor will be replaced with the adult labor with probability θ_H, and thus the average amount of child labor that remains (and can be potentially exposed to the public) becomes $d_E = 1 - \theta_H$. When the manufacturer chooses no inspections with θ_L, the third outcome in (2.13), $(w^{pre}, \theta^{pre}, d^{pre}) = (s_H - \Delta(\theta_L), \theta_L, 1)$, occurs in equilibrium. In this case, only the child labor is used in production so that $d_E = 1$.

To summarize, in the pre-Act scenario, the manufacturer chooses one of the following three strategies in equilibrium: pay a premium, conduct internal inspections, or do neither of the first two. In the rest of this paper, we refer to these strategies as "premium alone," "inspection alone," and "do-nothing" strategies, respectively. Note that $d_E = 0$ under the premium-alone strategy, $d_E = 1 - \theta_H$ under the inspection-alone strategy, and $d_E = 1$ under the do-nothing strategy. This suggests that the premium-alone strategy (resp., the do-nothing strategy) brings about the best

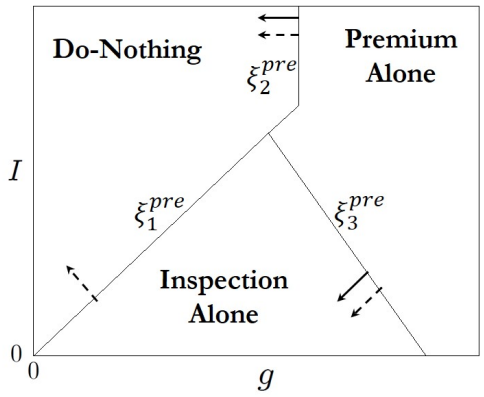

Figure 2.2: Equilibrium Outcomes in the Pre-Act Scenario. (*Note.* Solid arrows indicate how threshold lines change with γ, and dotted arrows indicate how threshold lines change with e.)

(resp., worst) outcome for child labor.

We next examine the factors that affect the equilibrium strategy. Figure 2.2 illustrates the three strategies divided by threshold lines ξ_1^{pre}, ξ_2^{pre} and ξ_3^{pre} defined in Proposition 1. First, we observe from Figure 2.2 that as the goodwill cost g increases, the manufacturer may change her strategy from do-nothing to premium-alone when the inspection cost I is high, or from do-nothing to inspection-alone and then to premium-alone when I is low. In both cases, as g increases, the expected amount of child labor d_E decreases (see discussion above). This is intuitive: a high goodwill cost incentivizes the manufacturer to combat child labor.

Second, as illustrated in Figure 2.2, as the manufacturer's inspection cost I decreases, the manufacturer may change her strategy from do-nothing to inspection-alone when the goodwill cost g is low, or from premium-alone to inspection-alone when g is high. While the former impact of I is intuitive, the latter impact of I is less so and merits some explanation. The latter impact implies that with a lower inspection cost I, the manufacturer may conduct internal inspections, while paying no premium to the supplier. This happens because by conducting inspections the manufacturer can lower the risk of exposing child labor, if any, to the public, and hence has less incentive to pay a premium to deter the supplier from hiring child labor. Thus, the price premium and internal inspections are used as strategic *substitutes* (as opposed to complements) in addressing the child labor issue. Due to this substitutability, as the inspection cost I decreases, the expected amount of child labor d_E may increase because the supplier is induced to hire child labor under

the inspection-alone strategy (although he is not under the premium-alone strategy), in spite of the high goodwill cost g. This finding implies that third-party organizations' increased support for internal monitoring may not always be effective and may even weaken the effect of public pressure exerted by these organizations. In the next section, we will see how supply chain transparency can help alleviate this adverse effect.

Third, solid arrows in Figure 2.2 illustrate that as the ratio of future orders to the current order, γ, increases, the threshold lines ξ_2^{pre} and ξ_3^{pre} move left. This means that when the value of future business is high, the manufacturer prefers the premium-alone strategy. The reason is as follows. With higher γ, the supplier incurs a larger loss from hiring child labor, leading to a lower price premium $(1/(\gamma e) - 1)\Delta(\theta_L)$ that the manufacturer pays to incentivize the supplier not to employ child labor. In this case, the premium-alone strategy enables the manufacturer to earn higher expected profits, and thus the manufacturer is more likely to use this strategy to combat child labor. This suggests that adding more value into business relations could help a manufacturer address her child labor issue. This finding is consistent with the practice of Nike: after suffering from a 69% fall in its earnings due to its scandal of severe labor rights violation in 1997, Nike instituted a new global value chain named "Future Vision," under which Nike provided its lead suppliers "an exclusive production relationship with guaranteed minimum monthly orders" (Lim and Phillips 2007).

Finally, dotted arrows in Figure 2.2 illustrate that as the external inspection level e increases, the threshold lines ξ_2^{pre} and ξ_3^{pre} move left as in the solid arrows; and furthermore the threshold line ξ_1^{pre} moves left as well. The former effect is the same as that of increasing γ, indicating that under a higher external inspection level e, the manufacturer is more likely to adopt the premium-alone strategy to combat child labor. This is easy to see from the price premium $(1/(\gamma e) - 1)\Delta(\theta_L)$, where γe appears in the denominator. The latter effect is unique to e, indicating that a higher e can also lead the manufacturer to choose the inspection-alone strategy to combat child labor. This is because under a higher external inspection level, the manufacturer faces a greater risk of losing consumer goodwill when child labor is employed, but internal inspections reduce the chance of child labor being exposed to the public. This result suggests that the increased external inspection level e can help reduce the expected amount of child labor d_E when the goodwill cost g is either low (by reducing the area where the do-nothing strategy is in equilibrium) or high (by expanding the area

where the premium-alone strategy is in equilibrium).

2.5 Equilibrium Analysis in the Post-Act Scenario

After the enactment of the Act, the manufacturer is required to publicly disclose her inspection policy, and thus can commit to her inspection level θ. After observing the manufacturer's decision on θ, the supplier determines his child labor decision on d. Throughout the paper we let the superscript "*post*" indicate results in the *post*-Act scenario. At the contract stage (S1), the manufacturer solves the following program to choose (w, θ) that maximizes her expected profit U:

$$\max_{w,\theta\in\{0,\theta_H\}} \quad U\left(w, \theta, d^{post}\left(\theta, w\right)\right) \tag{2.14}$$

$$s.t. \quad \Pi\left(w, \theta, d\right) \quad = w - s_H + \Delta\left(\theta\right) d^{post}(\theta, w) \geq 0 \tag{2.15}$$

$$d^{post}\left(\theta, w\right) \quad = \arg\max_{d\in\{0,1\}} \Pi\left(w, \theta, d\right) = \arg\max_{d\in\{0,1\}} \left\{1 - \gamma e\left(1 - \theta\right) d\right\}\left\{w - s_H + \Delta\left(\theta\right) d\right\}. \tag{2.16}$$

Constraint (2.15) ensures that the supplier earns non-negative expected profits. Constraint (2.16) requires that given the manufacturer's decisions on w and θ, the supplier chooses $d^{post}\left(\theta, w\right)$ to maximize his expected profit. In §2.5.1 we derive the subgame-perfect equilibrium of this program. In §2.5.2, by comparing the equilibrium in the post-Act scenario with that in the pre-Act scenario obtained in §2.4, we examine the impact of the Act on child labor.

2.5.1 Subgame-Perfect Equilibrium in the Post-Act Scenario

We first find the supplier's best response function $d^{post}\left(\theta, w\right)$ to the manufacturer's decision on (θ, w). Similar to the pre-Act scenario, the supplier determines his best response by comparing his expected profit when employing child labor with that when combating child labor. Thus the best response is the same as that in the pre-Act scenario; i.e., $d^{post}\left(\theta, w\right) = d^{pre}\left(\theta, w\right)$ given in (2.9). Next we substitute $d^{post}\left(\theta, w\right)$ into (2.14) and (2.15), and then determine the manufacturer's wholesale price w^{post} and inspection level θ^{post} by solving the following program:

$$\max_{w,\theta\in\{0,\theta_H\}} U\left(w, \theta, d^{post}\left(\theta, w\right)\right); s.t.\ (2.15). \tag{2.17}$$

Proposition 2 *The subgame-perfect equilibrium in the post-Act scenario is:*

$$(w^{post}, \theta^{post}, d^{post}) = \begin{cases} \left(s_H + \left(\frac{1}{\gamma e} - 1\right)\Delta(\theta_L), \theta_L, 0\right) & \text{if } g \geq \xi_2^{post}, I \geq \xi_3^{post}, I \geq \xi_5^{post}; \\ (s_H - \Delta(\theta_H), \theta_H, 1) & \text{if } I \leq \xi_1^{post}, I \leq \xi_3^{post}, g \leq \xi_4^{post}; \\ (s_H - \Delta(\theta_L), \theta_L, 1) & \text{if } I \geq \xi_1^{post}, g \leq \xi_2^{post}, I \geq \xi_6^{post}; \\ \left(s_H + \left\{\frac{1}{\gamma e(1-\theta_H)} - 1\right\}\Delta(\theta_H), \theta_H, 0\right) & \text{if } I \leq \xi_6^{post}, g \geq \xi_4^{post}, I \leq \xi_5^{post}, \end{cases} \tag{2.18}$$

where $\xi_1^{post} \equiv \Delta(\theta_H) - \Delta(\theta_L) + eg\theta_H$, $\xi_2^{post} \equiv \Delta(\theta_L)/(\gamma e^2)$, $\xi_3^{post} \equiv \Delta(\theta_H) + \Delta(\theta_L)/(\gamma e) - \Delta(\theta_L) - eg(1-\theta_H)$, $\xi_4^{post} \equiv \Delta(\theta_H)/\{\gamma e^2(1-\theta_H)^2\}$, $\xi_5^{post} \equiv \Delta(\theta_H) - \Delta(\theta_L) - \Delta(\theta_H)/\{\gamma e(1-\theta_H)\} + \Delta(\theta_L)/(\gamma e)$, *and* $\xi_6^{post} \equiv \Delta(\theta_H) + eg - \Delta(\theta_H)/\{\gamma e(1-\theta_H)\} - \Delta(\theta_L)$.

The first three equilibrium outcomes given in (2.18) correspond to the premium-alone, inspection-alone, and do-nothing strategies, respectively. In addition, the fourth equilibrium outcome shows a new "premium & inspection" strategy, under which the manufacturer conducts internal inspections (i.e., $\theta^{post} = \theta_H$) and at the same time pays a price premium (i.e., $w^{post} = s_H + \{1/(\gamma e(1-\theta_H)) - 1\}\Delta(\theta_H)$). Under this strategy, no child labor is used as in the premium-alone strategy (i.e., $d_E = 0$). Thus, compared to the inspection-alone strategy which induces child labor employment, the manufacturer does not face the risk of goodwill loss from child labor by paying a price premium to the supplier under the premium & inspection strategy. Moreover, compared to the premium-alone strategy, this strategy pays a lower price premium to the supplier by conducting internal inspections. This reflects the substitutability between the price premium and internal inspections discussed earlier in §2.4.2: the manufacturer substitutes some of the price premium with inspection efforts, thus leaving less surplus to the supplier.

To examine the factors that affect the equilibrium strategy, in Figure 2.3 we illustrate the four equilibrium strategies, which are divided by the threshold lines $\xi_1^{post}, \xi_2^{post}, ..., \xi_6^{post}$ defined in Proposition 2. Unlike the pre-Act scenario, the supplier's penalty of using child labor m (i.e., the amount of compensation the supplier pays to the child worker discovered during internal inspections) plays an important role. Specifically, depending on the value of m, three structures of the equilibrium outcome are possible as illustrated in Figure 2.3(a)-(c); see the proof of Proposition 2 in Appendix A for specific conditions on m. We can observe from Figure 2.3(a)-(c) that as m increases, the premium & inspection strategy is more likely to be in equilibrium, whereas the inspection-alone

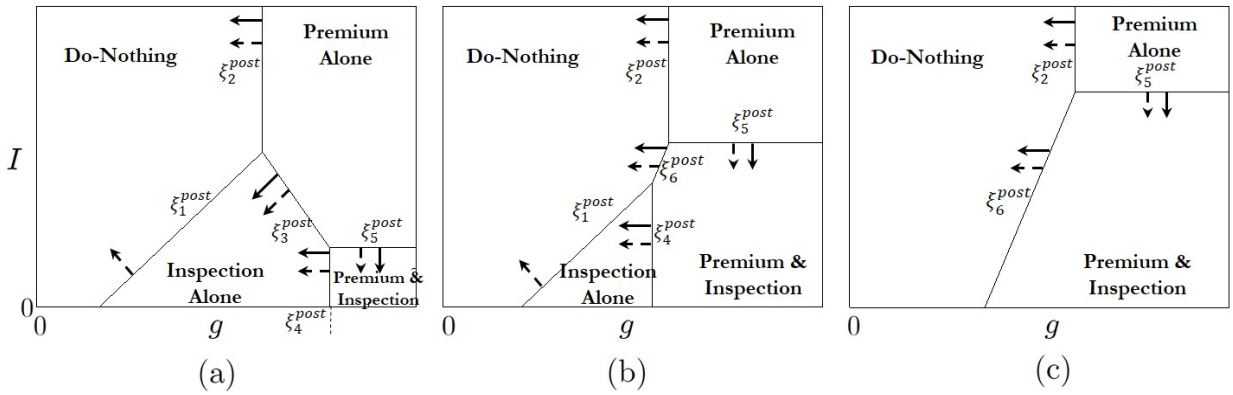

Figure 2.3: Equilibrium Outcome in the Post-Act Scenario with: (a) Low m, (b) Medium m, and (c) High m. (*Note*. Solid arrows indicate changes with γ, and dotted arrows indicate changes with e.)

strategy is less likely to be in equilibrium, and it is no longer in equilibrium when m is sufficiently large. The reason is as follows. When the manufacturer conducts internal inspections, an increase of the supplier's penalty m reduces the supplier's expected labor cost saving from using child labor (i.e., $\Delta(\theta_H)$ in (2.2) decreases with m). With the reduced benefit of using child labor, the manufacturer can deter the supplier from hiring child labor by paying a lower price premium (i.e., the price premium $(1/(\gamma e(1-\theta_H))-1)\Delta(\theta_H)$ decreases with m). Therefore, under the premium & inspection strategy, the manufacturer's expected profit increases with the penalty m. In contrast, under the inspection-alone strategy, an increase of m reduces $\Delta(\theta_H)$, which in turn increases the wholesale price $s_H - \Delta(\theta_H)$. This happens because under this strategy the supplier is induced to use child labor, and the manufacturer has to pay a higher wholesale price to compensate a loss in the supplier's expected labor cost saving from using child labor. Therefore, the manufacturer's expected profit decreases with the penalty m under the inspection-alone strategy. As a result, with a higher m, this strategy is less likely to be in equilibrium, whereas the premium & inspection strategy is more likely to be in equilibrium. This finding suggests that when the manufacturer can enforce a high penalty onto the supplier from internal inspections, the manufacturer can adopt the premium & inspection strategy and combat child labor more effectively.

It is not difficult to verify from Figures 2.2 and 2.3 that the impacts of g, γ and e on the expected amount of child labor d_E are the same in both pre-Act and post-Act scenarios. In contrast, the impact of inspection cost I on d_E differs in the post-Act scenario. Recall in the pre-Act scenario that

a lower inspection cost I may change the equilibrium strategy from premium-alone to inspection-alone, resulting in more child labor used on average. However, in the post-Act scenario, this is found only in Figure 2.3(a) for a smaller range of goodwill cost g (than that in Figure 2.2), but not in Figure 2.3(b)-(c). This suggests that in the post-Act scenario a lower inspection cost is less likely or even unlikely to result in more child labor and to weaken the effects of public pressure onto the manufacturer. This implies that enforcing supply chain transparency can help alleviate or even eliminate the adverse effect of a reduction in inspection cost. Therefore, if third-party organizations intend to combat child labor by providing support for firms' inspections but are concerned about the potential adverse effect, then they should consider pushing for more supply chain transparency on firms' inspection efforts.

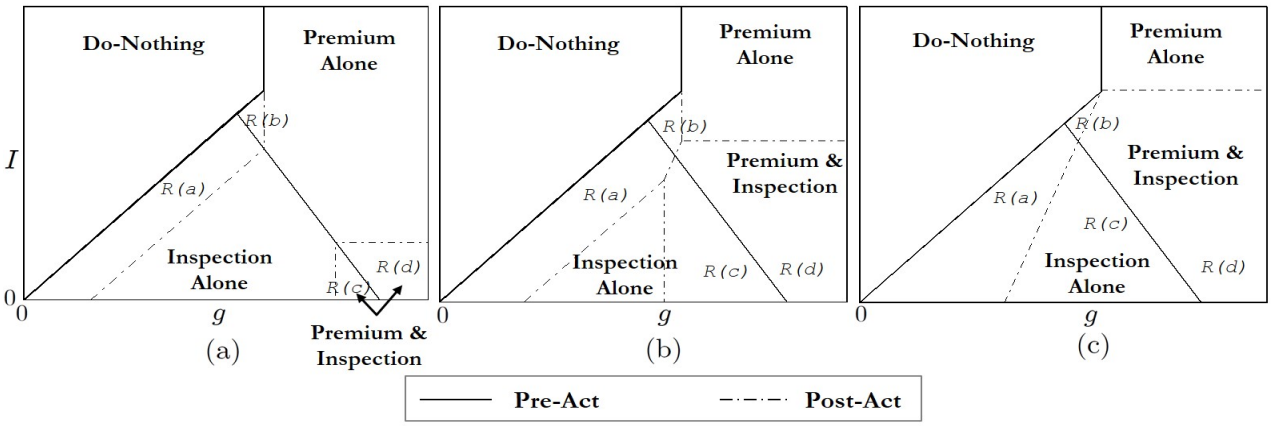

Figure 2.4: Comparison of Equilibrium Outcomes between the Pre-Act Scenario in Figure 2.2 and the Post-Act Scenario in Figure 2.3 with: (a) Low m, (b) Medium m, and (c) High m.

2.5.2 Effects of the Act

In this section we examine the effects of the Act by comparing the equilibrium outcomes in the pre-Act scenario with those in the post-Act scenario. Figure 2.4 which overlaps Figure 2.2 with Figure 2.3 illustrates that there are four areas in which the equilibrium outcomes are different between the two scenarios: $R(a)$, $R(b)$, $R(c)$ and $R(d)$. Table 2.1 provides precise conditions for each of these four areas.

Table 2.1: The Manufacturer's Equilibrium Strategy: the Pre-Act Scenario vs. the Post-Act Scenario

Area	Equilibrium Outcomes (Pre-Act $\rightarrow$ Post-Act)	Conditions
$R(a)$	Inspection Alone $\rightarrow$ Do-Nothing	$\max\left(\xi_1^{post}, \xi_6^{post}\right) \leq I \leq \min\left(\xi_1^{pre}, \xi_3^{post}\right)$
$R(b)$	Premium Alone $\rightarrow$ Do-Nothing	$\max\left(\xi_3^{post}, \xi_6^{post}\right) \leq I \leq \xi_1^{pre}$ and $g \leq \xi_2^{post}$
$R(c)$	Inspection Alone $\rightarrow$ Premium & Inspection	$I \leq \min\left(\xi_3^{post}, \xi_6^{post}\right)$ and $g \geq \xi_4^{post}$
$R(d)$	Premium Alone $\rightarrow$ Premium & Inspection	$\xi_3^{post} \leq I \leq \min\left(\xi_5^{post}, \xi_6^{post}\right)$

The comparison of the two scenarios reveals the following effects of the supply chain transparency Act on the manufacturer's profit, the supplier's profit, and child labor:

Proposition 3 *The Act will lead to the following results, ceteris paribus:*
(a) The supplier's profit remains the same in $R(a)$, decreases in $R(b)$ or $R(d)$, and increases in $R(c)$.
(b) The manufacturer's profit increases in all areas.
(c) The expected amount of child labor d_E increases in $R(a)$ or $R(b)$, decreases in $R(c)$, and remains the same in $R(d)$.

Proposition 3(a) indicates that the effect of the Act on the supplier's profit varies depending on how the Act changes the manufacturer's equilibrium strategy in each area. When the manufacturer pays no premium in $R(a)$ before and after the Act, the supplier's profit remains the same. On the other hand, when the manufacturer eliminates the premium after the Act in $R(b)$ or $R(d)$, the Act hurts the supplier's profit. Finally, when the manufacturer changes her strategy from paying no premium into paying a premium in $R(c)$, the supplier benefits from the Act.

Contrary to the effect of the Act on the supplier, Proposition 3(b) states that the Act always benefits the manufacturer. This is because the Act provides the manufacturer with the ability to credibly commit to her inspection level at the contract stage (S1) (see Figure 2.1). This can be easily seen by inspecting program (2.4)-(2.7) which includes an additional (incentive) constraint (2.6) with respect to the manufacturer's inspection level as compared with program (2.14)-(2.16).

Most importantly, Proposition 3(c) suggests that the Act may not necessarily reduce child labor

in supply chains. Specifically, Table 2.1 shows that the Act induces the manufacturer to change her equilibrium strategy from the inspection-alone strategy in $R(a)$ or the premium-alone strategy in $R(b)$ (both of which combat child labor) to the do-nothing strategy (which takes advantage of low-cost child labor). This change to the do-nothing strategy occurs when the manufacturer faces a low goodwill cost (i.e., $g \leq \xi_2^{post}$) and has a low inspection cost (i.e., $I \leq \xi_1^{pre}$). Prior to the Act, however, the manufacturer could not adopt the do-nothing strategy in $R(a)$ or $R(b)$ because this strategy is not credible to the supplier in the absence of a commitment device. In this case, the supplier believes that the manufacturer would deviate from the do-nothing strategy to the inspection-alone strategy after production starts, because the manufacturer could afford to conduct internal inspections in order to reduce the chance of child labor being exposed to the public and thereby reduce the risk of a goodwill loss.[9]

Proposition 3(c) also shows that the Act indeed reduces the use of child labor in $R(c)$. In this case, driven by a high goodwill cost and a low inspection cost, the manufacturer has a strong incentive to commit to combating child labor by using both inspection and premium after the Act. However, in the pre-Act scenario, such commitment is not credible because once the supplier accepts the contract and decides not to employ child labor, the manufacturer would recognize that it is unnecessary to conduct inspections even when the inspection cost is low.

The above analysis indicates that the Act may lead to unexpected consequences. The ability of the manufacturer to commit to her inspection policy under the Act not only helps the manufacturer combat child labor through the premium & inspection strategy, but also enables the manufacturer to exploit child labor more effectively by committing to the do-nothing strategy in certain situations. This finding cautions policy makers and third-party organizations that they should pay extra attention to the conditions under which the Act may backfire, and that they should try to minimize such adverse effects before pushing for similar transparency acts or legislations.[10]

In order to alleviate the adverse effects of supply chain transparency, one needs to reduce

[9]Prior to the Act, if the manufacturer could afford to conduct internal inspections, one may wonder why she would not commit to the inspection-alone strategy in the first place. This is because if committing to the inspection-alone strategy rather than the do-nothing strategy before production starts, the manufacturer needs to pay a higher wholesale price to recover the supplier's increased labor cost in expectation. In contrast, by pretending to adopt the do-nothing strategy before production starts but then deviating to the inspection-alone strategy afterwards, the manufacturer can not only pay a low wholesale price, but also reduce the risk of a goodwill loss.

[10]For example, a national bill entitled "the Business Transparency on Trafficking and Slavery Act" (H.R. 2759) was introduced in the House of Representative in 2011.

the manufacturer's incentive to employ child labor through the do-nothing strategy in the two areas $R(a)$ and $R(b)$. There are several intuitive measures that can help achieve this goal. First, increasing the goodwill cost g by arousing more consumer awareness or calling for more consumer boycotting will be helpful because this will increase the manufacturer's potential loss from the incidence of child labor. Second, if third-party organizations can help the manufacturer to reduce the inspection cost I by facilitating industry collaboration or improving a monitoring system (see §1), then the manufacturer would be more willing to conduct inspections to combat child labor. One counter-intuitive measure is to have a manufacturer freely choose the supplier's penalty of employing child labor, which will be discussed in §2.6.2.

2.6 Extensions

In §2.6.1 and §2.6.2, we extend our analysis by incorporating two other possible measures of third-party organizations, respectively: one is to educate consumers to pay more for a product sold by a socially-responsible manufacturer, and the other is to allow a manufacturer to endogenously determine the value of compensation paid by the supplier to a child laborer detected during internal inspections instead of enforcing a fixed one based on local living and education costs. In §2.6.3 we consider various penalty schemes for the supplier who employs child labor.

2.6.1 Socially Conscious Consumers

The Act informs consumers (as well as suppliers) of firms' efforts in combating child labor in their supply chains. As a result, some consumers may have higher valuations on a product sold by a manufacturer who conducts internal inspections, and they may be willing to pay a higher price for such a product. Specifically, after the introduction of the Act, suppose consumers' valuation increases from a fixed value v to $v' \equiv v + \delta\theta$, where $\delta \geq 0$ represents consumers' willingness to pay for the manufacturer's inspection effort level θ.[11]

It is easy to see that this increase of consumers' valuation creates additional incentives for the manufacturer to conduct inspections. As a result, the areas for the inspection-alone strategy and the

[11]For simplicity, as in Krishna and Rajan (2009), we assume that all consumers are willing to pay a higher price for a product sold by a socially-responsible manufacturer. In practice, only a portion of consumers may be socially conscious. This may be modeled by introducing a fixed ratio of socially conscious consumers as in Guo et al. (2014). Our results remain unchanged for this case.

premium & inspection strategy expand in Figure 2.3; specifically, we can show that four thresholds ξ_1^{post}, ξ_3^{post}, ξ_5^{post} and ξ_6^{post} are increased by $\delta\theta_H$. This result has two implications. On one hand, it helps to *mitigate* the adverse effect of the Act by discouraging the manufacturer from switching from the inspection-alone or premium-alone strategy to the do-nothing strategy (i.e., in Figure 2.4, the total area of $R(a)$ and $R(b)$, defined by $\max\left(\xi_1^{post}, \xi_3^{post}, \xi_6^{post}\right) \leq I \leq \xi_1^{pre}$ and $g \leq \xi_2^{post}$, shrinks as ξ_1^{post} and ξ_6^{post} increase). On the other hand, it can inadvertently introduce more child labor by inducing the manufacturer to rely more on imperfect inspections than to pay the price premium that deters the employment of child labor (i.e., in Figure 2.3(a), as ξ_3^{post} increases, the area for the inspection-alone strategy expands, whereas the area for the premium-alone strategy shrinks). Nevertheless, this issue would be resolved under a sufficiently high penalty m (i.e., ξ_3^{post} does not exist in Figure 2.3(b)-(c) where m is sufficiently high). This is because when the manufacturer can enforce a high penalty onto the supplier employing child labor, she is more likely to adopt the premium & inspection strategy rather than the inspection-alone strategy (see our earlier discussion in §2.5.1). Therefore, the Act will be more effective with socially-conscious consumers when the manufacturer can enforce a high penalty on the supplier's non-compliance.

2.6.2 Decision of Compensation in Corrective Actions

In our base model, when a child laborer is found during internal inspections, the supplier is required to pay compensation m to the child laborer. In this subsection, we consider the case when the manufacturer can choose the amount of m between low m_L and high m_H, where $0 \leq m_L < m_H$. We assume that the manufacturer makes this decision when determining inspection level θ. The choice of m affects the supplier's labor cost saving from using child labor $\Delta(\theta)$ given in (2.2), which is now replaced with $\Delta(\theta, m)$. Consequently, subgame-perfect equilibria in both pre-Act and post-Act scenarios involve the manufacturer's decision on m. Below we discuss the main implications of this extension, while presenting details in the online supplement.

In the pre-Act scenario, we show that the manufacturer still chooses one of the three strategies in equilibrium – do-nothing, inspection-alone or premium-alone. When the manufacturer conducts no inspections (i.e., $\theta = \theta_L$) under the do-nothing strategy or the premium-alone strategy, the manufacturer is obviously indifferent between the two values m_L and m_H because without internal inspections there will be no chance that the supplier needs to pay the compensation. When the

manufacturer chooses to conduct inspections (i.e., $\theta = \theta_H$) under the inspection-alone strategy, the manufacturer prefers low compensation m_L to m_H because this lowers the supplier's expected cost of employing child labor. This will allow the manufacturer to pay a lower wholesale price to the supplier, and hence induce some manufacturer to switch from the premium-alone strategy to the inspection-alone strategy (i.e., lower m increases ξ_3^{pre} in Figure 2.2). Because inspections are less effective in reducing child labor than the premium approach (see §2.4.2), this change leads to more child labor used in supply chains. Therefore, without supply chain transparency Act, third-party organizations should try to enforce the industry to pay a fixed amount of compensation paid to child laborers found, or at least set a minimum amount based on local living and education costs.

In the post-Act scenario, we find that the manufacturer still chooses low compensation m_L under the inspection-alone strategy (because her expected profit decreases with m under this strategy as discussed in §2.5.1). However, contrary to the pre-Act scenario, this may not necessarily increase the use of child labor because some manufacturer under low external pressure would change from the do-nothing strategy to the inspection-alone strategy (i.e., lower m increases ξ_1^{post} in Figure 2.3). In addition, we find that the manufacturer chooses high compensation m_H under the premium & inspection strategy because the manufacturer's expected profit increases with m under this strategy (see §2.5.1). This will incentivize some manufacturer to switch from the do-nothing strategy to the premium & inspection strategy, and thereby reduce the use of child labor (i.e., higher m increases ξ_6^{post} in Figure 2.3). Ultimately, this mitigates the adverse effect of the Act on child labor (i.e., the increase of ξ_1^{post} and ξ_6^{post} reduces the total area of $R(a)$ and $R(b)$ in Figure 2.4). Therefore, with the Act in place, third-party organizations may allow manufacturers to freely determine the compensation to child workers because such flexibility enables manufacturers to use internal inspections more effectively.

2.6.3 Restricted Penalty Schemes

This subsection considers two other plausible penalty schemes. Under the first scheme, the manufacturer requires the supplier to carry out only corrective actions (without terminating contracts) when child labor is detected during internal or external inspections. This scheme may be used in practice when it is difficult for a manufacturer to find a substitute supplier or when a manu-

facturer is eager to help fix the root cause of a child labor problem.[12] Under the second scheme, which is sometimes called a "zero-tolerance" policy, the manufacturer terminates contracts with the supplier when child labor is detected during internal inspections as well as during external inspections.[13] Below we discuss main results under each scheme, while providing more details in the online supplement.

Under the penalty scheme that requires only corrective actions, the supplier's expected cost saving from hiring child labor is revised from (2.2) to: $\Delta(\theta) = (s_H - s_L)(1-\theta)(1-e) - m\{\theta + (1-\theta)e\}$. In this case, our analysis shows that the supplier's decision on child labor is simplified as follows: the supplier employs child labor as long as the supplier can save his labor cost in expectation from hiring child labor (i.e., $\Delta(\theta) > 0$). This is because when child labor is found during internal or external inspections, the supplier will incur only one-time costs of hiring an adult worker and paying compensation to a child worker without losing future profits. Thus, under this scheme, the manufacturer can influence the supplier's decision only by her decision on internal inspections, but not by the wholesale-price contract.

Under the zero-tolerance penalty scheme, whenever a child laborer is found, the manufacturer will terminate her business with the supplier without offering the supplier a chance to correct its malpractice. Thus, the supplier's labor cost is s_H if $d = 0$ or s_L if $d = 1$, and consequently his cost saving from using child labor is simply $\Delta(\theta) = s_H - s_L$ for any $\theta \in \{\theta_L, \theta_H\}$. We obtain the following results under this scheme. First, as expected, this new scheme has no impact on do-nothing and premium-alone strategies, since the manufacturer does not conduct internal inspections under these strategies. (Recall that this scheme differs from the base scheme in that the manufacturer terminates contracts, instead of requiring corrective actions, when child labor is found during internal inspections.) Second, under the inspection-alone strategy, the equilibrium wholesale price becomes $s_H - \Delta(\theta_H) = s_H - (s_H - s_L) = s_L$, which is lower than the wholesale price in the base model, while being equal to the wholesale price under the do-nothing strategy. This is because under the new scheme the supplier will not have a chance to replace a child laborer

[12] For example, even after the news about children' illegal mining is released to the public, Apple continued to source tin from Indonesia because unilateral refusal of any tin from Indonesian mines may not improve the situation (Bilton 2014).

[13] For example, Samsung has a zero-tolerance policy on child labor although there was an exception when child workers were found to have used forged identification (BBC 2014). Similarly, Walmart adopts a zero-tolerance policy on subcontracting after tragic Bangladesh factory fire (D'Innocenzio 2013).

with an adult laborer after internal inspections, so the supplier's labor cost is the same under both inspection-alone and do-nothing strategies. As a result of this change in the wholesale price, the adverse effect of the Act (that introduces more child labor by inducing the manufacturer to switch from the inspection-alone strategy or the premium-alone strategy to the do-nothing strategy) does not occur under this scheme; i.e., $R(a)$ and $R(b)$ do not exist in Figure 4. This implies that the supply chain transparency Act will help reduce child labor in supply chains under the zero-tolerance policy. Lastly, we find that under the premium & inspection strategy, the wholesale price can be higher or lower than that in the base model. As a result, the new scheme does not have unambiguous effects on the area under which the manufacturer chooses this strategy in equilibrium (i.e., $R(c)$ or $R(d)$ can be larger or smaller than the respective area in the base model).

2.7 Conclusion

This paper studies child labor in a global supply chain. It models a multinational firm in a developed nation selling a product made by a supplier in a developing country where child labor is prevalent. The firm deals with child labor through her supply contract and internal inspections. We analyze the firm's strategies to mitigate her supplier's use of child labor, and examine how the firm's incentives to use these strategies are affected by various measures of third party organizations. These measures include imposing public pressure, providing support towards internal inspections, and most importantly, promoting legislation that enforces supply chain transparency on firms' inspection efforts.

We find that while imposing public pressure onto global manufacturers can effectively reduce the use of child labor in global supply chains, it may be compromised by other measures that provide support for internal inspections. Fortunately, this may be alleviated by the presence of supply chain transparency. On the other hand, the supply chain transparency Act may inadvertently introduce more child labor by enabling a manufacturer to credibly commit to undertaking no inspections when public pressure is low. We further find that this adverse effect can be eliminated under a zero-tolerance policy, or at least mitigated by additional measures such as calling for consumers to pay more for products of firms that have disclosed significant inspection efforts or allowing firms to choose the amount of compensation paid to child labor found during internal inspections.

Chapter 3

The Benefit of Introducing Variability in Quality Based Service Domains

3.1 Introduction

In this paper we propose a static service differentiation policy for a single-server queueing system, serving a group of homogeneous customers. Our policy randomly places customers into different service grades with different service rates. Few researches have considered such a policy, because conventional wisdom held that this sort of differentiation would harm system performance by increasing service time variance, which increases waiting time for a First-Come-First-Served single-server queue. (For Poisson arrivals this is a result of the classic P-K formula: $E[W] = \frac{\lambda E[S^2]}{2(1-\rho)}$.) However, we show that this conventional wisdom is *false*: We prove that despite this increased variance, a centralized single-server system can benefit from such a static service differentiation policy that randomly assigns customers different service rates independent of system state.

Our static differentiation policy works as follows: We modify the original identically distributed jobs into stochastically shorter and longer groups with different grades, initially holding the mean service time constant. Such differentiation increases service time variability, and thus increases the mean excess of work in service, lengthening waiting. However, this differentiation also creates service rate information, enabling the implementation of SEPT (shortest expected processing time) scheduling, reducing the expected waiting time for jobs in queue. When correctly implemented, the

contribution of the SEPT rule outweighs the impact of increased variability, and the total expected waiting time decreases. Furthermore, the more grades we introduce, the greater the reduction.

This benefit of reducing waiting time without changing the mean service time can be applied in various service areas. We demonstrate the use of such a differentiation policy in the context of *quality-based service* domains, in which service quality is directly related to the length of service time, with customers gaining more value from longer service. Typical examples include primary health care, call centers, consulting, education, personal care and repair services (Alizamir *et al.* 2013, Wang *et al.* 2010, 2012, Ren and Zhou 2008). When choosing service rates in quality-based settings, the provider faces a *speed-quality* tradeoff between customers' service value and waiting cost: reducing the service rate increases service value, but lengthens customer waiting. Our static service differentiation policy can benefit a quality-based system by reducing the total waiting without affecting the total service value (by keeping the mean service time constant).

There are two additional advantages of applying our differentiation policy in quality-based domains. First, in quality-based domains, the mean service time is allowed to vary. So, in addition to possibly reducing waiting time, the service provider could further increase service value by adjusting the aggregate average service time and thus generate still greater improvement. We consider this possibility as well. Second, our analysis shows that with static differentiation, customers who are assigned different service rates will have to wait different amounts of time. It turns out that longer waiting times are assigned to customers with slower service, so those customers who endure longer waiting are compensated with greater service value. From this perspective, the quality-based domain is one in which the static differentiation technique likely can be applied most successfully.

Note that our static differentiation policy is different from dynamic service rate control, which chooses the service rate based on the system state, following the basic idea of speeding up when the system is busy and slowing down when it is not (e.g., George and Harrison 2001). Such policies may be impractical in some settings, due to the effort/cost/time needed to obtain system state information (Heyman 1977, Harchol-Balter *et al.* 2003, Moreno 2009, Wang, Yang and Pearn 2010). In other settings, consumers may contract for service rates and priorities before they (randomly) arrive at the system. In this case a state-dependent policy cannot be utilized, and a system-independent service rate control policy is again required.

To summarize, our paper makes four primary contributions.

1. We prove that offering non-uniform service rates to homogeneous jobs can reduce average waiting time compared with uniform service, assuming both policies maintain the same aggregate average service time (i.e., the same system load). The superiority of differentiated service holds even in static (state-independent) and classical (service value independent) settings.

2. We provide the structure of the optimal policies for service differentiation into any fixed number of service grades. First for each grade we identify the optimal service rate and the corresponding static allocation probabilities of customers, which minimize the total amount of waiting time when system load is fixed. We find that both vectors — of service rates and allocation probabilities — form geometric sequences, with correlated ratios that depend on the system load. We then derive the optimal system load and the corresponding optimal system performance, which depend on customer characteristics such as marginal service value and marginal waiting cost, but are independent of customers' arrival rate.

3. We compute the asymptotically optimal system performance as the number of service grades grows, numerically providing a bound on the benefits of service rate differentiation ($\sim$5%).

4. We numerically explore the sensitivity of the benefit of service rate differentiation with respect to customers' characteristics such as marginal waiting cost and job size variation. This highlights the settings in which service rate differentiation is most likely to be valuable.

Our paper is organized as follows: In section 2 we review the related literature. In section 3 we introduce our queueing model and our policy of service rate differentiation. Section 4 presents all the analytical results, including the dominance of service rate differentiation, closed-form solutions for the optimal differentiation policies, and the asymptotically optimal system performance. We then perform a numerical sensitivity analysis in section 5, which illustrates the value of service rate differentiation in various settings. We conclude and discuss potential extensions in section 3.6.

3.2 Related Literature

Our work lies at the intersection of service rate control and service differentiation.

Service Rate Control. There is a rich body of research studying the characterization and computation of optimal service rate policies for queueing systems. Most papers aim to minimize waiting and capacity costs by dynamically changing service rates based on system state, beginning with Crabill (1972, 1974), and followed by papers including Stidham and Weber (1989), George and Harrison (2001), Ata and Shneorson (2006), and Adusumilli and Hasenbein (2010). Others aim to find a balance between waiting time and service value which increases with service time, including Hopp *et al.* (2007), Anand *et al.* (2011), Kostami and Rajagopalan (2013). All of the above papers adopt either a state-dependent control strategy or a uniform service rate. Our paper uses a totally different static service rate differentiation policy.

Service Differentiation. The common definition of service differentiation refers to distinguishing heterogenous customers and offering corresponding services. Prior literature primarily studied whether and how to assign different types of customers to different service grades. Most papers design policies for a decentralized system to induce strategic consumers to self-select the centralized optimal choices. Examples include Mendelson and Whang (1990), Rao and Petersen (1998) and van Mieghem (2000). None of these papers mentions offering multiple service grades to the same type of customers, as our model does.

Another related research stream focuses on large-scale service systems, where service level differentiation is obtained by server allocation and scheduling. Among them, Gurvich *et al.* (2008) and Gurvich and Whitt (2010) both enforce service-level constraints for heterogeneous customers. Armony and Mandelbaum (2011), considers homogeneous customers, studying how to match these customers with heterogeneous servers to asymptotically optimize system performance. Armony (2005) shows the benefits of heterogeneity in the many-server heavy-traffic regime: a large-scale service system with heterogeneous servers under the Faster Service First (FSF) policy generates a lower waiting time than a homogenous server system, when both systems have the same total capacity. The intuition is that the FSF policy uses heterogeneous servers more efficiently, which is similar to the intuition behind our result. But we prove the benefits of service time variance through a single server system without the heavy-traffic assumption.

The benefit of high variability is also found in Lin *et al.* (2011), which shows, again in the heavy traffic regime, under preemptive SRPT (shortest remaining processing time) scheduling, the growth rate of the average waiting time is much smaller if the service time distribution is unbounded

(compared to bounded job size). Again, our result doesn't depend on the heavy-traffic assumption.

In conclusion, to our knowledge, there is no prior research applying service grade differentiation with homogeneous customers in a static setting. This is possibly because it was assumed, erroneously, that such differentiation could only be detrimental.

3.3 Our Queueing Model

Classic queueing models typically choose a uniform service rate, or less commonly dynamically adjust service rates according to system state. Our model departs from these by permitting a single server to offer a set of discrete service rates, which correspond to different service grades within a static setting: Each arriving job or request is immediately assigned to a service grade with a corresponding probability, which is completely independent of system state.

In this section we introduce the basic elements of our model. We first explain our method of service rate differentiation, then our service rate allocation policies and job scheduling rules. We lastly describe customers' characteristics and performance metrics of a specific application of our policy within a quality-based service system.

Service Rate Differentiation. In order to model variable service rates we must first specify the relationship between the distributions of the differentiated (new) services and that of the original (base) service requirement. In some literature related to service time control, the service provider adopts a dynamic policy by which the job in service can be released (deemed complete) at any moment (in some papers subject to a minimum service requirement, e.g. Hopp *et al.* 2007). Thus in these settings the original distribution of service time is truncated. In other literature different service grades have the same base service distribution, which can be scaled according to differentiated service rates (i.e. by working faster or slower). An example is Debo *et al.* (2008). We also use this scaling approach for differentiation (but do not restrict ourselves to exponentially distributed service times). Outside of scaling we do not consider altering the original service distribution form.

The arriving jobs' original processing time is an i.i.d. random variable denoted by X_o, with a cumulative density function $F(\cdot)$; we let $\mu_o = 1/E[X_o] \in \Re^+$. We consider $1 \leq K \leq \infty$ different service grades, which are indicated by the subscript $k = 1, ..., K$. The service provider determines

the service rates of the K grades, denoted by the vector $\overrightarrow{\mu} = (\mu_1, ..., \mu_k, ..., \mu_K)$. Without loss of generality, we assume that $\mu_1 > ...\mu_k > ...\mu_K$. According to our scaling method, the corresponding service time distribution of jobs in service grade k, denoted by X_k, satisfies $X_k \sim \frac{\mu_o}{\mu_k} X_o$. Thus the average service time in grade k is $E[X_k] = \frac{\mu_o E[X_o]}{\mu_k} = \frac{1}{\mu_k}$.

We argue that such a scaling method is often very reasonable: A job's properties, including its processing time distribution, may be inherently determined. In such a situation the service provider is able to accelerate or slow down the service procedure by adjusting its efforts, but cannot change the entire service time distribution.

Rate Allocation and Job Scheduling Rule. With the designation of different service grades, the originally homogeneous jobs are artificially divided into different "types" receiving different service rates. In our model, the assignment of arriving jobs to different grades follows a static (time-invariant and system state independent) *resource allocation rule:* when a job arrives at the system, the server assigns it to the kth service grade with probability p_k; thus the allocation decision is specified by the K-dimensional probability vector $\overrightarrow{p} = (p_1, ..., p_k, ..., p_K)$ where $\sum_{k=1}^{K} p_k = 1$.

The server interacts with tasks based on a static and non-preemptive[1] *scheduling rule*, denoted by r. The scheduling rule r determines the service sequences of different service grades. Within each grade, without loss of generality, we assume FCFS is adopted. We also assume work conservation: Whenever there is a job waiting, the server is working; no unnecessary idling is allowed.

Jobs. Homogenous jobs arrive according to a Poisson process with rate λ. Therefore, the jobs comprising the kth service grade also form a Poisson process with an arrival rate $\lambda_k = \lambda p_k$. The arrival rates for the K service grades are represented by the vector $\overrightarrow{\lambda} = \lambda \overrightarrow{p} = (\lambda_1, ..., \lambda_k, ..., \lambda_K)$. Thus the system forms an $M/G/1$ queue with arrival rate λ and service time X, with distribution:

$$X \sim \begin{cases} X_1 = \frac{\mu_o}{\mu_1} X_o \text{ w.p. } p_1; \\ ...; \\ X_K = \frac{\mu_o}{\mu_K} X_o \text{ w.p. } p_K. \end{cases} \tag{3.1}$$

where $E[X] = \sum_{k=1}^{K} \frac{\mu_o E[X_o] p_k}{\mu_k} = \sum_{k=1}^{K} \frac{p_k}{\mu_k}$.

[1]Unless otherwise noted, all scheduling rules considered in this paper are non-preemptive. The adoption of a preemptive scheduling rule is discussed as an extension in Section 3.6.2.

Performance Metrics. All jobs suffer from waiting time in queue, and within a quality-based domain they may also gain value from their length of service. The gain and the loss are characterized by two functions, respectively: a non-decreasing service value function $G(X(j))$, and a strictly increasing delay cost function $C(W(j))$, where $X(j)$ and $W(j)$ are the service time and waiting time of a generic job j. We assume that both G and C are linear;[2] and we consequently define $G(X(j)) = u \cdot (X(j))$ and $C(W(j)) = h \cdot (W(j))$, where $u \geq 0$ and $h > 0$. Therefore, the corresponding expected values are $E[G(X(j))] = uE[X(j)]$ and $E[C(W(j))] = hE[W(j)]$. We denote by W_k the generic waiting time random variable of a job in grade k and define $w_k := E[W_k]$. Thus $E[G(X_k)] = uE[X_k] = u/\mu_k$ and $E[C(W_k)] = hE[W_k] = hw_k$.

Model. We define system performance as the net difference between average service value and average waiting cost per unit time, i.e., $V = \lambda\left(\sum_{k=1}^{K} p_k E[G(X_k)] - \sum_{k=1}^{K} p_k E[C(W_k)]\right)$. Based on our linearity assumptions, the system performance can be represented as:

$$V(\overrightarrow{\mu}, \overrightarrow{p}, r) = \lambda(u\sum_{k=1}^{K} p_k/\mu_k - h\sum_{k=1}^{K} p_k w_k(\overrightarrow{\mu}, \overrightarrow{p}, r)). \tag{3.2}$$

The server aims to maximize system performance by choosing service strategies consisting of a service rate vector $\overrightarrow{\mu} = (\mu_1, ..., \mu_k, ..., \mu_K)$, a grade allocation vector $\overrightarrow{p} = (p_1, ..., p_k, ..., p_K)$ and a scheduling rule, r. The average waiting time w_k is determined by $\overrightarrow{\mu}$, $\overrightarrow{p}$ and r, so we rewrite w_k as $w_k = w_k(\overrightarrow{\mu}, \overrightarrow{p}, r)$. Also note that the service strategy affects the service value—the first term in (3.2)—only through the aggregate average service time $E[X] = \sum_{k=1}^{K} p_k/\mu_k$.

3.4 Analytical Results

In this section we specify the optimal service strategies for our model. First, in section 4.1, we derive the optimal scheduling rule, which plays a decisive role in the characterization of the average waiting time. Although we have exogenously defined a service rate differentiation model, the advisability of offering multiple service rates has not yet been demonstrated. Hence, in section 4.2, we prove the benefits of service differentiation by showing how the average waiting in a system may decrease with the number of service grades K increasing while keeping the aggregate average service time

[2]Nonlinear service value and waiting cost functions are discussed as an extension in Section 3.6.2.

constant. In section 4.3, for a given number of service grades K, we first derive the optimal service rates and service allocation probabilities that minimize the average waiting for a given system load. We then solve for the optimal system load as well as the optimal system performance. Finally in section 4.4, we give the asymptotic performance as K goes to infinity.

For convenience, we define a service policy with only one service grade ($K = 1$) as *pure service*; we define *mixed service* ($K > 1$) as a policy offering more than one service grade. We will demonstrate the benefits of service rate differentiation by comparing mixed service with pure service.

3.4.1 Optimal Scheduling Rule

As in our system jobs are artificially differentiated into multiple grades, a scheduling rule can either make use of the grade information (*grade-based* policies) or not (*non-grade-based* policies). Before we derive the optimal scheduling rule — which will be proved to be grade-based — for comparison purposes we first show the optimal service rates and allocation probabilities under any non-grade-based scheduling rule. We will find that these two types of scheduling rules — grade-based and non-grade-based — produce completely opposite results about the optimal service policies. All proofs not included in the text are provided in the appendix.

Proposition 4 *Under any non-grade-based scheduling rule, the optimal service strategy is always pure service.*

Proposition 4 confirms the general intuition that service time variation degrades system performance. However, later we will show that the optimal scheduling rule is grade-based, under which mixed service becomes superior to pure service.

Our maximization problem contains three groups of decision variables: the service rate vector $\overrightarrow{\mu} = (\mu_1, ..., \mu_k, ..., \mu_K)$, the service allocation rule $\overrightarrow{p} = (p_1, ..., p_k, ..., p_K)$ and the scheduling rule r:

$$\max_{\overrightarrow{\mu}, \overrightarrow{p}, r} V = \lambda(u \sum_{k=1}^{K} p_k/\mu_k - h \sum_{k=1}^{K} p_k w_k(\overrightarrow{\mu}, \overrightarrow{p}, r)), \tag{3.3}$$

where the first term in the parentheses represents the expected service value and the second term represents the expected delay cost.

From (3.3) we can observe that the scheduling rule r affects the system performance only through the delay cost. Thus the maximization problem can be solved in two steps:

$$\max_{\overrightarrow{\mu},\overrightarrow{p}} V = \lambda(u\sum_{k=1}^{K} p_k/\mu_k - \min_{r|\overrightarrow{\mu},\overrightarrow{p}} h\sum_{k=1}^{K} p_k w_k(\overrightarrow{\mu},\overrightarrow{p},r)).$$

Hence the optimal scheduling rule should minimize the aggregate average waiting time of all K service grades with given service rates $\overrightarrow{\mu}$ and grade allocation probabilities $\overrightarrow{p}$. Such a rule is trivial to find as research about this topic is quite complete (Wolff 1989, Chapter 5):[3]

Lemma 1 *The optimal scheduling rule r^* is Shortest Expected Processing Time (SEPT) First, which minimizes the total waiting time, as well as the total delay cost in model (3.3).*

Under the SEPT scheduling rule, scheduling priority decreases with the grade's mean service time, or, increases with the grade's service rate. Based on our assumption that $\mu_1 > ...\mu_k > ...\mu_K$, SEPT requires the server always to choose jobs in grade 1 first, grade 2 second, ..., and finally grade K. Under such a rule, an arriving job will either receive service with a lower service value (a higher service rate) but a shorter waiting time, or wait longer for higher-quality service. Such a priority scheduling rule could easily be adopted in systems in which customers have no expectation of receiving service according to a FCFS rule, or when customers cannot observe the true service sequence. This is true in all online or telephone transactions, and *also* in many live transactions, such as auto repair shops, in which few customers will watch the whole repair processes.

3.4.2 The Dominance of Mixed Service

Before we derive the optimal service rates and service allocation rule, we need to validate the benefits of service grade differentiation, i.e., the dominance of mixed service over pure service. In this section we will use induction to show how the system performance may increase with the number of service grades K by showing how the average waiting time may decrease with K increasing while the aggregate average service time is held constant.

[3]In Wolff (1989) the optimal rule is the cμ-rule, under which the priority increases with the product of the delay cost rate c and the service rate μ. As in our model all grades have the same marginal delay cost, i.e. the same c (h in this paper), the cμ rule reduces to the SEPT rule under which the priority depends on the service rates μ only.

First, we derive the characterization of waiting time under the SEPT priority scheduling rule; the average waiting time in the kth service grade is (Harchol-Balter 2013, Chapter 33):

$$w_k = \frac{\lambda E[X^2]}{2(1-\sum_{i=1}^{k}\rho_i)(1-\sum_{i=1}^{k-1}\rho_i)},$$

where $E[X^2]$ denotes the second moment of the random service time X, and $\rho_i = \lambda p_i/\mu_i$ stands for the load of the ith service grade. For our model, as X is defined in (3.1), we have $E[X^2] = \mu_o^2 E[X_o^2]\sum_{k=1}^{K} p_k/\mu_k^2$. Hence the average waiting time of the kth service grade is:

$$w_k(\overrightarrow{\mu}, \overrightarrow{p}, r^*) = \frac{\lambda\mu_o^2 E[X_o^2]}{2(1-\sum_{i=1}^{k}\rho_i)(1-\sum_{i=1}^{k-1}\rho_i)}\left(\sum_{k=1}^{K}\frac{p_k}{\mu_k^2}\right). \tag{3.4}$$

To prove the dominance of mixed service , we start by showing that at least the simplest kind of mixed service (i.e. $K = 2$) is superior to pure service. The result is summarized in Proposition 2:

Proposition 5 *Under the optimal priority scheduling rule, for any given arrival rate λ, a mixed service policy having two service grades with service rates μ_1 and μ_2 and corresponding job allocation probabilities p_1 and p_2, has the same aggregate average service time but generates a shorter waiting time than pure service with service rate μ_o, if and only if*

$$\frac{\mu_2}{\mu_1} > 1-\rho, \tag{3.5}$$

where $\mu_1 > \mu_2$, $p_1 + p_2 = 1$, $\dfrac{p_1}{\mu_1} + \dfrac{p_2}{\mu_2} = \dfrac{1}{\mu_o}$ and $\rho = \dfrac{\lambda}{\mu_o}$.

Proof. Proof: Since the mixed service policy has the same expected service time as the pure service policy with service rate μ_o, we have

$$\frac{p_1}{\mu_1} + \frac{p_2}{\mu_2} = \frac{1}{\mu_o}, \tag{3.6}$$

and both mixed and pure service policies have the same system load as $\rho = \lambda/\mu_o$.

Next we compute the average waiting time for pure service, and then for mixed service. Finally we prove the dominance of mixed service by backward deductions from condition (3.5).

For pure service ($K = 1$) with service rate μ_o, there is no scaling of the original processing time. Thus the service time random variable of pure service $X_{(1)}$ satisfies $X_{(1)} = X_o$, $E[X_{(1)}^2] = E[X_o^2]$, and the load $\rho_{(1)} = \rho$. Based on (3.4) we have the average waiting time of pure service:

$$E[w]_{(1)} = \frac{\lambda E[X_{(1)}^2]}{2(1-\rho_{(1)})} = \frac{\lambda E[X_o^2]}{2(1-\rho)}. \tag{3.7}$$

For a two-grade mixed service, the corresponding service time random variable $X_{(2)}$ satisfies:

$$X_{(2)} = \begin{cases} \frac{\mu_o}{\mu_1} X_o \text{ w.p. } p_1; \\ \frac{\mu_o}{\mu_2} X_o \text{ w.p. } p_2. \end{cases}$$

Thus

$$E[X_{(2)}^2] = \mu_o^2 E[X_o^2](\frac{p_1}{\mu_1^2} + \frac{p_2}{\mu_2^2}).$$

As grade 1 is always served before grade 2:

$$E[w_1] = \frac{\lambda E[X_{(2)}^2]}{2(1-\rho_1)},$$
$$E[w_2] = \frac{\lambda E[X_{(2)}^2]}{2(1-\rho_1)(1-\rho_1-\rho_2)},$$

which, together with $\rho_1 = \frac{p_1\lambda}{\mu_1}, \rho_2 = \frac{p_2\lambda}{\mu_2}$ and $\rho = \rho_1 + \rho_2$, suggest that the total average waiting time of the two-grade mixed service satisfies:

$$E[w]_{(2)} = \frac{\lambda\mu_o^2 E[X_o^2]}{2(1-\rho_1)}(\frac{p_1}{\mu_1^2} + \frac{p_2}{\mu_2^2})\frac{1-p_1\rho}{1-\rho}. \tag{3.8}$$

We can compute $\frac{E[w]_{(2)}}{E[w]_{(1)}}$ based on (3.7) and (3.8):

$$\frac{E[w]_{(2)}}{E[w]_{(1)}} = (\frac{p_1}{\mu_1^2} + \frac{p_2}{\mu_2^2})\frac{\mu_o^2}{1-\rho_1}(1-p_1\rho). \tag{3.9}$$

Now we use condition (3.5) to show it implies (3.9) < 1:

$$\frac{\mu_2}{\mu_1} > 1 - \rho \Leftrightarrow \frac{\rho}{\mu_2} + \frac{1}{\mu_1} - \frac{1}{\mu_2} > 0. \tag{3.10}$$

Multiplying (3.10) with $p_1 p_2 > 0$ and $\frac{1}{\mu_1} - \frac{1}{\mu_2} < 0$, it becomes:

$$\begin{aligned}\frac{\mu_2}{\mu_1} > 1 - \rho &\Leftrightarrow p_1 p_2 \frac{\rho}{\mu_2}(\frac{1}{\mu_1} - \frac{1}{\mu_2}) + p_1 p_2 (\frac{1}{\mu_1} - \frac{1}{\mu_2})^2 < 0 \\ &\Leftrightarrow p_1 \rho (\frac{p_2}{\mu_2}\frac{1}{\mu_1} - \frac{p_2}{\mu_2^2}) + p_1 p_2 (\frac{1}{\mu_1} - \frac{1}{\mu_2})^2 < 0.\end{aligned}$$

Based on (3.6), we substitute $\frac{p_2}{\mu_2}$ with $\frac{1}{\mu_o} - \frac{p_1}{\mu_1}$ into the above inequality:

$$\begin{aligned}&\Leftrightarrow p_1 \rho ((\frac{1}{\mu_o} - \frac{p_1}{\mu_1})\frac{1}{\mu_1} - \frac{p_2}{\mu_2^2}) + p_1 p_2 (\frac{1}{\mu_1} - \frac{1}{\mu_2})^2 < 0 \\ &\Leftrightarrow p_1 \rho \frac{1}{\mu_o \mu_1} - p_1 \rho (\frac{p_1}{\mu_1^2} + \frac{p_2}{\mu_2^2}) + p_1 p_2 (\frac{1}{\mu_1} - \frac{1}{\mu_2})^2 < 0 \end{aligned} \tag{3.11}$$

As we have:

$$p_1 \rho \frac{1}{\mu_o \mu_1} = p_1 \frac{\lambda}{\mu_o}\frac{1}{\mu_o \mu_1} = \rho_1 \frac{1}{\mu_o^2}$$

and

$$p_1 p_2 (\frac{1}{\mu_1} - \frac{1}{\mu_2})^2 = (\frac{p_1}{\mu_1^2} + \frac{p_2}{\mu_2^2}) - \frac{1}{\mu_o^2},$$

we can replace the above two equalities into (3.11):

$$\begin{aligned}(3.11) &\Leftrightarrow \rho_1 \frac{1}{\mu_o^2} - p_1 \rho (\frac{p_1}{\mu_1^2} + \frac{p_2}{\mu_2^2}) + (\frac{p_1}{\mu_1^2} + \frac{p_2}{\mu_2^2}) - \frac{1}{\mu_o^2} < 0 \\ &\Leftrightarrow (\frac{p_1}{\mu_1^2} + \frac{p_2}{\mu_2^2})(1 - p_1 \rho)\frac{\mu_o^2}{1 - \rho_1} < 1.\end{aligned}$$

Comparing the above inequality with (3.9), finally we show that:

$$(3.5) \Leftrightarrow \frac{E[w]_{(2)}}{E[w]_{(1)}} < 1$$

■

Note that condition (3.5) can always be satisfied for any $\rho \in (0, 1)$. Thus for any λ and μ, there is a continuum of two-service grade policies that are superior to pure service.

Proposition 4 concludes that under the FCFS rule mixed service is inferior to pure service because of higher variance. But in Proposition 5, mixed service *under the SEPT rule* generates lower waiting times than pure service. The underlying intuition is as follows: The total waiting time (referred to as W) can be divided into two parts: waiting time for unfinished work in queue (referred to as WQ) and waiting time for unfinished work at the server (referred to as WS), where the latter part, WS, is often called the excess service time.

Since mixed service has a higher variance, its mean excess service time ($E[S_e] = \frac{E[S^2]}{2E[S]}$) is also higher and generates a higher WS. But the SEPT rule reduces WQ by serving short jobs before long jobs: The WQ of short jobs is reduced by the service times of the long jobs in queue "ahead" of them, while that of the long jobs is increased by the service times of short jobs in queue "behind" them. Since long jobs are longer, the total WQ is reduced. (This is the same principle, *pairwise interchange*, which is commonly used in proofs of the optimality of SEPT). Under some conditions, the reduced WQ outweighs the increased WS, and the total waiting time can be reduced.

It is not surprising that condition (3.5) bounds the ratio of the two service rates, because a significant difference between these two rates will increase the variance and thus the WS, to a level that even the SEPT rule will be unable to compensate for. Condition (3.5) also implies that the bound $1 - \rho$ becomes less restrictive in a more congested system, as the SEPT scheduling rule is more effective with more jobs waiting.

Now that we have proved that mixed service with two grades may perform better than pure service, we will investigate the effects of the number of service grades K on the system performance. It is not difficult to guess that the larger the number of service grades, the better the performance. But proving this conjecture is more complicated than the proof of Proposition 5, in which we separate one service grade into two but keep the average rate unchanged: If we consider comparing $K = 2$ and $K = 3$, for example, we can also choose one service grade, say, the first service grade with service rate μ_1, and change it into two new grades with service rates μ_{1a} and μ_{1b}, while keeping the aggregate average service time and assigned jobs' arrival rate unchanged. Similar to Proposition 5, we know that the two service grades with rates μ_{1a} and μ_{1b} have the same aggregate average service time but generate a shorter average waiting time than the original single service grade with

rate μ_1. But as the service time variance increases when K increases from 2 to 3, the average waiting time of jobs in the *unchanged* grade also increases. Thus we must derive new conditions that ensure that adding a grade improves performances:

Proposition 6 *Under the optimal priority scheduling rule, for any given arrival rate λ, mixed service having $K+1$ service grades with service rate vector $\overrightarrow{\mu} = (\mu_{1a}, \mu_{1b}, \mu_2..., \mu_k, ..., \mu_K)$ and grade allocation rule $\overrightarrow{p} = (p_{1a}, p_{1b}, p_2..., p_k, ..., p_K)$, has the same aggregate average service time but generates a shorter waiting time than mixed service having K service grades with service rate vector $\overrightarrow{\mu} = (\mu_1, \mu_2..., \mu_k, ..., \mu_K)$ and grade allocation rule $\overrightarrow{p} = (p_1, p_2..., p_k, ..., p_K)$, if and only if*

$$\frac{1}{\mu_{1b}} - \frac{1}{\mu_{1a}} < \frac{Ap_1\lambda}{B(1-\rho_{1a})(1-\rho_1)},$$

which suggests a simple sufficient condition as

$$\frac{\mu_{1a}}{\mu_{1b}} < 1 + \frac{\rho_1}{(1-\rho_1)\,B},$$

where $A = \left(\sum_{k=1}^{K} \frac{p_k}{\mu_k^2} + \frac{p_{1a}p_{1b}}{p_1}\left(\frac{1}{\mu_{1b}} - \frac{1}{\mu_{1a}}\right)^2\right)$, $B = \sum_{k=1}^{K} \frac{p_k}{(1-\sum_{i=1}^{k}\rho_i)(1-\sum_{i=1}^{k-1}\rho_i)}$, $p_{1a} + p_{1b} = p_1$, $\frac{p_{1a}}{\mu_{1a}} + \frac{p_{1b}}{\mu_{1b}} = \frac{p_1}{\mu_1}$ *and* $\rho_i = \frac{\lambda p_i}{\mu_i}$ *for all* $i = 1, 2, ...K$.

Similar to Proposition 5, Proposition 6 also provides a bound on the difference between the two new service rates, which can always be attained for any mixed service with K service grades. Propositions 5 and 6 together lead to the following corollary.

Corollary 1 *Having the same aggregate average service time, mixed service can generate a shorter average waiting time than pure service. In addition, increasing the number of service grades can generate a policy with shorter average waiting time.*

Corollary 1 concludes that offering additional service grades lowers expected waiting time. Note that the waiting time is reduced not because all customers are assigned to the fastest service grade: Customers are differentiated into either faster or slower service grades so that the aggregate average service time spent by the service provider remains unchanged. Since the expected waiting time is reduced, the system utility is improved. This superiority of mixed service holds in not only

quality-based settings, but also in other settings where customers derive a fixed value from service ("constant-quality" systems). Thus, mixed service can benefit the quality-based system and the constant-quality system by reducing average waiting time while not affecting the aggregate average service time, which is formally stated in the following theorem:

Theorem 1 *Mixed service dominates pure service in quality-based (and constant-quality) domains, and the benefit increases with the number of service grades.*

In the previous proofs, we show the benefits of service rate differentiation when holding the aggregate average service time taken over all jobs constant. Furthermore, it is not difficult to verify that, with the configuration of service rates and allocation probabilities in Proposition 5 and Proposition 6, the service time variance increases with the number of service grades. Thus, the superiority of mixed service shows the value of introducing service time variability into the system. Moreover, this value also extends to constant-quality domains such as systems where the objective is to minimize waiting time subject to a fixed service capacity, or service load. We will interpret this potential benefit of introducing variability in detail in Section 3.6.1.

3.4.3 Optimal Strategy

The previous analysis proves the dominance of mixed service over pure service by showing that mixed service can generate a shorter average waiting time while retaining the same average service value. However, the optimal structure of a mixed service policy remains unclear. In this section we derive the characteristics of *optimal* service rates and service allocation polices, by solving the system performance maximization problem with a given $K \geq 1$ number of service grades.

Based on the optimality of SEPT the original maximization model (3.3) can be stated as:

$$\begin{aligned} \underset{\vec{\mu},\vec{p}}{max}\ V = \lambda \left(u \sum_{k=1}^{K} p_k/\mu_k - h \sum_{k=1}^{K} \frac{p_k \lambda \mu_o^2 E[X_o^2]}{2(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)} \left(\sum_{k=1}^{K} \frac{p_k}{\mu_k^2} \right) \right), \\ s.t.\ \sum_{k=1}^{K} p_k = 1, \end{aligned} \tag{3.12}$$

where $\rho_i = \lambda p_i/\mu_i$ denotes the load of the ith service grade, and as assumed we have $\mu_1 > ...\mu_k > ...\mu_K$, and $0 < p_k < 1$ for all $k = 1, ...K$. If $p_k = 0$, the kth service grade is abandoned, and the

strategy degenerates to $K-1$ service grades; if $p_k = 1$, the strategy becomes pure service.

We first substitute the service grades load vector $\overrightarrow{\rho}$ for the service rate vector $\overrightarrow{\mu}$ in (3.12) based on the following definitions:

$$\mu_k = \lambda p_k / \rho_k \quad \text{for } k = 1, ...K. \tag{3.13}$$

Hence the corresponding optimization function can be formulated with new decision variables:

$$\begin{aligned} \max_{\overrightarrow{\rho},\overrightarrow{p}} V = u\sum_{k=1}^{K}\rho_k - C\left(\sum_{k=1}^{K}\frac{p_k}{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}\right)\left(\sum_{k=1}^{K}\frac{\rho_k^2}{p_k}\right), \\ s.t.\ \sum_{k=1}^{K} p_k = 1, \end{aligned} \tag{3.14}$$

where we denote $\dfrac{h\mu_o^2 E[X_o^2]}{2} = C > 0$, a constant which only depends on jobs' exogenous characteristics. Noting that the positive part in (3.14), $u\sum_{k=1}^{K}\rho_k$, representing the system service value, is linear with the total system load $\rho = \sum_{k=1}^{K}\rho_k$, we introduce the system load ρ as a decision variable in place of the Kth service grade's load, ρ_K.

Using ρ, the maximization problem can be decomposed into two parts. The inner part is equivalent to the minimization problem of the average waiting time in the system for a fixed ρ:

$$w^* := \min_{\overrightarrow{p},\rho_1,...\rho_{K-1}|\rho} \frac{C}{\lambda h}\left(\sum_{k=1}^{K}\frac{p_k}{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}\right)\left(\sum_{k=1}^{K}\frac{\rho_k^2}{p_k}\right) \tag{3.15}$$

$$s.t.\ \sum_{k=1}^{K} p_k = 1; \tag{3.16}$$

$$\sum_{k=1}^{K}\rho_k = \rho. \tag{3.17}$$

The outer part aims to maximize the system performance through ρ only:

$$\max_{\rho} V = u\rho - h\lambda w^*. \tag{3.18}$$

We will first solve the inner minimization problem (3.15) with constraints (3.16) and (3.17). Observing that (3.16) only contains decision variables $\overrightarrow{p}$, the minimization problem (3.15) can be

further decomposed into two steps: For a fixed $\rho_1, ...\rho_{K-1}$ we determine the optimal expression for $\overrightarrow{p}$. Then using this expression we determine the optimal values of $\rho_1, ...\rho_{K-1}$.

The minimization problem for $\overrightarrow{p}$ is equivalent to:

$$\begin{aligned} \underset{\overrightarrow{p}|\{\rho_1,...\rho_{K-1},\rho\}}{min} \quad & \left(\sum_{k=1}^{K} \frac{p_k}{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}\right)\left(\sum_{k=1}^{K}\frac{\rho_k^2}{p_k}\right), \\ s.t. \quad & \sum_{k=1}^{K} p_k = 1. \end{aligned} \tag{3.19}$$

Problem (3.19) can be solved directly by the Cauchy-Schwarz inequality:

Lemma 2 *With given $\{\rho_1, ...\rho_{K-1}, \rho\}$, the optimal solution $\overrightarrow{p}$ of problem (3.19) is*

$$p_k = \frac{\rho_k\sqrt{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}}{\sum_{k=1}^{K}\rho_k\sqrt{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}} \quad \textit{for } k = 1, ...K. \tag{3.20}$$

Substituting (3.20) into (3.15), the inner minimization problem becomes

$$\underset{\rho_1,...\rho_{K-1}|\rho}{min}\left(\sum_{k=1}^{K}\frac{\rho_k}{\sqrt{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}}\right)^2. \tag{3.21}$$

Remark 1 *Constraint (3.16) does <u>not</u> affect the optimal value of problem (3.19). This means, even with $\sum_{k=1}^{K} p_k < 1$ (i.e. rejecting some jobs), while the optimal solution may change, the objective of problem (3.19) will not. Through the following analysis, we will find that allowing jobs to be rejected likewise will not influence the optimal values of $\rho_1, ...\rho_{K-1}$ and ρ as long as $\sum_{k=1}^{K} p_k > 0$.*

It is trivial to verify that the minimization problem (3.21) is equivalent to

$$\begin{aligned} \min_{\rho_1,\rho_2,...\rho_K} \quad & \sum_{k=1}^{K}\frac{\rho_k}{\sqrt{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}}, \\ s.t. \quad & \sum_{i=1}^{K}\rho_i = \rho. \end{aligned} \tag{3.22}$$

Problem (3.22) is actually a service load allocation problem which seeks to minimize the average waiting time with a fixed system load ρ. While there is no obvious structure (convexities or

concavities) to the objective function, the special case of problem (3.22) with only two service grades suggests that the optimal solutions $\{\rho_k\}$ satisfy what we call the *chain rule*. Such a rule can be extended to the general case with multiple service grades, as an optimization of the multi-grade case can always be achieved by partial adjustments of any two adjacent service grades' loads. The chain rule is defined in the following lemma.

Lemma 3 *The optimal solutions of (3.22) satisfy the following chain rule:*

$$\left(1-\sum_{i=1}^{k}\rho_i\right)^2=\left(1-\sum_{i=1}^{k+1}\rho_i\right)\left(1-\sum_{i=1}^{k-1}\rho_i\right)\quad \textit{for } k=1,...K-1.$$

Proof. Proof: Initially we focus on a special case of problem (3.22) with only two grades:

$$\min_{\rho_k,\rho_{k+1}}\ \frac{\rho_k}{\sqrt{\left(1-\sum_{i=1}^{k-1}\rho_i\right)\left(1-\sum_{i=1}^{k}\rho_i\right)}}+\frac{\rho_{k+1}}{\sqrt{\left(1-\sum_{i=1}^{k}\rho_i\right)\left(1-\sum_{i=1}^{k+1}\rho_i\right)}}.$$

As in the above model all the variables are fixed except for ρ_k and ρ_{k+1}, both $\sum_{i=1}^{k-1}\rho_i$ and $\sum_{i=1}^{k+1}\rho_i$ are constants. We simplify the objective as follows:

$$\begin{aligned}&\frac{\rho_k}{\sqrt{\left(1-\sum_{i=1}^{k-1}\rho_i\right)\left(1-\sum_{i=1}^{k}\rho_i\right)}}+\frac{\rho_{k+1}}{\sqrt{\left(1-\sum_{i=1}^{k}\rho_i\right)\left(1-\sum_{i=1}^{k+1}\rho_i\right)}}\\&=\left(\sqrt{1-\sum_{i=1}^{k-1}\rho_i}-\sqrt{1-\sum_{i=1}^{k+1}\rho_i}\right)\left(\frac{1}{\sqrt{1-\sum_{i=1}^{k}\rho_i}}+\frac{\%sqrt1-\sum_{i=1}^{k}\rho_i}{\sqrt{1-\sum_{i=1}^{k-1}\rho_i}\sqrt{1-\sum_{i=1}^{k+1}\rho_i}}\right).\end{aligned}$$

When both $\sum_{i=1}^{k-1}\rho_i$ and $\sum_{i=1}^{k+1}\rho_i$ are constants, based on the Cauchy-Schwarz inequality, the objective is minimized if and only if

$$\begin{aligned}&\frac{1}{\sqrt{1-\sum_{i=1}^{k}\rho_i}}=\frac{\sqrt{1-\sum_{i=1}^{k}\rho_i}}{\sqrt{1-\sum_{i=1}^{k-1}\rho_i}\sqrt{1-\sum_{i=1}^{k+1}\rho_i}}\\&\Leftrightarrow\left(1-\sum_{i=1}^{k}\rho_i\right)^2=\left(1-\sum_{i=1}^{k+1}\rho_i\right)\left(1-\sum_{i=1}^{k-1}\rho_i\right).\end{aligned}\tag{3.23}$$

Thus for any two adjacent service grades, say k and $k+1$, given the constraint that all the other service grades' loads are fixed, the aggregate average waiting time is minimized if and only if the

two decision variables ρ_k and ρ_{k+1} satisfy the chain rule.

Next we deduce that such a rule solves problem (3.22). The key idea is that if any two adjacent service grades' loads do not satisfy the chain rule (3.23), a reallocation of loads between these two service grades following the chain rule can always improve the objective without affecting the other service grades. Similarly we can adjust all the grades' loads $\rho_1, \rho_2, ...\rho_K$ until

$$\left(1 - \sum_{i=1}^{k} \rho_i\right)^2 = \left(1 - \sum_{i=1}^{k+1} \rho_i\right)\left(1 - \sum_{i=1}^{k-1} \rho_i\right) \quad \text{for } k = 1, ...K-1.$$

■

The chain rule tells us that the summations of different grades' loads are correlated in a particular "chain-like" fashion. With the chain rule all the service grades' loads $\rho_1, \rho_2, ...\rho_K$ can be derived sequentially and expressed in terms of ρ_1. Such a derivation is in accordance with the principle that under a non-preemptive priority scheduling rule a job's average waiting time primarily depends on those jobs with equal or higher priorities. In addition, the chain rule implies the following relationship between the waiting times of all the service grades.

Corollary 2 *All the service grades have the same "weighted waiting time" under the optimal allocation policy:*

$$p_k w_k = p_{k+1} w_{k+1} \quad \textit{for } k = 1, ...K-1,$$

where $w_k = E[W_k]$ denotes the average waiting time of the kth service grade.

Corollary 2 indicates that the optimal strategies of our model generate a balance of average waiting time for arriving jobs. In other words, the contribution of each term in the sum $E[W] = \sum_{k=1}^{K} p_k w_k$ is equivalent among all the service grades, although the expected service values are *not* equal (which is implied by the fact that the values of ρ_k under the chain rule are not the same).

Under the chain rule, the load of each service grade can be expressed in terms of ρ:

Corollary 3 *The optimal load of the kth grade ρ_k satisfies:*

$$\rho_k = (1-\rho)^{\frac{k-1}{K}}\left(1 - (1-\rho)^{\frac{1}{K}}\right).$$

The above equation defines the geometric form of the optimal load allocation vector $\overrightarrow{\rho}$. Substituting the optimal ρ_k into the optimal probability allocation given in (3.20) and the optimal service rate given in (3.13) also generates a geometric sequence of $\overrightarrow{p}$ and $\overrightarrow{\mu}$:

Proposition 7 *The optimal strategy of the average waiting time minimization problem (3.15) with K service grades requires that the service rate vector $\overrightarrow{\mu} = (\mu_1, \mu_2, ..., \mu_K)$, service allocation vector $\overrightarrow{p} = (p_1, p_2, .., p_K)$ and correspondingly service load vector $\overrightarrow{\rho} = (\rho_1, \rho_2, .., \rho_K)$ form geometric sequences respectively, i.e.,*

$$\frac{p_{k+1}}{p_k} = \left(\frac{\mu_{k+1}}{\mu_k}\right)^2 = \left(\frac{\rho_{k+1}}{\rho_k}\right)^2 = (1-\rho)^{\frac{2}{K}} \quad for \;\; k = 1, ...K-1. \tag{3.24}$$

Remark 2 *The geometric structure in Proposition 7 minimizes the average waiting time (3.15) under a fixed system load ρ. Based on this, it can be inferred that the geometric structure can also apply in constant-quality settings such as minimizing the average waiting time subject to a fixed or maximal service capacity, service load, or service value. In addition, Proposition 7 shows the benefit of introducing service time variability even without adding more service grades: The service policy that minimizes the average waiting time requires a specific geometric structure of service rates, and the geometric ratio, as a function of system load ρ, is less than 1. Since the geometric ratio is other than 1, the service time variance of the optimal policy for a fixed number of service grades is not minimized. Thus, there exist an infinite number of systems with the same number of service grades (being served according to SEPT) having* lower *service time variability but generating* higher *waiting times than the optimal policy. By changing these service policies to the optimal structure in (3.24) while keeping the system load unchanged, service time variability is increased and the average waiting is reduced, but no more service grades are introduced.*

Next we will derive the optimal system load ρ^* that maximizes the system performance V as shown in (3.18). Based on the optimal grade load ρ_k given in Corollary 3, the optimal average waiting time w^* defined in (3.15) can be simplified as

$$w^* = \frac{C}{\lambda h} K^2 \left((1-\rho)^{-\frac{1}{K}} + (1-\rho)^{\frac{1}{K}} - 2 \right). \tag{3.25}$$

Now (3.18) becomes an unconstrained maximization problem with only one decision variable, ρ:

$$\max_{\rho} V = u\rho - CK^2\left((1-\rho)^{-\frac{1}{K}} + (1-\rho)^{\frac{1}{K}} - 2\right), \tag{3.26}$$

where $0 < \rho < 1$. Problem (3.26) can be solved with its first order condition.

Proposition 8 *The optimal system load ρ^* for the problem (3.26) satisfies the following equation:*

$$u(1-\rho) + CK\left((1-\rho)^{\frac{1}{K}} - (1-\rho)^{-\frac{1}{K}}\right) = 0; \tag{3.27}$$

such a ρ^* always exists, and is unique.

From (3.27) we can see that the optimal system load ρ^* only depends on three parameters: K, u, and $C = \frac{h\mu_o^2 E[X_o^2]}{2}$. Thus we can conclude that ρ^*, and consequently the optimal system load vector $\overrightarrow{\rho}^*$, are independent of jobs' arrival rate λ. And, as the optimal grade assignment vector $\overrightarrow{p}^* = \{p_k^* : k = 1, 2...K\}$ given by (3.20) only depends on $\overrightarrow{\rho}^*$, we can conclude that $\{p_k^*\}$ is also independent of the arrival rate λ. In contrast, based on the definition of $\mu_k = \lambda p_k/\rho_k$, μ_k^* is linear in λ for all $k = 1, 2...K$. So the optimal system responds to a change in arrival rate by changing the service rates proportionally, but leaving the allocation scheme unchanged.

With the optimal solution ρ^*, it is trivial to derive the optimal system performance V^*:

$$V^* = u\rho^* - CK^2\left((1-\rho^*)^{-\frac{1}{K}} + (1-\rho^*)^{\frac{1}{K}} - 2\right). \tag{3.28}$$

The fact that the optimal system performance V^* contains only ρ^* implies that V^* in turn only depends on u, K and $C = \frac{h\mu_o^2 E[X_o^2]}{2}$:

Proposition 9 *The optimal system performance depends only on the optimal system load ρ^* and is insensitive to the arrival rate λ.*

The independence of V^* to arrival rate λ is striking. This results from the optimal strategy in which the server adjusts the service rates to recapture the optimal load ρ^* if faced by a different λ. In addition, using the envelope theorem it is not difficult to verify that V^* increases with K.

3.4.4 Asymptotic System Performance

Having established the optimal differentiation strategies, the corresponding optimal system performance, and the fact that the optimal system performance increases with the number of service grades K, we now evaluate the optimal system performance as K approaches infinity.

We first identify the asymptotically optimal service differentiation strategy, which consists of the asymptotically optimal system load $\widetilde{\rho}$, and then compute the asymptotically minimal average waiting time $\widetilde{w}$ and optimal system performance $\widetilde{V}$. Finally, we study the asymptotic performance improvement of the static differentiation policy compared to pure service.

By observing the formulae of w^* given in (3.25) and ρ^* given in (3.27), we see that both parts containing K have the same pattern: $a^{\frac{1}{K}} - a^{-\frac{1}{K}}$, where $a > 0$. Thus we define a function $f : \mathbb{R} \to \mathbb{R}$ as $f(x) = a^x - a^{-x}$ with $a > 0$. Based on Taylor expansion, $f(x)$ can be expanded at $x = 0$ as follows:

$$a^x - a^{-x} = 2(\ln a)\, x + \frac{2(\ln a)^3}{3!}x^2 ... + \frac{(\ln a)^n}{n!}x^{n-1}\left(1 + (-1)^{n-1}\right) + ...$$

The limit of $\dfrac{f(x)}{x}$ is:

$$\lim_{x \to 0} \quad \frac{1}{x}\left(a^x - a^{-x}\right) = 2 \ln a.$$

Thus the asymptotic limit of the latter portion of the expression for ρ^* defined in (3.27) is:

$$\lim_{K \to \infty} K\left((1-\rho)^{\frac{1}{K}} - (1-\rho)^{-\frac{1}{K}}\right) = 2\ln(1-\rho).$$

Substituting the above term into (3.27) suggests that the asymptotically optimal system load $\widetilde{\rho}$ satisfies

$$u - \frac{h\mu_o^2 E[X_o^2]}{1-\widetilde{\rho}} \ln \frac{1}{1-\widetilde{\rho}} = 0. \tag{3.29}$$

Similarly we can derive the limit of the minimal average waiting time w^* given in (3.25):

$$\lim_{K \to \infty} K^2\left((1-\rho)^{-\frac{1}{K}} + (1-\rho)^{\frac{1}{K}} - 2\right) = \lim_{K \to \infty} K^2\left((1-\rho)^{\frac{1}{2K}} - (1-\rho)^{-\frac{1}{2K}}\right)^2$$
$$= \lim_{K \to \infty} \frac{1}{4}\left(\frac{(1-\rho)^{\frac{1}{2K}} - (1-\rho)^{-\frac{1}{2K}}}{1/2K}\right)^2 = \frac{1}{4}\left(2\ln(1-\rho)\right)^2 = \left(\ln(1-\rho)\right)^2.$$

Substituting the above limit value into (3.25) and (3.28), respectively, we obtain the asymptotic

average waiting time and system performance, respectively:

$$\widetilde{w} = \lim_{K\to\infty} \quad w^*(K) = \frac{\mu_o^2 E[X_o^2]}{2\lambda}\left(\ln(1-\rho)\right)^2\Big|_{\rho=\widetilde{\rho}}; \tag{3.30}$$

$$\widetilde{V} = \lim_{K\to\infty} \quad V^*(K) = u\rho - \frac{h\mu_o^2 E[X_o^2]}{2}\left(\ln(1-\rho)\right)^2\Big|_{\rho=\widetilde{\rho}}, \tag{3.31}$$

Note that the asymptotic value $\widetilde{w}$ given in (3.30) suggests that when the system load $\rho \to 1$, the growth rate of the expected waiting time under the differentiation policy is of order $\left(\ln(1-\rho)\right)^2$. This is similar to a result in Lin *et al.* (2011), where the authors find that under preemptive SRPT (as opposed to our SEPT scheduling), the growth rate of the average waiting time could be as slow as the order of $(-\ln(1-\rho))$ as $\rho \to 1$ when the service time distribution is unbounded.

Finally we define the *dominance index* R_V as a measure of the ratio of the optimal system performance of K-grade mixed service to that of pure service. Correspondingly, the asymptotic dominance index, denoted by $\widetilde{R_V}$, is the ratio of the asymptotically optimal system performance under service differentiation to that of pure service, i.e. $\widetilde{R_V} = \widetilde{V}/V(1)$. In Appendix B we show that $\widetilde{R_V}$ depends on the external parameters u, h and $\mu_o^2 E[X_o^2]$ only through the term $\tau := \dfrac{u}{h\mu_o^2 E[X_o^2]}$. Numerical analysis suggests that the first order condition $\dfrac{d\widetilde{R_V}}{d\tau} = 0$ has a unique positive root at $\tau = 52.626$, at which $\widetilde{R_V} = 1.0528$, indicating an improvement of 5% compared to pure service. We numerically verify this result in Section 3.5.

3.5 Numerical Analysis

Even with our analytical solution of the problem, two questions still remain: how many service grades are "enough" and how the benefit of differentiation changes with system parameters. In this section, we answer the first question by studying the behavior of the dominance index R_V with respect to the number of service grades K. We also show how the optimal expected service value and average waiting cost change with K. To address the second question, we illustrate the values of $\widetilde{R_V}$ under different settings of parameters u, h and the coefficient of variation CV. As a part of this study we also illustrate the different service values and waiting costs of individual service grades.

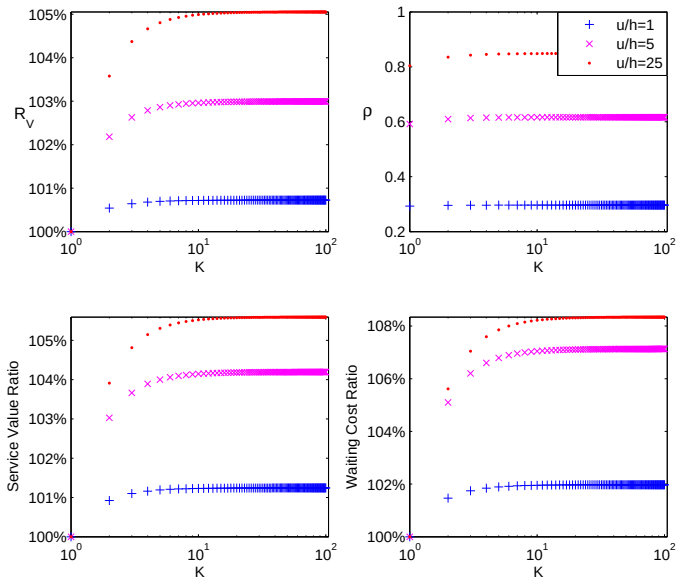

Figure 3.1: How R_V, ρ, service value and average waiting cost change with number of service grades K

Unless otherwise noted, our parameters are as follows[4]: $\lambda = 1$, $h = 1$, $u = 5$ and $CV = 1$. For simplicity, we study the sensitivity of system performance to u/h instead of u and h individually.

First, we numerically study the effects of the number of service grades K on performance metrics, including the dominance index R_V, optimal system load ρ, service value ratio (the ratio of service value of mixed service to that of pure service), and waiting cost ratio (the ratio of waiting cost of mixed service to that of pure service). All values are computed based on the optimal service strategy for corresponding K. The three different types of markers in Figure 3.1 represent three cases, under different ratios of marginal service value u to marginal delay cost h. The upper-left plot in Figure 3.1 validates our theorem that the system performance increases with K, and shows that under all three settings the optimal system performance has almost converged when $K = 10$. Likewise the two graphs in the bottom of Figure 3.1 show that both service value and waiting cost increase as K increases. This increment of service value arises from an increase in system load ρ, as seen in the upper-right graph.

Figure 3.2 illustrates how the performance of static differentiation is affected by jobs' character-

[4]Different sets of parameters generate the same patterns as this set.

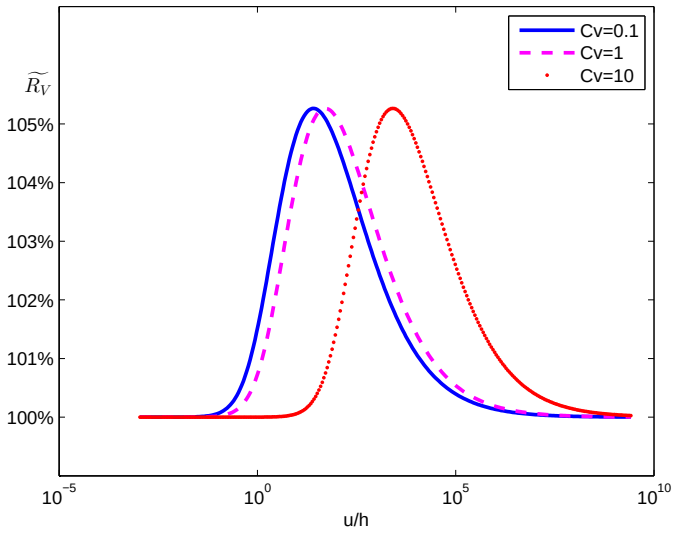

Figure 3.2: Asymptotic dominance index $\widetilde{R_V}$ changes with u/h and CV

istics, which include marginal service value u, marginal waiting cost h and variance of jobs' original processing time, represented by the coefficient of variation $CV = \sqrt{\mu_o^2 E[X_o^2] - 1}$. Figure 3.2 shows that for a fixed CV, with u/h increasing, $\widetilde{R_V}$ increases before reaching a peak, and then gradually decreases. In addition, when CV increases, the whole pattern shifts to the right, but the peak values remain the same, around 105%. These observations coincide with the findings presented in Section 3.4.4 that the asymptotic ratio $\widetilde{R_V}$ only depends on $\tau = \dfrac{u}{h\,(CV^2+1)}$ and the function has only one stationary point $\tau = 52.626$, at which $\widetilde{R_V} = 1.0528$.

The sensitivity of $\widetilde{R_V}$ with respect to u/h can be understood as follows. On the one hand, when $u/h \to 0$, the asymptotic optimal system load $\widetilde{\rho}$ approaches 0, under which the SEPT works least effectively in reducing the average waiting time compared to the uniform service. On the other hand, when $u/h \to \infty$ the delay cost becomes so insignificant that even though service differentiation can greatly reduce waiting, it contributes little in improving the system performance. The same reasoning can be applied to understand the sensitivity of $\widetilde{R_V}$ with respect to CV.

Finally, we show the distribution of grades' performances and probability allocations in Figure 3.3, in which service grades $1, 2, ...K = 10$ are represented by the rectangles from left to right. The width of the kth rectangle represents the allocation probability p_k. The height of the white rectangle above the x-axis illustrates the net performance of each grade, and that of the opaque

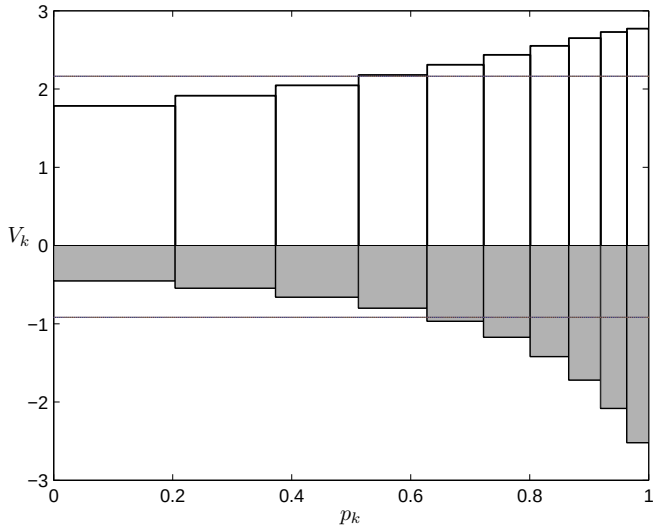

Figure 3.3: Grade performance V_k vs probability allocation p_k for the kth service grade with $K = 10$

rectangle below the x-axis depicts the average waiting cost of each grade, so the absolute height (white plus opaque) gives the mean service value. Figure 3.3 illustrates that the waiting time, the service value and the grade performance are all increasing as the grade's priority decreases; jobs assigned lower priorities wait for a longer time, receive loner service and gain more net benefit (An analytical proof is given in Appendix B). The dashed lines show the average performance and waiting cost of the system respectively, from which we can see that half of the jobs obtain more net value than average.

3.6 Conclusions, Discussions and Future Work

3.6.1 Interpretations of Variance Benefits

Generally, people believe that when facing homogeneous jobs, uniform service is superior to mixed service, as the latter policy generates greater service time variability. Our work shows that adding variability to a system by introducing service rate differentiation, and serving jobs according to the SEPT rule can actually reduce the overall waiting time. Unlike traditional dynamic control of service rates, such differentiation is effective even if it is static and independent of system state.

The benefit of introducing service time variability could be attributed to the value of service

rate information created by differentiation. Specifically, while the differentiation policy increases service time variance and hence lengthens waiting, it also creates service rate information and thus enables the implementation of SEPT, which conversely reduces waiting. The former impact on the waiting time is the well-known variance effect, and the latter could be referred to as "information value." Our research shows that when properly implemented, the information value outweighs the variance effect, and thus a benefit arises despite increased variability. The benefit can be thought of as the surplus of the information value over the variance effect (see Example I, Appendix B.2).

It is worth noting that the service rate information used by our policy is created without effort because service rates are endogenously chosen and assigned by the system. In this sense, the information value comes for free. This is quite different from the "service rate information" referred to in most of the previous literature, where the service rate/value information is intrinsic and requires effort or proper incentives to be obtained (e.g. Van Mieghem 2000).

The value of the created service rate information is affected by the values of parameters chosen for differentiation in a complicated way. Specifically, while keeping the aggregate average service time fixed, the information value is affected by (at least) two factors: service time variance and the number of service grades. One may reasonably expect that information value should increase with the number of service grades, as more information is created. But we find that this may not be the case: Information value may decrease or increase with the number of service grades, and this happens even when service time variance is fixed (see Example II, Appendix B.2). The relationship between information value and service time variance is likewise complicated: Intuitively, information value should increase with service time variance, but we find that it may decrease as variance increases for systems with different number of service grades (see Example III, Appendix B).

Both variance effect and information value are well documented in literature, but, to the best of our knowledge, our work is the first to discuss their overall effect and arrive at the conclusion that a surplus may arise from introducing variability. In addition, we derive necessary and sufficient conditions under which the surplus occurs and characterize the optimal policies to maximize the surplus for a fixed system load — a geometric structure for service rates and assignment probabilities, with correlated ratios that depend on system load.

3.6.2 Future Work

We show the benefit of introducing service time variance through a static service differentiation policy, and illustrate it within quality-based service domains. This work can be extended in multiple ways. First, the benefit of static differentiation (as well as the optimal geometric structure) can be extended to more general settings of quality-based service as well as other non-quality-based settings. For quality-based service, in this paper, we adopt linear functions for service value and waiting cost. Future research can consider nonlinear function forms. Appendix B.3 shows that the dominance of mixed service can be easily extended to convex service value and linear waiting cost, but concave service value, or nonlinear waiting cost require more study.

The differentiation policy and the optimal geometric structure also extend to other service domains which minimize the average waiting time subject to a fixed service capacity, or service load. Since the optimal geometric structure only depends on a few parameters and thus can potentially be implemented, this related research has great application potential.

The service differentiation policy proposed in this paper can also be extended in several directions. Appendix B.3 illustrates two directions: preemptive scheduling and dynamic allocation. In the former case we can analytically prove that under the shortest expected remaining processing time first (SERPT) rule, for exponentially distributed jobs, any service rate differentiation with two service grades can *always* generate a shorter average waiting time than pure service with the same aggregate average service time (see Proposition 10 in Appendix B.3). Further research can consider characterizing conditions for other specific or a general service time distributions.

We also consider a dynamic service differentiation policy, under which the service provider chooses a service rate for each job when it starts service from a set of candidate rates, based on the system state. In this case the optimal policy is a threshold policy that assigns service rates based on queue length: The longer the queue, the faster the service rate allocated. (This is similar to the TP control proposed by Armony and Gurvich, 2010.) Based on this structure we can numerically compute the optimal dynamic policy of a two-grade system and demonstrate the benefits of the dynamic allocation rule by comparing it with a static one. A numerical example suggests that a dynamic allocation rule introduces more service time variability and generates up to 18% more system value (see Figure B.31 in Appendix B.3). Thus future research on characterizing

the optimal dynamic allocation rule will further benefit the system and increase the benefit of introducing service time variability.

Chapter 4

Dual Outsourcing Under Strategic Collaboration of Competing Servers

4.1 Introduction

"Multiple outsourcing" has been adopted by many service or make-to-order production firms (Levina and Su 2008); one of its main advantages is that a client can obtain fast service at a minimum cost by creating competition among multiple service providers or suppliers. A typical practice is for the client to allocate demand based on the performance of service providers or suppliers; for example, a call center may send more calls to a faster agent (de Vericourt and Zhou 2005). In a manufacturing example, Apple's two major suppliers Foxconn and Pegatron have both expanded their workforce to compete for Apple's iPhone 6 production orders (Clover 2014). Obviously, a client can benefit from such a performance-based demand allocation policy as it provides agents incentives to make effort and/or invest in capacities (Benjaafar 2007).

Meanwhile, close attention has also been paid to collaboration/cooperation among service providers or suppliers. For instance, some medical call centers have launched software that facilitates connections among agents, to encourage agents to share their experience and expertise when dealing with common medical problems. While it is believed that such collaboration would improve system efficiency and productivity, concerns arise that when an agent can rely on others to perform, s/he might lose incentives to contribute (Wright 2014). As a result, a tradeoff between efficiency and incentives arises in the presence of both competition and collaboration.

Although such a tradeoff has been examined by previous papers including Gilbert and Weng (1998) and Cachon and Zhang (2007), these papers addressed the problem only from the client's perspective. In particular, these papers aim to understand how a client could maximize its payoff through designing outsourcing contracts to induce different degrees of collaboration and competition. However, in practice service collaborations might not always be guided by the client: Call center agents may voluntarily share their expertise without the client noticing or even knowing about it; alternatively, make-to-order manufacturers may secretly subcontract their allocated orders if the buyer is not able to monitor their production process. In these cases, servers collaborate to maximize their own payoffs, irrespective of the client's wishes. We call such collaboration "*Strategic Collaboration*." To the best of our knowledge, such strategic collaboration has not been examined in the literature.

This chapter studies the impact of strategic collaboration in the context of a queueing model. We model collaboration in terms of capacity sharing: servers might transfer arriving jobs during their service or production process, choosing the degree of capacity sharing according to their own interests, not the client's. In particular, we aim to answer the following questions: First, how would such collaboration affect the suppliers' capacity investment? Second, what is the impact of this strategic collaboration on system performance statistics such as customers' waiting time? Lastly, how should the client respond to such strategic collaboration — should the client prevent such collaboration (e.g. through penalization) or encourage it (e.g. through rewards)?

To answer these research questions I conduct both theoretical and numerical analyses. In the theoretical analysis, I use a two dimensional Markov chain to model a queueing system with two servers (who represent suppliers or agents), and find a Markov Perfect Equilibrium (MPE) to characterize the collaboration equilibrium between the two servers. I find that the MPE of the collaboration game follows a non-monotonic structure: counter intuitively, a server might be less likely to collaborate when there are more jobs waiting for it, or when there are fewer jobs waiting for the other server. Further I conduct simulations to numerically examine the impact of strategic collaboration on a dual-outsourcing strategy. The numerical study shows that even though strategic collaboration reduces capacity investment, it could still reduce waiting by improving capacity utilization. As a result, the client could benefit from strategic collaboration and thus should encourage it.

4.2 Literature Review

This paper is closely related to the literature on dual/multi outsourcing design in service industries. Most papers in this stream focus on examining the performance of different job allocation policies and payment schemes. Benjaafar (2007) and Elahi et al. (2014) study how to enforce competition among multiple service suppliers (servers) via demand allocation and procurement price, but they don't allow servers to collaborate. In contrast, Gilbert and Weng (1998) and Cachon and Zhang (2007) examine the tradeoff between inducing competition and promoting collaboration by implementing state-dependent and state-independent job allocation policies. However, as mentioned in the introduction, both papers assume that the the degree of collaboration is adjusted by the client to maximize its own payoff, whereas we allow servers to cooperate (collaborate) in their own interests.

There are also some papers studying server cooperation in terms of capacity pooling (for example, Anily and Haviv (2009) and Yu et al. (2014)), which are similar to the setting of servers collaboration in our paper. However, while most cooperation papers aim to achieve capacity sharing at a full level, we will show that strategic pooling in our setting is almost always partial. On the other hand, He and Neuts (2002) also study internal job transferring in a partial way, but in their setting such transferring is exogenous, not endogenous as in our setting.

To our knowledge, we are the first to consider servers' strategic collaboration in a competitive environment. Similar to us, Hu et al. (2013) also analyze strategic inventory pooling behavior in a competitive environment, in which a supplier sells products to multiple buyers. Specifically, they examine how competing buyers' sharing demand would affect their common supplier's profit. But Hu et al. (2013) mainly study inventory pooling in a newsvendor setting (faced with static demand). In contrast, our paper analyzes capacity pooling within a queueing context (faced with dynamic demand).

4.3 Model

Model Setup. We consider a client firm outsourcing its service business to two external service providers (servers), which are indexed by $o \in \{i, j\}$. Jobs arrive following a Poisson process with

rate λ.[1] On each job server o spends an exponentially distributed service time with a rate μ_o, which is also referred to as server o's capacity. We denote by $\boldsymbol{\mu} := (\mu_i, \mu_j)$ the service rate vector. Server o with capacity μ_o incurs a capacity cost associated with establishing, maintaining and managing the capacity, regardless of whether the server is busy or idle. The rate of capacity cost is denoted by a function of the capacity $g_o(\mu_o)$; we adopt a linear cost function $g_o(\mu_o) := c_o\mu_o$, where the parameter c_o is the marginal capacity cost of agent o. This satisfies the desirable properties, $g_o(0) = 0$, $g_o'(\cdot) > 0$ and $g_o''(\cdot) \geq 0$.

The client allocates jobs following a static allocation rule; i.e., the client allocates an arriving job to server o with a probability p_o, independent of system state (Benjaafar et al. 2007).[2] In particular, we assume that the probability of an arriving job being allocated to server o is linear in server o's capacity μ_o; i.e., $p_o = \frac{\mu_o}{\mu_i+\mu_j}$, where $o \in \{i, j\}$. We denote by $\mathbf{p} := (p_i, p_j)$ the assignment probability vector. For each job a server serves, the client pays the server a reward R; meanwhile the server pays a penalty h per unit waiting time for each job (Hasija et al. 2008).

The sequence of the events is as follows: First, the client determines R and h to minimize the total payment while maintaining a fixed service level (e.g., a fixed average waiting time). Second, given the allocation rule and payment mechanism, each server determines its service rate and builds capacity accordingly. Finally, after capacities are build, jobs randomly arrive and the two servers start to process their allocated jobs, potentially collaborating with each other in their own interests. We assume that each server's objective is to maximize its expected discounted payoff from its service.

Strategic Collaboration. During the service process, the two servers collaborate by transferring their allocated jobs between themselves, with the aim of maximizing their own payoff, not the client's. While collaboration might take different forms, for analytical tractability we assume that: (i) When a job is transferred, its associated reward and subsequent waiting penalty are also transferred; (ii) Each job may be transferred only upon its arrival (i.e., no jobs in queue are transferred.)[3]

[1] We assume all jobs are eventually served and no jobs renege.

[2] A client could allocate jobs dynamically (i.e., allocating customers based on the number of jobs waiting for each server), but this may not always be possible, for example when the jobs are allocated before they arrive at the system (one such example would be insurance companies that assign customers to hospitals and clinics long before the customers needs to see a doctor).

[3] We adopt this assumption because if a job in queue were to be transferred, it would be often difficult to assign the waiting cost incurred by the job before it is transferred.

This strategic collaboration can be modeled as a stochastic game, which is played as follows: Each time when a job arrives at server $i(j)$, server $i(j)$ needs to decide whether to transfer the newly arrived job to server $j(i)$ or not. Meanwhile, server $j(i)$ decides whether to accept a job that might be transferred from server $i(j)$ or not. A job would be successfully transferred from server $i(j)$ to server $j(i)$ if and only if server $i(j)$ decides to transfer a job and server $j(i)$ decides to accept a transferred job.

Markov Perfect Equilibrium. Each server's strategy specifies its decisions regarding transferring/keeping a job arriving from outside or accepting/rejecting a job transferred by the other server. These actions could be characterized as a function of the preceding history of states and actions. But, for analytical tractability we restrict attention to Markov strategies where servers' decisions depend on the current state alone, allowing us to ignore all other details of the history and thus consider a much smaller set of variables. This is logical, given the Markovian nature of our system. Moreover, we focus on stationary Markov perfect equilibria (MPE), which requires that the Markov strategy yields a Nash equilibrium in every proper subgame. Such Markov perfect equilibria are commonly used in both research and applications due to analytical and/or computational advantages (Mailath and Samuelson 2006).

A stationary Markov perfect equilibrium allows the current action to depend only on the current state, which is represented by m_o: the numbers of jobs waiting for and/or being served by server o, $o \in \{i, j\}$. Thus, server o's strategy could be represented by a function $a_o(m_i, m_j)$. In particular, we use $a_o(m_i, m_j) = 1$ to represent the case that server o would transfer an incoming job to the other server or reject (not accept) a job transferred by the other server, whereas $a_o(m_i, m_j) = 0$ indicates that server o would keep (not transfer) an incoming job to the other server or accept a job transferred by the other server.[4]

Embedded Markov Chain. Since servers collaborate under a Markov perfect equilibrium and both interarrival and service times are exponentially distributed, at times between the arrival and departure event epochs the service process evolves as a continuous time Markov chain. The continuous time Markov Chain can be further transformed into discrete time Markov chain through

[4]Note that it can not happen that in one state a server would choose to transfer an incoming job and also accept a job transferred by the other server. This is because in any state a server would make only two choices when a job arrives: either to add the job into its own queue or leave the job to the other server; in the former case, the server would keep the job if it arrives at the server or accept the job if it arrives at the other server, whereas in the latter case the server would transfer the job if it arrives at the server or reject the job if it arrives at the other server.

uniformization (Puterman 1994). The uniformization rate is denoted by $\Gamma := \lambda_i + \lambda_j + \mu_i + \mu_j$. As a result, $\lambda_i := \frac{\lambda_i}{\Gamma}$ and $\mu_i := \frac{\mu_i}{\Gamma}$ become the probability that the next uniformized transition is an arrival at server i or a departure from server i, respectively. Similarly, $h := \frac{h}{\Gamma}$ is the expected discounted penalty per waiting job in each time unit. Given these uniformized system parameters, the service process can be modeled as two-dimensional Markov Chain with state (m_i, m_j), where m_i and m_j represents the number of jobs waiting for or being served by server i and j, respectively.

Optimal Value Function. Next we provide the expressions for the optimal value function and best response function of each server given the other server's strategy. We take server i as an example. Given server j's strategy $a_j(m_i, m_j)$ (abbreviated as a_j), let $V_i(m_i, m_j, a_j)$ denote server i's optimal value function at state (m_i, m_j), and $\widetilde{a}_i(m_i, m_j, a_j)$ (abbreviated as $\widetilde{a}_i$) denote server i's best response function to server j's strategy. Thus, $V_i(m_i, m_j, a_j)$ satisfies the optimality equation (Ross 1983)

$$\begin{aligned} V_i(m_i, m_j, a_j) &= \mu_i V_i(m_i - 1, m_j, a_j) + \mu_j V_i(m_i, m_j - 1, a_j) \\ &+ \lambda_i \left\{ \widetilde{a}_i(1 - a_j) V_i(m_i, m_j + 1, a_j) + (1 - \widetilde{a}_i(1 - a_j)) (R - hm_i/\mu_i + V_i(m_i + 1, m_j, a_j)) \right\} \\ &+ \lambda_j \left\{ (1 - a_j(1 - \widetilde{a}_i)) V_i(m_i, m_j + 1, a_j) + a_j(1 - \widetilde{a}_i) (R - hm_i/\mu_i + V_i(m_i + 1, m_j, a_j)) \right\}, \end{aligned} \tag{4.1}$$

where server i's best response $\widetilde{a}_i(m_i, m_j, a_j)$ at state (m_i, m_j) is determined as:

$$\widetilde{a}_i(m_i, m_j, a_j) = \begin{cases} 1 & \text{if } R - hm_i/\mu_i + V_i(m_i + 1, m_j, a_j) < V_i(m_i, m_j + 1, a_j); \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

The four terms of the right hand side of (4.1) represent the probabilities and the profits-to-go associated with a job completion and arrival, respectively. For instance, when a job arrives at server i at rate λ_i, $\widetilde{a}_i(1 - a_j)$ represents the probability that the job would be transferred by server i to server j, whereas $1 - \widetilde{a}_i(1 - a_j)$ represents the probability that the job would be kept and served by server i. Note that by convention, when a job arrives at or is transferred to server i with a queue length of m_i, server i both receives the reward R and pays the expected waiting cost for the expected waiting time of the job, hm_i/μ_i.

The characterization of the best-response function in (4.2) is derived based on two facts: First,

when $a_j(m_i, m_j) = 0$, server j would not transfer any incoming job but would accept a transferred job, so server i only needs to decide whether to transfer an incoming job or not, and server i would choose to transfer (i.e., $\widetilde{a}_i(m_i, m_j, a_j) = 1$) if and only if its payoff from keeping the job, $R + V_i(m_i + 1, m_j, a_j)$, is less than its payoff from transferring the job, $V_i(m_i, m_j + 1, a_j)$. Second, when $a_j(m_i, m_j) = 1$, server j would always transfer an incoming job but would reject any transferred job, so server i only needs to decide whether to reject a job transferred by server j, and server i would choose to reject (i.e., $\widetilde{a}_i(m_i, m_j, a_j) = 1$) if and only if its payoff from accepting the job, $R + V_i(m_i + 1, m_j, a_j)$, is less than its payoff from rejecting the job, $V_i(m_i, m_j + 1, a_j)$. For ease of exposition, let

$$\Delta V_i(m_i, m_j, a_j) := V_i(m_i, m_j + 1, a_j) - V_i(m_i + 1, m_j, a_j), \tag{4.3}$$

represent the marginal *future* value server i could obtain at state (m_i, m_j) by successfully transferring an arriving job to server j compared to not transferring it (or by rejecting a job transferred by server j compared to accepting it). Thus, $\widetilde{a}_i(m_i, m_j, a_j)$ could be redefined as

$$\widetilde{a}_i(m_i, m_j, a_j) = \begin{cases} 1 & \text{if } R - hm_i/\mu_i < \Delta V_i(m_i, m_j, a_j); \\ 0 & \text{otherwise.} \end{cases} \tag{4.4}$$

4.4 Equilibrium Analysis

In this section we study the collaboration equilibrium under a finite horizon with n periods. In each period k, server $i(j)$'s decision is denoted by $a^k_{i(j)}(m_i, m_j)$ (abbreviated as $a^k_{i(j)}$): $a^k_{i(j)} = 1$ means that server $i(j)$ would transfer an incoming job to server $j(i)$ or reject a job transferred by server $j(i)$, whereas $a^k_{i(j)} = 0$ indicates that server $i(j)$ would keep an incoming job or accept a job transferred by server $j(i)$. Similarly, in each period k, let $V^k_{i(j)}(m_i, m_j)$ denote server $i(j)$'s value function at state (m_i, m_j), and $\widetilde{a}^k_{i(j)}\left(m_i, m_j, a^k_{j(i)}\right)$ (abbreviated as $\widetilde{a}^k_{i(j)}$) denote server $i(j)$'s best response function to the other server $j(i)$'s strategy. We first analytically study the two servers' collaboration strategies in equilibrium and then numerically demonstrate the structure and

properties of the collaboration equilibrium.

4.4.1 Analytical Study

In this section we will analytically study the two servers' strategies in each period, proceeding backwards. In each period, we first analyze each server's value function, as well as the server's best response function, and then summarize the collaboration equilibrium in each period.

Last period n

We start with server i. At the last period n, server i's payoff is merely the surplus from serving a job, $R-hm_is_i$, where $s_i = 1/\mu_i$ represents the average service time of each job at server i. Moreover, given server i's decision $a_i^n(m_i, m_j)$ and server j's decision $a_j^n(m_i, m_j)$, server i could serve a job if (i) a job arrives at server j but is transferred by server j and accepted by server i, which occurs with a probability $\lambda_j\left(1-a_i^n(m_i,m_j)\right)a_j^n(m_i,m_j)$; (ii) a job arrives at server i and is not transferred to server j successfully, which occurs with a probability $\lambda_i\left(1-a_i^n(m_i,m_j)\right)a_j^n(m_i,m_j)$. As a result, server i's value function in period n is

$$
\begin{aligned}
&V_i^n(m_i,m_j)\\
&=\left\{\lambda_j\left(1-a_i^n(m_i,m_j)\right)a_j^n(m_i,m_j)+\lambda_i\left(1-\left(1-a_j^n(m_i,m_j)\right)a_i^n(m_i,m_j)\right)\right\}(R-hm_is_i)\\
&=-\left\{\lambda_j a_j^n(m_i,m_j)+\lambda_i\left(1-a_j^n(m_i,m_j)\right)\right\}a_i^n(m_i,m_j)(R-hm_is_i)+\lambda_j a_j^n(m_i,m_j)(R-hm_is_i).
\end{aligned}
$$

It is straightforward to derive server i's best response $\widetilde{a}_i^n(m_i,m_j)$ in period n:

$$
\widetilde{a}_i^n(m_i,m_j)=\begin{cases}0 & \text{if } m_i\le\overline{m}_i\\ 1 & \text{if } m_i>\overline{m}_i\end{cases},
$$

where

$$
\overline{m}_i\equiv R/(hs_i)
$$

is a threshold value at which server i makes zero surplus from serving a job. The intuition is that server i would like to keep an incoming job or accept a transferred job (i.e., $\widetilde{a}_i^n(m_i,m_j)=0$) if and only if server i would make a non-negative profit from serving the job (i.e., $R-hs_im_i\ge 0\Leftrightarrow m_i\le$

$\overline{m}_i$), otherwise server i would prefer to transferring an incoming job or rejecting a transferred job (i.e., $\widetilde{a}_i^n(m_i, m_j) = 1$).

Similarly server j's value function at the last period n is

$$\begin{aligned} & V_j^n(m_i, m_j) \\ & = \left\{\lambda_j\left(1-\left(1-a_i^n(m_i,m_j)\right)a_j^n(m_i,m_j)\right)+\lambda_i\left(1-a_j^n(m_i,m_j)\right)a_i^n(m_i,m_j)\right\}(R-hm_js_j), \end{aligned}$$

where, again $s_j = 1/\mu_j$ represents the average service time of each job at server j.

Accordingly, server j's best response in period n:

$$\widetilde{a}_j^n(m_i, m_j) = \begin{cases} 0 & \text{if } m_j \leq \overline{m}_j; \\ 1 & \text{if } m_j > \overline{m}_j, \end{cases}$$

where

$$\overline{m}_j \equiv R/(hs_j)$$

is a threshold value at which server j makes zero surplus from serving a job.[5]

As a result, we have the following result for the collaboration equilibrium in period n:

Lemma 4 *In period n, two servers collaborate in terms of transferring jobs under the following conditions*

(a) $m_i \leq \overline{m}_i$ and $m_j > \overline{m}_j$, under which an job arriving at server j would be transferred by server j and accepted by server i (i.e., $a_i^n(m_i, m_j) = 0$ and $a_j^n(m_i, m_j) = 1$);

(b) $m_i > \overline{m}_i$ and $m_j \leq \overline{m}_j$, under which an job arriving at server i would be transferred by server i and accepted by server j (i.e., $a_i^n(m_i, m_j) = 1$ and $a_j^n(m_i, m_j) = 0$).

In other words, server j would transfer an incoming job to server i and server i would accept the job if and only if server i has no more waiting jobs than the benchmark $\overline{m}_i$ and server j has more waiting jobs than the benchmark $\overline{m}_j$, and vice versa. Lemma 4 indicates that in period n, each server's strategy follows a constant threshold structure, and the value of the threshold only depends on the reward R, the waiting penalty h, and the average service time of the server itself.

[5]For analytical simplicity, here we assume that $\overline{m}_i$ and $\overline{m}_j$ are both integers.

Correspondingly, server i's and server j's value functions in period n are

$$V_i^n(m_i, m_j) = \begin{cases} (\lambda_i + \lambda_j)(R - hs_i m_i) & \text{if } m_i \le \overline{m}_i \text{ and } m_j > \overline{m}_j; \\ \lambda_i (R - hs_i m_i) & \text{if } \{m_i \le \overline{m}_i \text{ and } m_j \le \overline{m}_j\} \text{ or } \{m_i > \overline{m}_i \text{ and } m_j > \overline{m}_j\}; \\ 0 & \text{if } m_i > \overline{m}_i \text{ and } m_j \le \overline{m}_j. \end{cases} \tag{4.5}$$

$$V_j^n(m_i, m_j) = \begin{cases} 0 & \text{if } m_i \le \overline{m}_i \text{ and } m_j > \overline{m}_j; \\ \lambda_j (R - hs_j m_j) & \text{if } \{m_i \le \overline{m}_i \text{ and } m_j \le \overline{m}_j\} \text{ or } \{m_i > \overline{m}_i \text{ and } m_j > \overline{m}_j\}; \\ (\lambda_i + \lambda_j)(R - hs_j m_j) & \text{if } m_i > \overline{m}_i \text{ and } m_j \le \overline{m}_j. \end{cases} \tag{4.6}$$

By analyzing the value functions we can infer the impact of collaboration in period n. Here we take server i as an example and compare the payoff in (4.5) with server i's payoff in the absence of collaboration, $\lambda_i (R - hs_i m_i)$. The comparison indicates that server i's value function is affected by the collaboration in the following ways. First, through collaboration server i could make an extra profit from serving jobs transferred by server j, especially when server i has a few jobs waiting and server j has many (i.e., $m_i \le \overline{m}_i$ and $m_j > \overline{m}_j$). Second, server i could also benefit from collaboration by avoiding a high waiting penalty through transferring jobs to server j when server i is overloaded (i.e., $m_i > \overline{m}_i$). However, this can only occur when server j has a few jobs waiting (i.e., $m_j \le \overline{m}_j$); Otherwise when server j also has many jobs (i.e., $m_j > \overline{m}_j$), server j would reject the jobs transferred by server i, leaving server i with a loss (i.e., $\lambda_i (R - hs_i m_i) \le 0$ for $m_i > \overline{m}_i$).

Similar to (4.3), we define $\Delta V_i^n(m_i, m_j) \equiv V_i^n(m_i, m_j + 1) - V_i^n(m_i + 1, m_j)$, which represents the marginal value server i could obtain at state (m_i, m_j) in period n by successfully transferring an incoming job to server j compared to keeping the job (or by rejecting a job transferred by server j compared to accepting the job) in period $n-1$, one period before period n. Referring to (4.5),

there are six possible values of $\Delta V_i^n(m_i, m_j)$, depending on six different sets of $\{m_i, m_j\}$:

$$\Delta V_i^n(m_i, m_j) \tag{4.7}$$

$$= V_i^n(m_i, m_j + 1) - V_i^n(m_i + 1, m_j)$$

$$= \begin{cases} (\lambda_i + \lambda_j) hs_i & \text{(i) } m_i \leq \overline{m}_i - 1 \text{ and } m_j > \overline{m}_j \\ \lambda_j (R - hs_i m_i) + \lambda_i hs_i & \text{(ii) } \{m_i = \overline{m}_i \text{ and } m_j > \overline{m}_j\} \text{ or } \{m_i < \overline{m}_i \text{ and } m_j = \overline{m}_j\} \\ (\lambda_i + \lambda_j)(R - hs_i m_i) & \text{(iii) } m_i = \overline{m}_i \text{ and } m_j = \overline{m}_j \\ \lambda_i hs_i & \text{(iv) } \{m_i > \overline{m}_i \text{ and } m_j > \overline{m}_j\} \text{ or } \{m_i < \overline{m}_i \text{ and } m_j < \overline{m}_j\} \\ \lambda_i (R - hs_i m_i) & \text{(v) } \{m_i = \overline{m}_i \text{ and } m_j < \overline{m}_j\} \text{ or } \{m_i > \overline{m}_i \text{ and } m_j = \overline{m}_j\} \\ 0 & \text{(vi) } m_i > \overline{m}_i \text{ and } m_j < \overline{m}_j \end{cases}. \tag{4.8}$$

From (4.8) we can infer how server i's value in period n is changed by server i's decision to transfer/reject a job compared to keep/accept a job in period $n-1$. In brief, if server i transfers or rejects a job in period $n-1$, in period n server i would have one fewer job in its queue but leave one more job in server j's queue. According to (4.8), such a change could bring a difference to server i's value in period n in the following four ways: First, due to one fewer job in server i's queue, server i could reduce the waiting cost of jobs served by server i in period n by an amount of hs_i per job (see case (i), (ii) and (iv)). Second, since server j has one more job waiting and server i has one fewer, server i could earn an chance to receive jobs from server j in period n (see case (ii) and (iii)). Thirdly, due to one more job in server j's queue, server i might incur a loss caused by server j's rejection of jobs transferred by server i (see case (v) in which $\lambda_i (R - hs_i m_i) < 0$ for $m_i > \overline{m}_i$). Lastly, server i might not be affected at all because no matter whether server i gives away a job in period $n-1$ or not, in period n server i would transfer an incoming job due to a large enough m_i and would not receive any job from server j due to a low enough m_j (see case (vi)).

Similarly, define $\Delta V_j^n(m_i, m_j) \equiv V_j^n(m_i + 1, m_j) - V_j^n(m_i, m_j + 1)$, which represents the marginal value server j could obtain at state (m_i, m_j) in period n by successfully transferring an incoming job to server i compared to keeping the job (or by rejecting a job transferred by server i compared

to accepting the job) in period $n-1$. Referring to (4.6),

$$\Delta V_j^n(m_i, m_j) \equiv V_j^n(m_i+1, m_j) - V_j^n(m_i, m_j+1)$$
$$= \begin{cases} 0 & \text{(i) } m_i \le \overline{m}_i - 1 \text{ and } m_j > \overline{m}_j \\ \lambda_j(R - hs_j m_j) & \text{(ii) } \{m_i = \overline{m}_i \text{ and } m_j > \overline{m}_j\} \text{ or } \{m_i < \overline{m}_i \text{ and } m_j = \overline{m}_j\} \\ (\lambda_i + \lambda_j)(R - hs_j m_j) & \text{(iii) } m_i = \overline{m}_i \text{ and } m_j = \overline{m}_j \\ \lambda_j hs_j & \text{(iv) } \{m_i > \overline{m}_i \text{ and } m_j > \overline{m}_j\} \text{ or } \{m_i < \overline{m}_i \text{ and } m_j < \overline{m}_j\} \\ \lambda_i(R - hs_j m_j) + \lambda_j hs_j & \text{(v) } \{m_i = \overline{m}_i \text{ and } m_j < \overline{m}_j\} \text{ or } \{m_i > \overline{m}_i \text{ and } m_j = \overline{m}_j\} \\ (\lambda_i + \lambda_j) hs_j & \text{(vi) } m_i > \overline{m}_i \text{ and } m_j < \overline{m}_j \end{cases}. \tag{4.9}$$

The second to last period $n-1$

Similar to the analysis in period n, we start with server i.

Server i**.** At the second last period $n-1$, similar to the optimality equation (4.1), given server i's decision $a_i^{n-1}(m_i, m_j)$ and server j's decision $a_j^{n-1}(m_i, m_j)$, server i's value function is

$$\begin{aligned} V_i^{n-1}(m_i, m_j) = {} & \lambda_j\left(1 - a_i^{n-1}(m_i, m_j)\right) a_j^{n-1}(m_i, m_j)\left(R - hm_i s_i + V_i^n(m_i+1, m_j)\right) \\ & + \lambda_j\left(1 - \left(1 - a_i^{n-1}(m_i, m_j)\right) a_j^{n-1}(m_i, m_j)\right) V_i^n(m_i, m_j+1) \\ & + \lambda_i\left(1 - \left(1 - a_j^{n-1}(m_i, m_j)\right) a_i^{n-1}(m_i, m_j)\right)\left(R - hm_i s_i + V_i^n(m_i+1, m_j)\right) \\ & + \lambda_i\left(1 - a_j^{n-1}(m_i, m_j)\right) a_i^{n-1}(m_i, m_j) V_i^n(m_i, m_j+1) \\ & + \mu_i\left(\mathbb{1}(m_i = 0) V_i^n(m_i, m_j) + (1 - \mathbb{1}(m_i = 0)) V_i^n(m_i - 1, m_j)\right) \\ & + \mu_j\left(\mathbb{1}(m_j = 0) V_i^n(m_i, m_j) + (1 - \mathbb{1}(m_j = 0)) V_i^n(m_i, m_j - 1)\right), \end{aligned}$$

where $\mathbb{1}(m_i = 0)$ and $\mathbb{1}(m_j = 0)$ are indicator functions such that

$$\mathbb{1}(m_i = 0) = \begin{cases} 1 & \text{for } m_i = 0 \\ 0 & \text{for } m_i > 0 \end{cases} \quad \text{and} \quad \mathbb{1}(m_j = 0) = \begin{cases} 1 & \text{for } m_j = 0 \\ 0 & \text{for } m_j > 0 \end{cases}.$$

The indicator functions are introduced to include the special cases that server $i(j)$ has an empty queue (i.e., $m_{i(j)} = 0$), under which at the rate $\mu_{i(j)}$ the state (m_i, m_j) would not change.

Again, in period $n-1$ server i would accept a job if and only if server i would not incur any loss from serving the job (i.e., $R - hs_i m_i + V_i^n(m_i+1, m_j) \geq V_i^n(m_i, m_j+1)$). Thus server i's best response is

$$\widetilde{a}_i^{n-1}(m_i, m_j) = \begin{cases} 0 & \text{if } R - hs_i m_i \geq \Delta V_i^n(m_i, m_j) \\ 1 & \text{if } R - hs_i m_i < \Delta V_i^n(m_i, m_j) \end{cases}.$$

Further referring to (4.8), we can derive the condition for $\widetilde{a}_i^{n-1}(m_i, m_j) = 0$ as follows:

$$\begin{aligned}
&R - hs_i m_i \geq \Delta V_i^n(m_i, m_j) \\
&\Leftrightarrow R - hs_i m_i \geq \\
&\begin{cases} (\lambda_i + \lambda_j) hs_i & \text{if } m_i \leq \overline{m}_i - 1 \text{ and } m_j > \overline{m}_j \\ \lambda_j (R - hs_i m_i) + \lambda_i hs_i & \text{if } \{m_i = \overline{m}_i \text{ and } m_j > \overline{m}_j\} \text{ or } \{m_i < \overline{m}_i \text{ and } m_j = \overline{m}_j\} \\ (\lambda_i + \lambda_j)(R - hs_i m_i) & \text{if } m_i = \overline{m}_i \text{ and } m_j = \overline{m}_j \\ \lambda_i hs_i & \text{if } \{m_i > \overline{m}_i \text{ and } m_j > \overline{m}_j\} \text{ or } \{m_i < \overline{m}_i \text{ and } m_j < \overline{m}_j\} \\ \lambda_i (R - hs_i m_i) & \text{if } \{m_i = \overline{m}_i \text{ and } m_j < \overline{m}_j\} \text{ or } \{m_i > \overline{m}_i \text{ and } m_j = \overline{m}_j\} \\ 0 & \text{if } m_i > \overline{m}_i \text{ and } m_j < \overline{m}_j \end{cases} \\
&\Leftrightarrow \begin{cases} m_i \leq \overline{m}_i - (\lambda_i + \lambda_j) & \text{if } m_i \leq \overline{m}_i - 1 \text{ and } m_j > \overline{m}_j \\ m_i \leq \overline{m}_i - \frac{\lambda_i}{1-\lambda_j} & \text{if } \{m_i = \overline{m}_i \text{ and } m_j > \overline{m}_j\} \text{ or } \{m_i < \overline{m}_i \text{ and } m_j = \overline{m}_j\} \\ m_i \leq \overline{m}_i & \text{if } m_i = \overline{m}_i \text{ and } m_j = \overline{m}_j \\ m_i \leq \overline{m}_i - \lambda_i & \text{if } \{m_i > \overline{m}_i \text{ and } m_j > \overline{m}_j\} \text{ or } \{m_i < \overline{m}_i \text{ and } m_j < \overline{m}_j\} \\ m_i \leq \overline{m}_i & \text{if } \{m_i = \overline{m}_i \text{ and } m_j < \overline{m}_j\} \text{ or } \{m_i > \overline{m}_i \text{ and } m_j = \overline{m}_j\} \\ m_i \leq \overline{m}_i & \text{if } m_i > \overline{m}_i \text{ and } m_j < \overline{m}_j \end{cases} \\
&\Leftrightarrow \begin{cases} m_i \leq \overline{m}_i - 1 \text{ and } m_j > \overline{m}_j \\ m_i \leq \overline{m}_i - \frac{\lambda_i}{1-\lambda_j} \text{ and } m_j = \overline{m}_j \\ m_i = \overline{m}_i \text{ and } m_j = \overline{m}_j \\ m_i \leq \overline{m}_i - \lambda_i \text{ and } m_j < \overline{m}_j \\ m_i = \overline{m}_i \text{ and } m_j < \overline{m}_j \end{cases}.
\end{aligned}$$

Note that $\lambda_i < 1$ and $\frac{\lambda_i}{1-\lambda_j} \leq 1$. Thus, the above condition could be further simplified as

$$R - hs_i m_i \geq \Delta V_i^n (m_i, m_j) \Leftrightarrow \begin{cases} m_i \leq \overline{m}_i - 1 \text{ and } m_j > \overline{m}_j \\ m_i \leq \overline{m}_i - 1 \text{ and } m_j = \overline{m}_j \\ m_i \leq \overline{m}_i - 1 \text{ and } m_j < \overline{m}_j \\ m_i = \overline{m}_i \text{ and } m_j \leq \overline{m}_j \end{cases},$$

which leads to the following result on server i's best response function:

$$\widetilde{a}_i^{n-1} (m_i, m_j) = \begin{cases} 0 & \text{if } \{m_i \leq \overline{m}_i - 1 \text{ and } m_j > \overline{m}_j\} \text{ or } \{m_i \leq \overline{m}_i \text{ and } m_j \leq \overline{m}_j\}; \\ 1 & \text{otherwise.} \end{cases} \tag{4.10}$$

Obviously, server i's strategy in period $n-1$ is different from that at period n in that the threshold decreases from $\overline{m}_i$ to $\overline{m}_i - 1$ as server j's queue length exceeds $\overline{m}_j$.

Server j. Server j's value function is similar to server i: given server i's decision $a_i^{n-1} (m_i, m_j)$, server j's value function is:

$$\begin{aligned} V_j^{n-1} (m_i, m_j) = {} & \lambda_j \left(1 - \left(1 - a_i^{n-1} (m_i, m_j)\right) a_j^{n-1} (m_i, m_j)\right) \left(R - hm_j s_j + V_j^n (m_i + 1, m_j)\right) \\ & + \lambda_j \left(1 - a_i^{n-1} (m_i, m_j)\right) a_j^{n-1} (m_i, m_j) V_j^n (m_i, m_j + 1) \\ & + \lambda_i \left(1 - a_j^{n-1} (m_i, m_j)\right) a_i^{n-1} (m_i, m_j) \left(R - hm_j s_j + V_j^n (m_i + 1, m_j)\right) \\ & + \lambda_i \left(1 - \left(1 - a_j^{n-1} (m_i, m_j)\right) a_i^{n-1} (m_i, m_j)\right) V_j^n (m_i, m_j + 1) \\ & + \mu_i \left(\mathbb{1}(m_i = 0) V_j^n (m_i, m_j) + (1 - \mathbb{1}(m_i = 0)) V_j^n (m_i - 1, m_j)\right) \\ & + \mu_j \left(\mathbb{1}(m_j = 0) V_j^n (m_i, m_j) + (1 - \mathbb{1}(m_j = 0)) V_j^n (m_i, m_j - 1)\right). \end{aligned}$$

Similar to server i, server j's best response is

$$\widetilde{a}_j^{n-1} (m_i, m_j) = \begin{cases} 0 & \text{if } R - hs_j m_j \geq \Delta V_j^n (m_i, m_j) \\ 1 & \text{if } R - hs_j m_j < \Delta V_j^n (m_i, m_j) \end{cases}.$$

Referring to (4.9), server j's best response $\widetilde{a}_j^{n-1}(m_i, m_j) = 0$ if and only if

$$
\begin{aligned}
& R - hs_j m_j \geq \Delta V_j^n(m_i, m_j) \\
& \Leftrightarrow R - hs_j m_j \geq \\
& \begin{cases}
0 & \text{if } m_i \leq \overline{m}_i - 1 \text{ and } m_j > \overline{m}_j \\
\lambda_j (R - hs_j m_j) & \text{if } \{m_i = \overline{m}_i \text{ and } m_j > \overline{m}_j\} \text{ or } \{m_i < \overline{m}_i \text{ and } m_j = \overline{m}_j\} \\
(\lambda_i + \lambda_j)(R - hs_j m_j) & \text{if } m_i = \overline{m}_i \text{ and } m_j = \overline{m}_j \\
\lambda_j hs_j & \text{if } \{m_i > \overline{m}_i \text{ and } m_j > \overline{m}_j\} \text{ or } \{m_i < \overline{m}_i \text{ and } m_j < \overline{m}_j\} \\
\lambda_i (R - hs_j m_j) + \lambda_j hs_j & \text{if } \{m_i = \overline{m}_i \text{ and } m_j < \overline{m}_j\} \text{ or } \{m_i > \overline{m}_i \text{ and } m_j = \overline{m}_j\} \\
(\lambda_i + \lambda_j) hs_j & \text{if } m_i > \overline{m}_i \text{ and } m_j < \overline{m}_j
\end{cases} \\
& \Leftrightarrow \begin{cases} m_i \leq \overline{m}_i \text{ and } m_j \leq \overline{m}_j \\ m_i > \overline{m}_i \text{ and } m_j \leq \overline{m}_j - 1 \end{cases}
\end{aligned}
$$

Similarly, server j's best response function:

$$
\widetilde{a}_j^{n-1}(m_i, m_j) = \begin{cases} 0 & \text{if } \{m_i \leq \overline{m}_i \text{ and } m_j \leq \overline{m}_j\} \text{ or } \{m_i > \overline{m}_i \text{ and } m_j \leq \overline{m}_j - 1\}; \\ 1 & \text{otherwise.} \end{cases} \tag{4.11}
$$

(4.11) indicates that in period $n-1$ server j's threshold decreases from $\overline{m}_j$ to $\overline{m}_j - 1$ as server i's queue length exceeds $\overline{m}_i$.

The two servers' best response functions (4.10) and (4.11) together indicate that

Lemma 5 *In period $n-1$, two servers collaborate in terms of transferring jobs under the following conditions*

(a) $m_i \leq \overline{m}_i - 1$ and $m_j > \overline{m}_j$, under which an job arriving at server j would be transferred by server j and accepted by server i (i.e., $a_i^n(m_i, m_j) = 0$ and $a_j^n(m_i, m_j) = 1$);

(b) $m_i > \overline{m}_i$ and $m_j \leq \overline{m}_j - 1$, under which an job arriving at server i would be transferred by server i and accepted by server j (i.e., $a_i^n(m_i, m_j) = 1$ and $a_j^n(m_i, m_j) = 0$).

Comparing the collaboration structure in the above Lemma with that in Lemma 4, we can see that the state space for successful transferring shrinks in period $n-1$ compared to that in period n, indicating that less transferring occurs when the time period gets further away from the end.

Correspondingly, server i and server j's value functions are

$$
\begin{aligned}
&V_i^{n-1}(m_i, m_j) \\
&= \begin{cases} (\lambda_i+\lambda_j)\left(R - hm_i s_i + V_i^n(m_i+1, m_j)\right) & \text{if } m_i \le \overline{m}_i - 1 \text{ and } m_j > \overline{m}_j; \\ (\lambda_i+\lambda_j) V_i^n(m_i, m_j+1) & \text{if } m_i > \overline{m}_i \text{ and } m_j \le \overline{m}_j - 1; \\ \lambda_i\left(R - hm_i s_i + V_i^n(m_i+1, m_j)\right) + \lambda_j V_i^n(m_i, m_j+1) & \text{otherwise.} \end{cases} \\
&\quad +\mu_i\left(\mathbb{1}(m_i=0) V_i^n(m_i, m_j) + (1-\mathbb{1}(m_i=0)) V_i^n(m_i-1, m_j)\right) \\
&\quad +\mu_j\left(\mathbb{1}(m_j=0) V_i^n(m_i, m_j) + (1-\mathbb{1}(m_j=0)) V_i^n(m_i, m_j-1)\right)
\end{aligned} \tag{4.12}
$$

$$
\begin{aligned}
&V_j^{n-1}(m_i, m_j) \\
&= \begin{cases} (\lambda_i+\lambda_j) V_i^n(m_i+1, m_j) & \text{if } m_i \le \overline{m}_i - 1 \text{ and } m_j > \overline{m}_j; \\ (\lambda_i+\lambda_j)\left(R - hm_j s_j + V_j^n(m_i, m_j+1)\right) & \text{if } m_i > \overline{m}_i \text{ and } m_j \le \overline{m}_j - 1; \\ \lambda_i V_j^n(m_i+1, m_j) + \lambda_j\left(R - hm_j s_j + V_j^n(m_i, m_j+1)\right) & \text{otherwise.} \end{cases} \\
&\quad +\mu_i\left(\mathbb{1}(m_i=0) V_j^n(m_i, m_j) + (1-\mathbb{1}(m_i=0)) V_j^n(m_i-1, m_j)\right) \\
&\quad +\mu_j\left(\mathbb{1}(m_j=0) V_j^n(m_i, m_j) + (1-\mathbb{1}(m_j=0)) V_j^n(m_i, m_j-1)\right)
\end{aligned} \tag{4.13}
$$

Similar to period n, we could derive the expression of $\Delta V_i^{n-1}(m_i, m_j)$ and $\Delta V_j^{n-1}(m_i, m_j)$ based on (4.12) and (4.13). However, the complicated structure of the value functions $V_i^{n-1}(m_i, m_j)$ and $V_j^{n-1}(m_i, m_j)$ shown in (4.12) and (4.13) will yield a complicated structure of $\Delta V_i^{n-1}(m_i, m_j)$ and $\Delta V_j^{n-1}(m_i, m_j)$, and likewise for the two servers' best response functions in period $n-2$. We can also infer that the structure will become more complicated as the period gets further away from the end. In fact it is analytically intractable for a case with a general number of periods. Therefore, we turn to numerical analysis of the collaboration equilibrium in terms of server i's threshold for accepting and server j's threshold for transferring.

4.4.2 Numerical Study

Since theoretical results can only provide limited information about the structure of the collaboration equilibrium, in this section we aim to examine the equilibrium's structure, as well as the

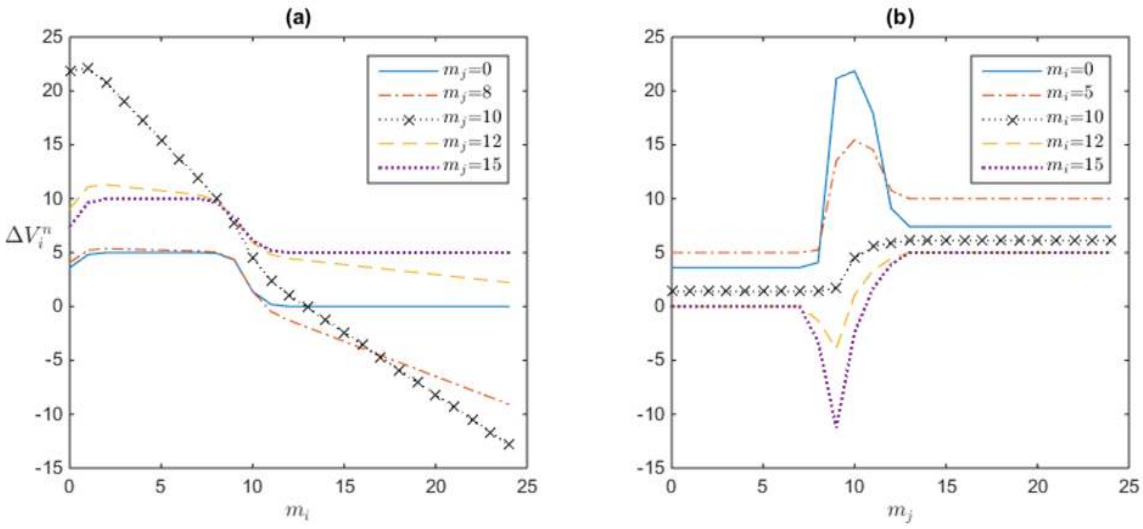

Figure 4.1: How $\Delta V_i^k(m_i, m_j)$ changes with m_i and m_j

equilibrium's sensitivity to other parameters, numerically. With this purpose, we solve the optimal value functions in each period in a backward manner, for a number of cases over a wide range of parameter combinations. We first show the properties of the difference in the value function $\Delta V_{i(j)}^k(m_i, m_j)$, then the structure of the thresholds $\overline{m}_i$ and $\overline{m}_j$, and finally the sensitivity of $\overline{m}_i$ and $\overline{m}_j$ with respect to other parameters: the queue length of the other server, the arrival rate λ_j, the two service rates μ_i and μ_j, the reward-to-penalty ratio R/h and the period index k.

Unless otherwise noted, our parameters are as follows: $\lambda_i = \lambda_j = 1$, $\mu_j = \mu_i = 2$, $R/h = 5$, and $N = 60$. For simplicity, we study the sensitivity of the equilibrium structure to R/h instead of R and h individually.

The results of this numerical study are summarized in the following observations.

Observation 1: *Fixing all the system parameters,*

(a) as $m_{i(j)}$ increases, the value difference $\Delta V_{i(j)}^k(m_i, m_j)$ first increases and then decreases;

(b) as $m_{j(i)}$ increases, the value difference $\Delta V_{i(j)}^k(m_i, m_j)$ first increases and then decreases for small $m_{i(j)}$, but first decreases and then increases for large $m_{i(j)}$.

Due to symmetry, here we only analyze $\Delta V_i^k(m_i, m_j)$ from server i's perspective, which is demonstrated in Figure 4.1.[6] As explained earlier, $\Delta V_i^k(m_i, m_j)$ can be interpreted as the marginal *future* value server i could obtain at state (m_i, m_j) by successfully transferring an incoming job to server

[6]Here we choose the third to last period (i.e. $k = N - 2$) for demonstration.

j compared to keeping the job (or by rejecting a job transferred by server j compared to accepting the job). We can also use $\Delta V_i^k(m_i, m_j)$ to represent server i's incentive to transfer/reject jobs. In particular, server i's future value would be affected by its transferring/rejection in the following three ways: (i) server i could benefit from transferring/rejection by pushing server j's queue length to approach and even exceed its threshold, leading server j to transfer a job in the future; (ii) server i might incur a loss from transferring/rejection as giving a job away to server j now might lead server j to reject jobs transferred by server i in the future when server i needs to transfer jobs to avoid high waiting penalties under a long queue; (iii) transferring/rejection might also bring a loss to server i by increasing the chance of server i being idle in the future, resulting a waste of server i's capacity. This occurs even though server i earns revenue only upon the arrival of jobs; if server i is idle the transition due to a completion at i is to the same state, lowering the incremental value of this transition.

Now we explain the impact of m_i on the function $\Delta V_i^k(m_i, m_j)$. Figure 4.1(a) suggests that there are two types of impacts. First, $\Delta V_i^k(m_i, m_j)$ usually decreases in m_i, indicating that under a larger m_i server i will typically obtain less future value from transferring/rejecting a job. This is because any benefit server i may receive from transferring/rejection by pushing server j to transfer jobs in the future decreases as m_i increases due to a higher waiting cost under a longer queue. Similarly, when server i incurs a loss from transferring/rejection by pushing server j to reject jobs in the future, the loss increases as m_i increases, again caused by a higher waiting cost under a longer queue. Second, Figure 4.1(a) suggests that when m_i is small $\Delta V_i^k(m_i, m_j)$ might increase in m_i. The reason is that when m_i is small (e.g, close to 0), server i faces a high risk of being idle in the future, which greatly dampens server i's incentive to transfer/reject jobs. But a larger m_i reduces the future risk of being idle, and thus cuts server i's loss (or increases its future value) from transferring/rejecting a job.

Next we study the impact of m_j through Figure 4.1(b), which suggests that the impact of m_j on $\Delta V_i^k(m_i, m_j)$ interacts with the magnitude of m_i: when m_i is small (e.g., $m_i = 0$ or 5) $\Delta V_i^k(m_i, m_j)$ first increases and then decreases in m_j, whereas when m_i is large (e.g., $m_i = 12$ or 15) $\Delta V_i^k(m_i, m_j)$ first decreases and then increases in m_j. These opposite impacts are caused by the different roles of server i in the collaboration game under different values of m_i.

On the one hand, when m_i is small, server i benefits more from accepting jobs than transferring

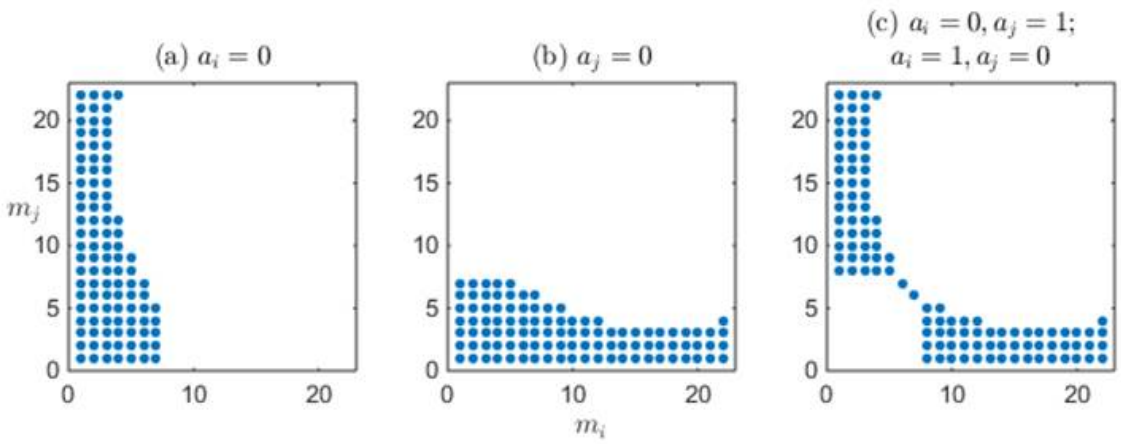

Figure 4.2: The structure of server i's best response $a_i^k(m_i, m_j)$, server j's best response $a_j^k(m_i, m_j)$, and collaboration equilibrium

jobs, and thus server i acts more like a job receiver rather than a job transferrer. In this case the marginal future value, $\Delta V_i^k(m_i, m_j)$, is significantly affected by server i's marginal benefit from rejection, which is achieved by pushing server j to transfer jobs in the future. Furthermore, the benefit would first increase in m_j, because when m_j is below the threshold server i's rejection is more likely to push m_j to exceed the threshold, leading server j to transfer jobs sooner in the future. However, once m_j reaches the threshold, as m_j further increases server i's benefit from rejection diminishes; it does not matter whether server i rejects a job or not since server j would be likely to transfer a job in the future either way.

On the other hand, when m_i is large server i acts more likely a job transferrer rather than a job receiver, in which case $\Delta V_i^k(m_i, m_j)$ is significantly affected by server i's loss from transferring, caused by pushing server j to reject jobs transferred by server i in the future. In this regard, when m_j is small (i.e., below the threshold), as m_j increases (i.e., approaches to the threshold) the loss becomes more significant as transferring would push server j to reject jobs sooner in the future. However, once m_j becomes large enough (i.e., larger than the threshold), as m_j further increases server i's loss diminishes because server j would always reject a job in the near future no matter whether server i transfers a job or not now.

Observation 2: *Fix all the system parameters, as $m_{j(i)}$ increases, typically the threshold $\overline{m}_{i(j)}$ first decreases, then increases and finally becomes a constant value (see Figure 4.2).*

In particular, the dots in Figure 4.2(a) represent the state space for server i to accept jobs, those in (b) represent the state space for server j to accept jobs, and those in (c) represent the overlapping state space in (a) and (b) for jobs to be successfully transferred between the two servers. From Figure 4.2 we can see that the two servers' decisions are symmetric because of the same parameter settings. Thus we focus on the example of server i to explain the structure of the thresholds.

Observation 2(a) suggests that server i's threshold $\overline{m}_i$ is non-monotonic in m_j. The non-monotonicity is reflected in Figure 4.2(a) as a hollow in the state space; we refer it as a "hollow effect." The hollow effect indicates that server i is more likely to transfer/reject jobs as m_j increases from 0, but becomes less likely to do so as m_j increases from a sufficiently large value. This is consistent with the discussion of Observation 1 where we noted that under a small m_i server i's future value from transferring/rejecting jobs, $\Delta V_i^k(m_i, m_j)$, first increases and then decreases in m_j, due to the fact that under a small m_i server i typically acts as a job receiver and thus benefits from rejecting by pushing server j to transfer jobs in the future, and the benefit first increases and then decreases in m_j. However, Figure 4.2 doesn't reflect the other part of Observation 1 that when m_i is large $\Delta V_i^k(m_i, m_j)$ might first decrease and then increase in m_j, in which case server i typically acts as a job transferrer and determines its threshold $\overline{m}_i$ through its incentive to transfer jobs. The comparison indicates that the value of the threshold $\overline{m}_i$ is dominated by server i's incentive to receive jobs rather than that to transfer jobs. Furthermore, in the following analysis of the threshold's sensitivity with respect to arrival rates, we will see that the hollow effect emerges only when a server acts as a job receiver but disappears when a server acts as a job transferrer.

As a result of the non-monotonic structure, there are two interesting observations in the collaboration equilibrium shown in Figure 4.2(c). First, Figure 4.2(c) suggests that given the same queue length for server i, jobs might be transferred successfully from j to i when server j has a short queue instead of a long queue (e.g., in Figure 4.2(c) jobs are transferred at state ($m_i = 4$, $m_j = 8$) but not at the state ($m_i = 4$, $m_j = 11$)). This is because as server j's queue length m_j increases (from a small value), server i's incentive to reject jobs increases, with a purpose to push server j to transfer more jobs in the future. Or, jobs might be transferred successfully from j to i when server i has a long queue instead of a short queue given the same queue length for server j (e.g., in Figure 4.2(c), jobs are transferred at state ($m_i = 5$, $m_j = 6$) but not at the state ($m_i = 0$, $m_j = 6$)), simply because when server i's queue length m_i decreases, server j's incentive

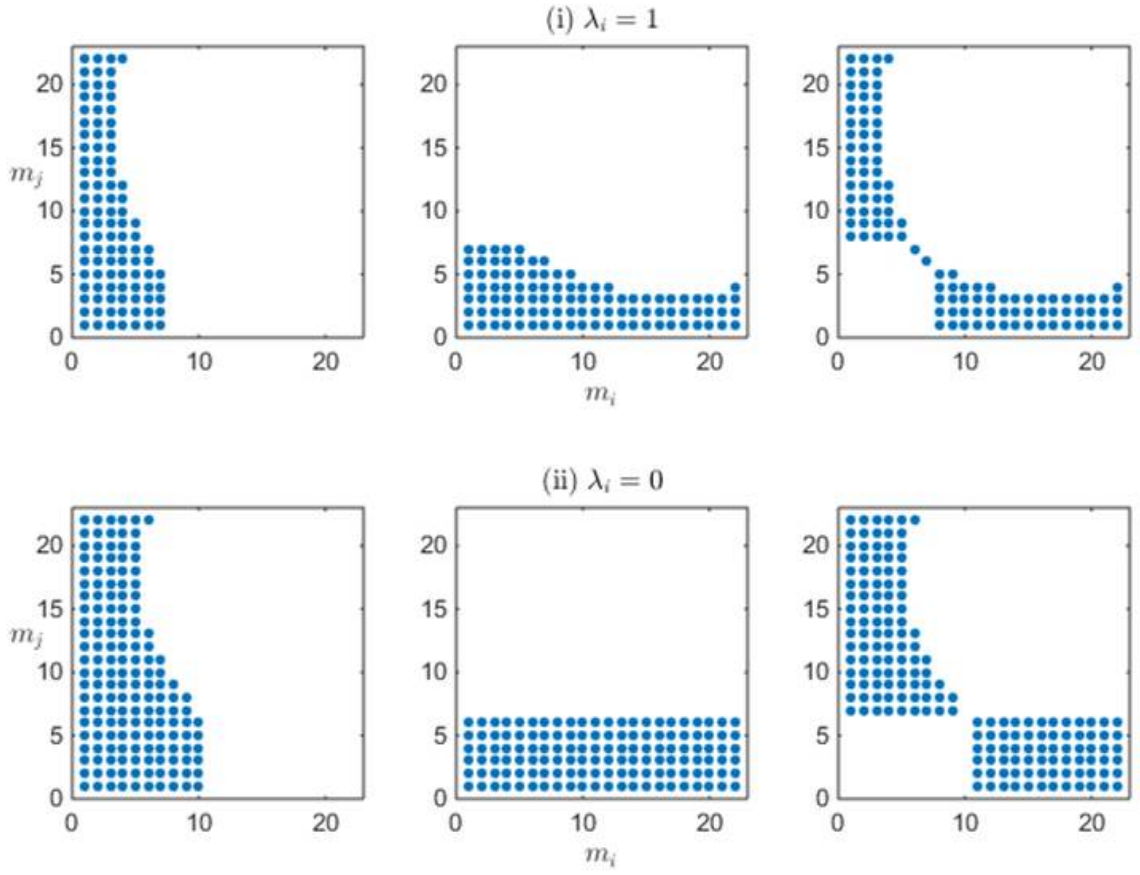

Figure 4.3: How λ_i affects the structure of server i's best response $a_i^k(m_i, m_j)$, server j's best response $a_j^k(m_i, m_j)$, and collaboration equilibrium

to transfer jobs diminishes, knowing that it is less likely to receive jobs from server i in the future. Such observations contradicts the intuition that the two servers would be more likely to transfer jobs when server j has more jobs waiting or server i has fewer jobs waiting.

Next we study how the two servers' strategy and the collaboration equilibrium are affected by the parameters including service rates μ_i and μ_j, arrival rates λ_i and λ_j, immediate reward R, waiting penalty h and period index k. Due to symmetry between server i and j, we only explain the impacts of server i's parameters as an example; those of server j's parameters can be inferred accordingly. The results are summarized as follows.

Observation 3: *While fixing all other parameters,*

(a) As λ_i decreases, both thresholds $\overline{m}_i$ and $\overline{m}_j$ typically increase (see Figure 4.3);

(b) As μ_i typically increases, both thresholds $\overline{m}_i$ and $\overline{m}_j$ typically increase (see Figure 4.4);

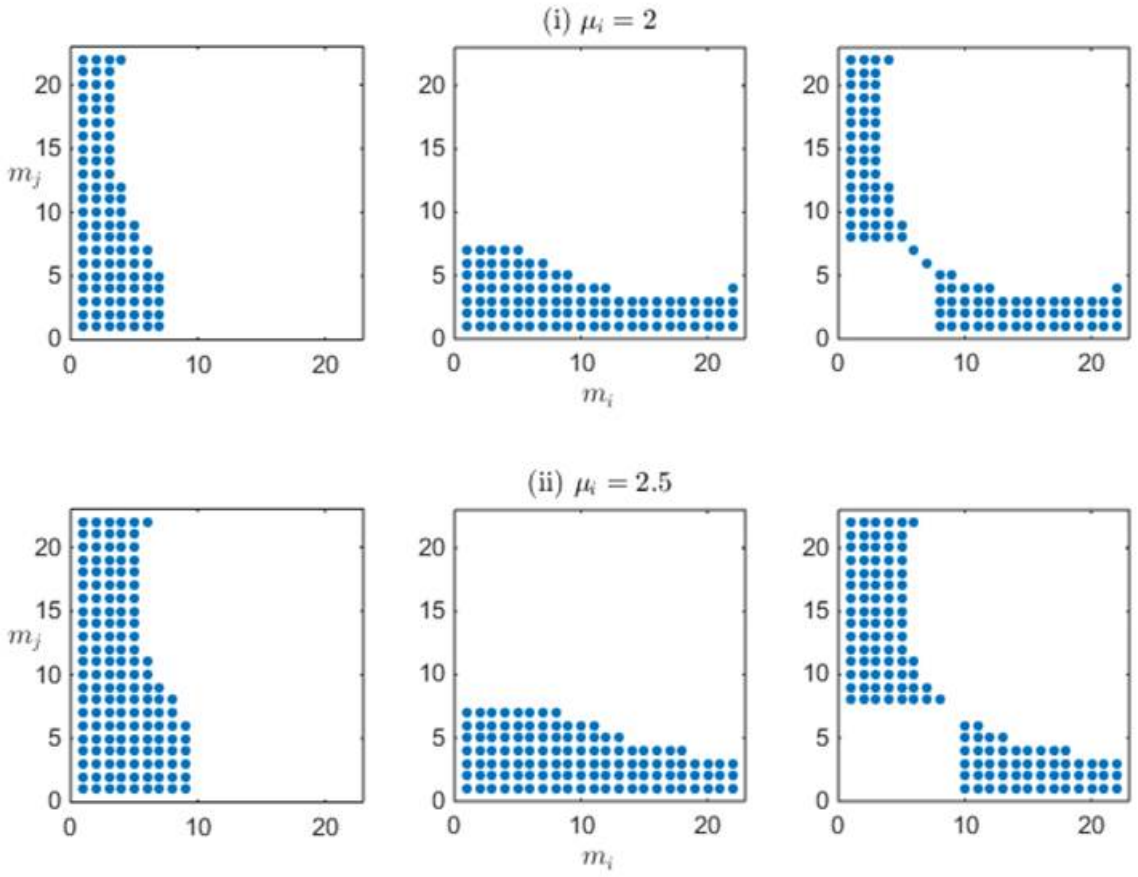

Figure 4.4: How μ_i affects the structure of server i's best response $a_i^k(m_i, m_j)$, server j's best response $a_j^k(m_i, m_j)$, and the collaboration equilibrium

(c) As both λ_i and μ_i increase proportionally, both thresholds $\overline{m}_i$ and $\overline{m}_j$ typically increase (see Figure 4.5);

(d) As R/h increases, both thresholds $\overline{m}_i$ and $\overline{m}_j$ increase (see Figure 4.6);

(e) As k increases, both thresholds $\overline{m}_i$ and $\overline{m}_j$ increase (see Figure 4.7).

We start with the arrival rate λ_i. In order to explicitly demonstrate the impact of λ_i, we present a special comparison in Figure 4.3 in which λ_i reduces from 1 to 0. Figure 4.3 shows that as λ_i decreases both $\overline{m}_i$ and $\overline{m}_j$ increase. It indicates that a smaller arrival rate λ_i dampens the two servers' incentive to transfer jobs and increases their incentive to accept jobs. This is intuitive: when there are fewer jobs coming to server i, server i is more likely to be idle and thus becomes more eager to accept jobs from server j and more conservative in transferring jobs; as a result of receiving fewer jobs from server i, server j also becomes more concerned about being idle, becoming

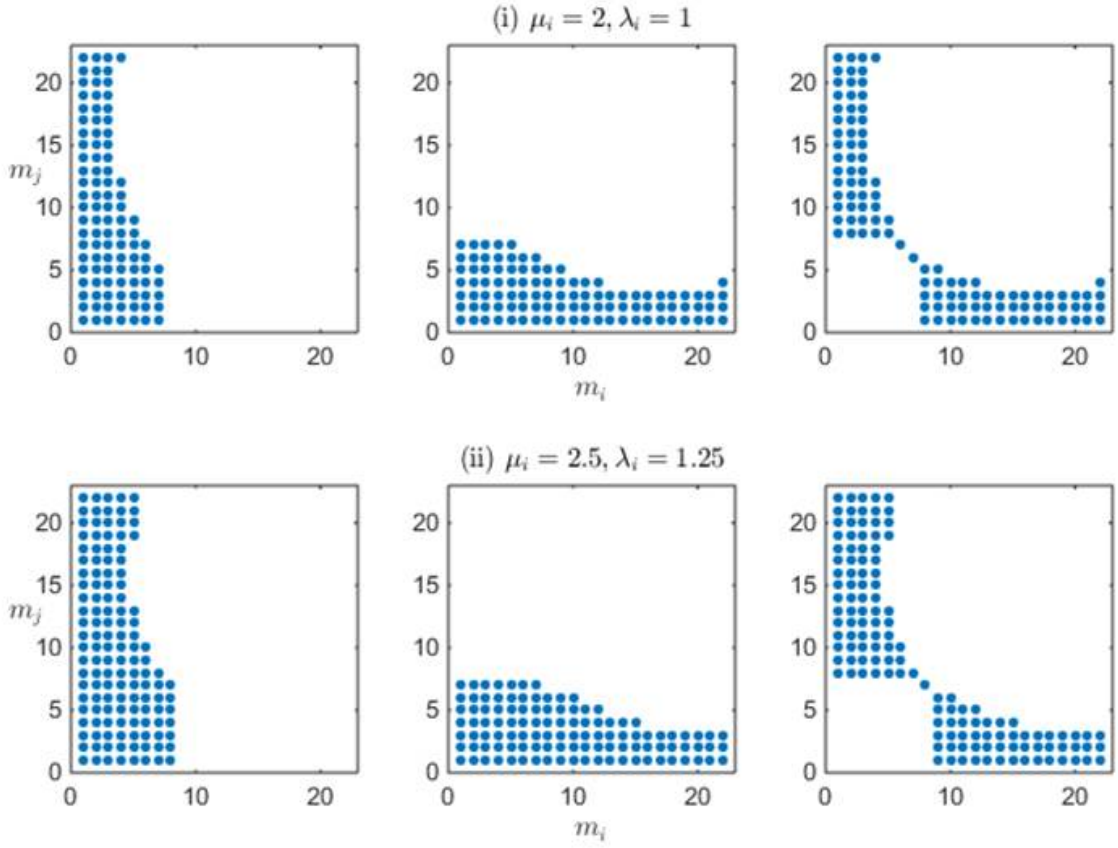

Figure 4.5: How λ_i and μ_i affect the structure of server i's best response $a_i^k(m_i, m_j)$, server j's best response $a_j^k(m_i, m_j)$, and collaboration equilibrium

less likely to transfer jobs. Furthermore, from Figure 4.3 we can observe that in case (ii) the hollow effect exists in server i's strategy but disappears in server j's strategy. Note that in case (ii) server i has no outside arrivals and only relies on server j to transfer jobs; in other words, server i is purely a job receiver whereas server j is purely a job transferrer. The observation supports our previous argument that the hollow effect is only driven by a job receiver's incentive to reject/accept jobs instead of a job transferrer's incentive to transfer/keep jobs.

Second we examine the impact of server i's service rate μ_i through Figure 4.4. Obviously the impacts of μ_i on the thresholds $\overline{m}_i$ and $\overline{m}_j$ are opposite to that of λ_i: the two servers are motivated to keep/accept more jobs under a higher service rate μ_i. This is simply because a larger service rate μ_i would lead server i to be more likely to be idle and thus incentivizes server i to keep/accept more jobs, as a result of which server j becomes more likely to be idle and thus more conservative in rejecting/transferring jobs. Furthermore, by comparing the changes in the strategies of service

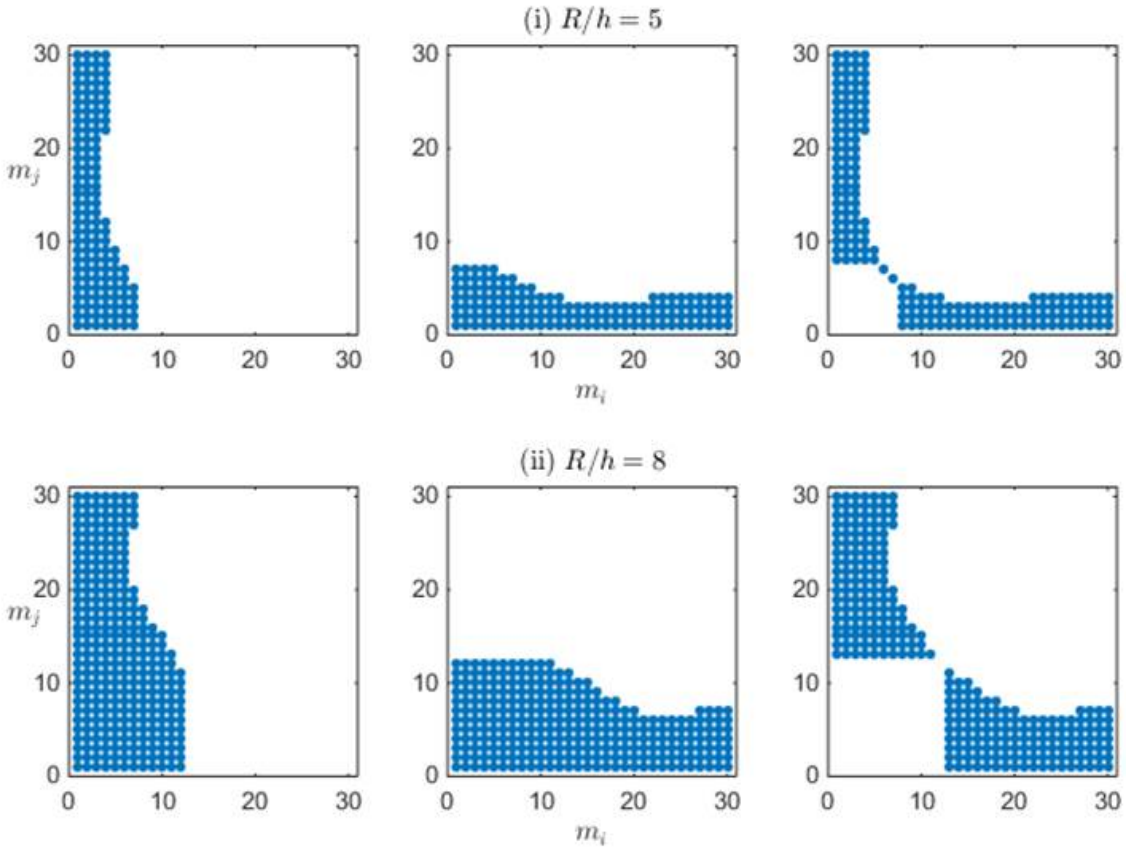

Figure 4.6: How the reward-to-penalty ratio R/h affects the structure of server i's best response $a_i^k(m_i, m_j)$, server j's best response $a_j^k(m_i, m_j)$, and collaboration equilibrium

i and j, we can observe that server i's threshold $\overline{m}_i$ is affected much more than $\overline{m}_j$ is. Obviously, when server i's capacity increases, it exerts its main effect on server i, and a more subtle external effect on server j through the two servers' interaction.

Thirdly we study the combined effect of arrival and service rates. Figure 4.5 suggests that as both λ_i and μ_i increase proportionally both $\overline{m}_i$ and $\overline{m}_j$ increase, indicating that the effect of an increased service rate dominates that of an increased arrival rate. The reason is that as suggested by classic queueing theory (e.g., an M/M/1 queue), when both λ_i and μ_i increase proportionally, the average waiting time for server i is reduced. In other words, server i is more capable in processing jobs, so server i is more likely to keep/accept jobs, which leads server j to be more conservative in transferring/rejecting jobs.

Lastly we briefly explain the impacts of the reward-to-penalty ratio R/h and the period index

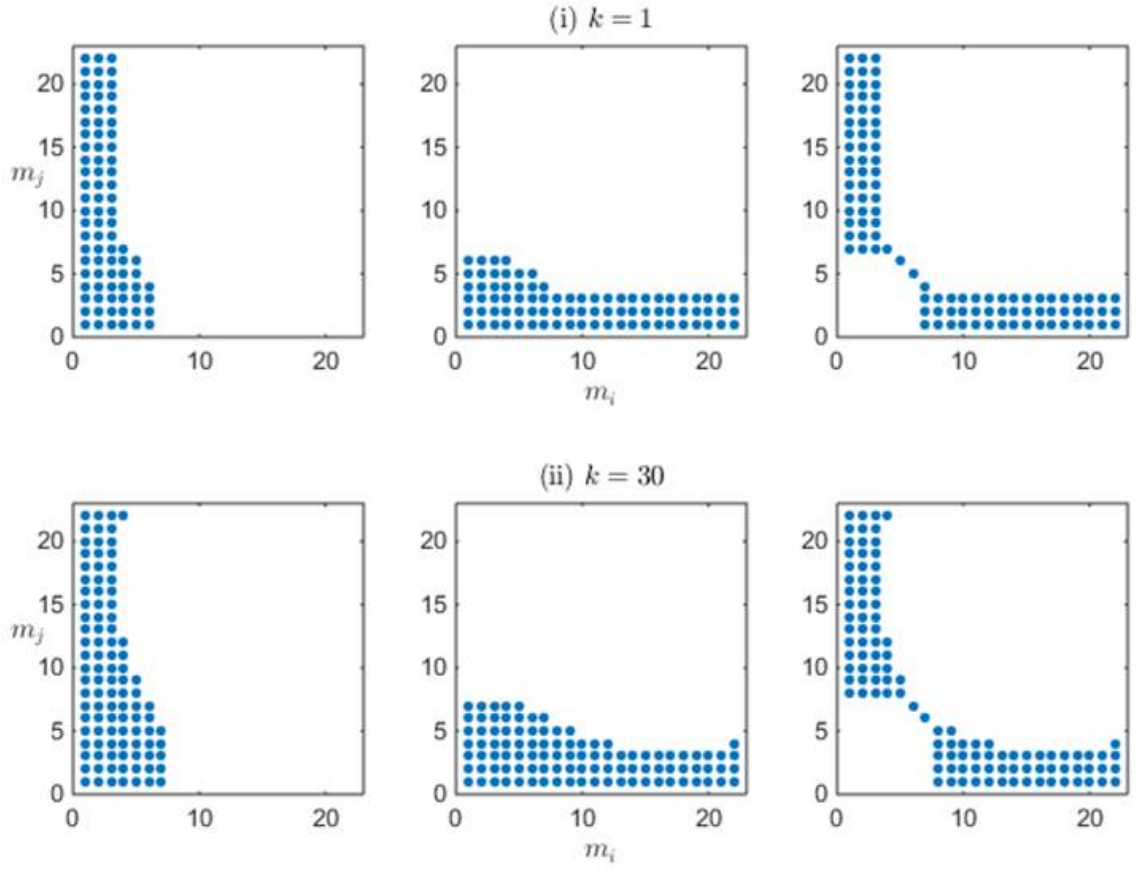

Figure 4.7: How the period index k affects the structure of server i's best response $a_i^k(m_i, m_j)$, server j's best response $a_j^k(m_i, m_j)$, and collaboration equilibrium

k. As expected, as R/h increases both $\overline{m}_i$ and $\overline{m}_j$ increase in Figure 4.6 because as jobs become more profitable, both servers become more reluctant to transfer jobs and more eager to accept jobs. Similar impacts could be observed in Figure 4.7 as the period index k increases: As it gets closer to the end, both servers value the instant payoff more than the diminished future effect.

It is worth noting that Figure 4.6 could also be used to demonstrate the impact of both λ_i and λ_j decreasing proportionally or both μ_i and μ_j increasing proportionally. The reasoning is straightforward: as both λ_i and λ_j decrease proportionally or both μ_i and μ_j increase proportionally, both servers become less congested, and thus become less likely to transfer jobs but more likely to accept jobs. As a result, both thresholds $\overline{m}_i$ and $\overline{m}_j$ would increase.

The numerical results indicate that the threshold structure is complex, being affected by a number of factors such as the other server's queue length, the arrival rates, the service rates,

the time period index, etc. This is because in a long term interaction, the two servers need to consider the impact of their decisions on their future payoff, and this future value includes several components such as the risk of being idle, the opportunity to transfer/receive jobs in the future, the effect of their own decisions on the other server's decisions, etc.

The complex structure indicated by the numerical study suggests that the analytical study on the equilibrium structure would be quite difficult. A possible solution might be to study the performance of a heuristic policy such as constant threshold: In the next section, we will examine the impact of collaboration through a constant threshold policy.

4.5 The Impact of Collaboration

In this section we use numerical results to further explore the collaboration equilibrium and its impact. Since we are interested in the design of mechanism of rewards and penalties, in the following we first analyze the sensitivity of the collaboration equilibrium with respect to the reward, under different service rates. Then we conduct comparative analysis to examine the impact of the strategic collaboration on servers' capacity investment, the average waiting time in the system, and a client's service outsourcing strategy.

In the simulation we consider two symmetric servers with the same capacity cost parameter (i.e., $c_i = c_j$). The two servers play a two-stage game backwards. During the service process the two servers play the collaboration game while all other parameters are fixed (including service rates (μ_i, μ_j), the total arrival rates λ, reward R and penalty rate h). Our theoretical analysis suggests that the two servers might collaborate following a threshold policy with a non-monotonic structure. However, since the state space of non-monotonic threshold functions is too large to evaluate, we adopt a constant threshold as an approximation. In the collaboration game, each server is allowed to choose a threshold from 1 to 10 (as we find that when the difference is larger than 10, in most cases servers have no transferring). Each server's threshold is chosen as a best response to the other server's choice.

At the capacity choice stage the two servers are allowed to choose from a total of eighteen different service rates, which correspond to 18 system utilizations evenly distributed between 10% and 95%. The service rate equilibrium is chosen when each server's choice of service rate forms

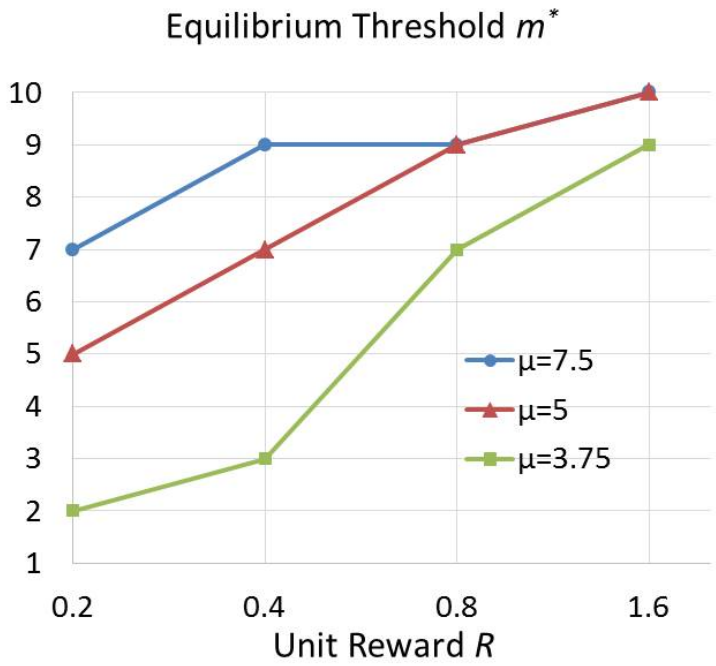

Figure 4.8: How the equilibrium threshold m^* changes with unit reward R and service rate μ

a best response to the other server's rate. Unless otherwise noted, our parameters are as follows: unit penalty rate $h = 3.2$, capacity cost rate $c = 0.034$, arrival rate $\lambda = 6$.

First, we examine how the collaboration equilibrium is affected by other parameters. The collaboration equilibrium is characterized by a constant threshold, denoted by $\bar{m}_o$ for server $o \in \{i, j\}$. In our experiments we allow μ_i to be different from μ_j but here we only show the results when $\mu_i = \mu_j = \mu$, which leads to the same threshold in equilibrium $\bar{m}_i = \bar{m}_j = m^*$. Figure 4.8 plots the value of m^* against two parameters: unit reward R and the service rate μ. In this figure, R is plotted on the x-axis and different values of service rate μ are represented by three lines with different markers. Figure 4.8 suggests two insight. First, m^* decreases as R decreases, indicating that servers have more collaboration under a lower reward. Obviously, servers are more likely to share their jobs when jobs are less valuable. Second, m^* decreases as μ decreases, suggesting that servers have more collaboration under lower service rates. This is because when servers are less capable of handling jobs, they more often face long queues, and thus are more likely to share jobs to alleviate this congestion.

Second, we study the impact of strategic collaboration on servers' capacity investment and the

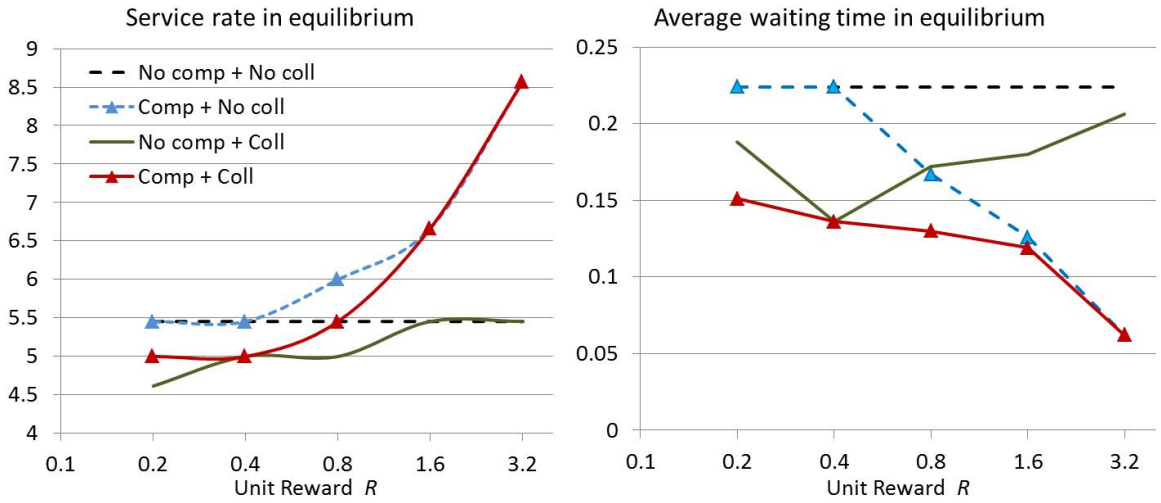

Figure 4.9: How the service rate μ and the average waiting time in equilibrium change with unit reward R in four cases: no competition nor collaboration, competition without collaboration, collaboration without competition, and competition with collaboration

average waiting of all incoming jobs. Since servers are symmetric, their service capacities turn out to be the same in equilibrium, and are represented by one single service rate. Figure 4.9 plots the equilibrium service rate on the left and the average waiting time on the right, both with the unit reward R on the x-axis. In Figure 4.9 the solid line with triangle markers represents the setting in our model, in which servers need to compete for jobs (i.e., the probability of an incoming job being allocated to server o is proportional to server o's service rates: $p_o = \frac{\mu_o}{\mu_i+\mu_j}$) and servers are allowed to collaborate according to their interests (i.e., $\bar{m}_i = \bar{m}_j = m^*$). In order to illustrate how the presence of strategic collaboration affects the two variables, Figure 4.9 also includes three benchmark cases: the dashed line without markers represents the case with no competition (i.e., $p_i = p_j = 1/2$, irrespective of service rates) nor collaboration (i.e., $\bar{m}_i$ and $\bar{m}_j$ are manually set to be infinite), the solid line without markers represents the case with collaboration but no competition, and the dashed line with triangle markers represents the case with competition but no collaboration.

Figure 4.9 suggests the following conclusions. By comparing the dashed and solid lines (without markers), we can see that the presence of strategic collaboration lowers servers' service rates and shortens the average waiting time in the system. This happens even in the presence of competition (which is shown by comparing the dashed and solid lines with triangle markers): Although the introduction of collaboration might dampen the competition effect by undermining servers' incen-

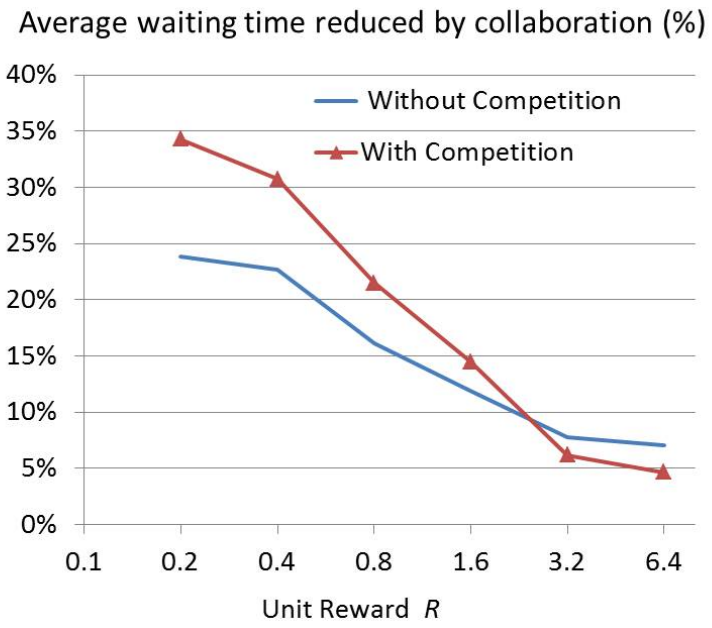

Figure 4.10: How the percentage of waiting time reduction led by strategic collaboration changes with unit reward R in two cases: without or with competition

tives to invest in capacities, it would indeed benefit the system in terms of leading to lower average waiting times. One possible explanation is that, as shown in Figure 4.9, competition often takes effect under a relative high reward, whereas collaboration takes affect mainly under a relatively low unit reward. As a result, when the reward is low, collaboration is highly active but competition is already weak, so the improved efficiency plays a major role and outweighs the undermined incentive. On the other hand, when the reward is high, competition is quite intense, but collaboration is almost non-existent and thus brings in little dampening impact. This suggests that collaboration might complement competition in terms of reducing system waiting time.

Figure 4.10 plots the (average)[7] percentage of the average waiting time reduced by the introduction of strategic collaboration under different rewards R. It considers two pairs of comparisons: the solid line without markers represents the waiting time reduction when there is no capacity competition (i.e. by comparing the dashed and solid lines without markers in Figure 4.9), and the solid line with triangle markers represents the waiting time reduction at the presence of capacity competition (i.e. by comparing the dashed and solid lines with triangle markers in Figure 4.9). The

[7]The statistics are computed by averaging results from different capacity cost parameters (i.e., c = 0.002, 0.003, 0.005, 0.007, 0.010, 0.015, 0.023, 0.034, 0.051, 0.077, 0.115, 0.173, 0.259, 0.389).

plot suggests two conclusions. First, both lines show that the waiting time reduction decreases as the reward R increases. This is consistent with our finding that collaboration is less active under a lower reward. Second, by comparing the two lines we can see that collaboration often reduces waiting time more when competition is present. This finding further validates our previous conjecture that collaboration complements competition. As a result, a client could often benefit from servers' strategic collaboration and thus should encourage it instead of prohibiting it; and the client should do so even when the client has implemented capacity competition between servers.

Chapter 5

Conclusion

This dissertation studies emerging operational problems in two areas: socially responsible management and service operations management. A main focus of the dissertation is to examine the role of information in operations management. The first chapter studies the impact of supply chain transparency on the use of child labor in global supply chain; I find that when more information regarding global firms' inspection efforts is disclosed to the public, this could backfire and lead to more child labor being used in the supply chain. The second chapter relates to service policy design for a single-server system; it shows that providing different service rates to homogeneous jobs might help reduce waiting in the system by creating more service time information. Finally, the last chapter studies strategic collaboration between two servers in service dual-outsourcing: when servers can share information regarding their queue lengths, they could transfer their allocated jobs in their own interests. In the last chapter, I have studied the structure of the two servers' collaboration strategies and numerically examined the impact of strategic collaboration on dual-outsourcing systems. In particular, I find that strategic collaboration could complement the commonly used capacity competition strategy in dual-outsourcing in reducing waiting and thus should be encouraged in dual-outsourcing.

Appendix A

Supplements for Chapter 2

A.1 Proofs of Analytical Results

Proof of Proposition 1. First, we simplify the program (2.12) by substituting $(\theta^{pre}(w), d^{pre}(w))$ given in (2.11) into (2.12), which gives

$$\max_{w} \begin{cases} U(w, \theta_L, 1) = v - w - eg & \text{if } s_H - \Delta(\theta_L) \leq w < s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L) \text{ and } I \geq \xi_1^{pre} \\ U(w, \theta_L, 0) = v - w & \text{if } w \geq s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L) \\ U(w, \theta_H, 1) = v - w - I - eg(1 - \theta_H) & \text{if } s_H - \Delta(\theta_H) \leq w < s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H) \\ & \text{and } I < \xi_1^{pre} \end{cases}$$

$$= \max_{w} \begin{cases} U(w, \theta_L, 1) = v - w - eg \text{ where } w = s_H - \Delta(\theta_L) & \text{if } I \geq \xi_1^{pre} \\ U(w, \theta_L, 0) = v - w \text{ where } w = s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L) & \\ U(w, \theta_H, 1) = v - w - eg(1 - \theta_H), \text{ where } w = s_H - \Delta(\theta_H) & \text{if } I < \xi_1^{pre}. \end{cases}$$

Next we solve for the equilibrium wholesale price w^{pre}. If $I \geq \xi_1^{pre}$, the manufacturer chooses $w = s_H + (1/(\gamma e) - 1)\Delta(\theta_L)$ when

$$U(s_H + (1/(\gamma e) - 1)\Delta(\theta_L), \theta_L, 0) \geq U(s_H - \Delta(\theta_L), \theta_L, 1) \Leftrightarrow g \geq \xi_2^{pre}.$$

If $I < \xi_1^{pre}$, the manufacturer chooses $w = s_H + (1/(\gamma e) - 1)\Delta(\theta_L)$ when

$$U(s_H + (1/(\gamma e) - 1)\Delta(\theta_L), \theta_L, 0) \geq U(s_H - \Delta(\theta_H), \theta_H, 1) \Leftrightarrow I \geq \xi_3^{pre}.$$

As a result, the outcome $(s_H + (1/(\gamma e) - 1)\Delta(\theta_L), \theta_L, 0)$ is in equilibrium if and only if $\{I \geq \xi_1^{pre}, g \geq \xi_2^{pre}\}$ or $\{I < \xi_1^{pre}, I \geq \xi_3^{pre}\}$, the outcome $(s_H - \Delta(\theta_H), \theta_H, 1)$ is in equilibrium if and only if $I < \xi_1^{pre}$ and $I < \xi_3^{pre}$, and the outcome $(s_H - \Delta(\theta_L), \theta_L, 1)$ is in equilibrium otherwise. ■ **Proof of Proposition 2.** We first simplify the program (2.17), and then solve for the subgame-perfect equilibrium through three lemmas. Substituting $d^{post}(\theta, w) = d^{pre}(\theta, w)$ given in (2.9) into (2.17) yields:

$$\max_{w,\theta} \begin{cases} U(w, \theta, 1) = v - w - I(\theta) - e(1-\theta)g \quad \text{if } s_H - \Delta(\theta) \leq w < s_H + \frac{\Delta(\theta)}{\gamma e(1-\theta)} - \Delta(\theta); \\ U(w, \theta_L, 1) = v - w - e(1-\theta_L)g \quad \text{if } s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H) \leq w < s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L); \\ U(w, \theta_H, 0) = v - w - I \quad \text{if } s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H) \leq w < s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L); \\ U(w, \theta, 0) = v - w - I(\theta) \quad \text{if } w \geq s_H + \frac{\Delta(\theta)}{\gamma e(1-\theta)} - \Delta(\theta), \end{cases}$$

which can be further simplified as follows:

$$\max_{w,\theta} \begin{cases} U(w, \theta, 1) = v - w - I(\theta) - e(1-\theta)g \quad \text{where } w = s_H - \Delta(\theta); \\ U(w, \theta_L, 1) = v - w - e(1-\theta_L)g \quad \text{where } w = s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H); \\ U(w, \theta_H, 0) = v - w - I \quad \text{where } w = s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H); \\ U(w, \theta, 0) = v - w - I(\theta) \quad \text{where } w = s_H + \frac{\Delta(\theta)}{\gamma e(1-\theta)} - \Delta(\theta). \end{cases}$$

In the above, the second $U(w, \theta_L, 1)$ is dominated by the first $U(w, \theta, 1)$ when $\theta = \theta_L$ because $w = s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H)$ is larger than $w = s_H - \Delta(\theta_L)$ by $s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H) > s_H > s_H - \Delta(\theta_L)$. Also, the third $U(w, \theta_H, 0)$ is a special case of the fourth $U(w, \theta, 0)$ when $\theta = \theta_H$. Therefore, we can simplify the above maximization program into the following:

$$\max_{\theta} \begin{cases} U(s_H - \Delta(\theta), \theta, 1) = v - (s_H - \Delta(\theta)) - I(\theta) - e(1-\theta)g; \\ U\left(s_H + \frac{\Delta(\theta)}{\gamma e(1-\theta)} - \Delta(\theta), \theta, 0\right) = v - \left(s_H + \frac{\Delta(\theta)}{\gamma e(1-\theta)} - \Delta(\theta)\right) - I(\theta). \end{cases} \tag{A.1}$$

From (A.1), we obtain $d^{post}=1$ if and only if:

$$U\left(s_H-\Delta\left(\theta_L\right),\theta_L,1\right)\geq U\left(s_H+\tfrac{\Delta(\theta_L)}{\gamma e(1-\theta_L)}-\Delta\left(\theta_L\right),\theta_L,0\right)\Leftrightarrow g\leq\xi_2^{post}\text{ when }\theta=\theta_L; \tag{A.2}$$

$$U\left(s_H-\Delta\left(\theta_H\right),\theta_H,1\right)\geq U\left(s_H+\tfrac{\Delta(\theta_H)}{\gamma e(1-\theta_H)}-\Delta\left(\theta_H\right),\theta_H,0\right)\Leftrightarrow g\leq\xi_4^{post}\text{ when }\theta=\theta_H. \tag{A.3}$$

In the rest of the proof, we provide the remaining conditions under three possible structures of the equilibrium outcome that depend on the value of m. Lemmas 1, 2 and 3 together prove Proposition 2. Figures 2.3(a), 2.3(b) and 2.3(c) illustrate the equilibrium presented in Lemmas 1, 2 and 3, respectively. We present the proof of Lemma 1 here, while providing the proofs of Lemma 7 and 8 in the online supplement. For convenience, we rewrite assumptions (A2) and (A3), respectively, as $m\leq\overline{m}\equiv\dfrac{\left(1-\theta_H\right)\left(s_H-s_L\right)}{\theta_H}$ and $m\geq\underline{m}\equiv\dfrac{\gamma e\left(1-\theta_H\right)\left(s_H-s_L\right)}{1-\gamma e\left(1-\theta_H\right)}$.

Lemma 6 *When* $\gamma e<\dfrac{1}{2-\theta_H}$ *and* $\underline{m}\leq m<m_1\equiv\left(1-\theta_H\right)\left(s_H-s_L\right)$*, the equilibrium outcome of program (A.1) is*

$$\left(w^{post},\theta^{post},d^{post}\right)=\begin{cases}\left(s_H+\left(\frac{1}{\gamma e}-1\right)\Delta\left(\theta_L\right),\theta_L,0\right) & \text{if } g\geq\xi_2^{post},I\geq\xi_3^{post},I\geq\xi_5^{post};\\ \left(s_H-\Delta\left(\theta_L\right),\theta_L,1\right) & \text{if } I\geq\xi_1^{post},g\leq\xi_2^{post};\\ \left(s_H-\Delta\left(\theta_H\right),\theta_H,1\right) & \text{if } I\leq\xi_1^{post},I\leq\xi_3^{post},g\leq\xi_4^{post};\\ \left(s_H+\left\{\frac{1}{\gamma e(1-\theta_H)}-1\right\}\Delta\left(\theta_H\right),\theta_H,0\right) & \text{if } g\geq\xi_4^{post},I\leq\xi_5^{post}.\end{cases} \tag{A.4}$$

Proof: The first condition $\gamma e<\dfrac{1}{2-\theta_H}$ ensures that there exists $m\in[\underline{m},m_1)$, and the second condition $m<m_1$ suggests that $\xi_4^{post}>\xi_2^{post}$. Given that $\xi_4^{post}>\xi_2^{post}$, program (A.1) can be simplified as follows:

$$\begin{aligned}&\max\left\{U\left(s_H-\Delta\left(\theta_L\right),\theta_L,1\right),U\left(s_H-\Delta\left(\theta_H\right),\theta_H,1\right)\right\}\quad\text{if } g\leq\xi_2^{post};\\ &\max\left\{U\left(s_H-\Delta\left(\theta_H\right),\theta_H,1\right),U\left(s_H+\tfrac{\Delta(\theta_L)}{\gamma e}-\Delta\left(\theta_L\right),\theta_L,0\right)\right\}\quad\text{if } \xi_2^{post}<g\leq\xi_4^{post};\\ &\max\left\{U\left(s_H+\tfrac{\Delta(\theta_L)}{\gamma e}-\Delta\left(\theta_L\right),\theta_L,0\right),U\left(s_H+\tfrac{\Delta(\theta_H)}{\gamma e(1-\theta_H)}-\Delta\left(\theta_H\right),\theta_H,0\right)\right\}\quad\text{if } g>\xi_4^{post}.\end{aligned}$$

Therefore, we obtain the following results: (i) $\left(s_H-\Delta\left(\theta_L\right),\theta_L,1\right)$ is the equilibrium outcome if and only if $g\leq\xi_2^{post}$ and

$$U\left(s_H-\Delta\left(\theta_L\right),\theta_L,1\right)\geq U\left(s_H-\Delta\left(\theta_H\right),\theta_H,1\right)\Leftrightarrow I\geq\xi_1^{post}. \tag{A.5}$$

(ii) $(s_H - \Delta(\theta_H), \theta_H, 1)$ is the equilibrium outcome if and only if $\left\{g \leq \xi_2^{post}, I \leq \xi_1^{post}\right\}$, or $\left\{\xi_2^{post} < g \leq \xi_4^{post},\ I \leq \xi_3^{post}\right\}$, where

$$U(s_H - \Delta(\theta_H), \theta_H, 1) \geq U\left(s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L), \theta_L, 0\right) \Leftrightarrow I \leq \xi_3^{post}. \tag{A.6}$$

We can show that these conditions can be further simplified to $g \leq \xi_4^{post}, I \leq \xi_3^{post}$, and $I \leq \xi_1^{post}$. (iii) $\left(s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L), \theta_L, 0\right)$ is the equilibrium outcome if and only if $\left\{\xi_2^{post} < g \leq \xi_4^{post},\ I \geq \xi_3^{post}\right\}$ or $\left\{g > \xi_2^{post},\ I \geq \xi_5^{post}\right\}$, where

$$U\left(s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L), \theta_L, 0\right) \geq U\left(s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H), \theta_H, 0\right) \Leftrightarrow I \geq \xi_5^{post}. \tag{A.7}$$

Similar to (ii), these conditions can be simplified to $g > \xi_2^{post}$, $I \geq \xi_3^{post}$ and $I \geq \xi_5^{post}$. (iv) $\left(s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H), \theta_H, 0\right)$ is the equilibrium outcome if and only if $g > \xi_4^{post}$ and $I \leq \xi_5^{post}$. ■

Lemma 7 *When* $\gamma e < \dfrac{1}{2-\theta_H}$ *and* $m_1 \leq m \leq m_2 \equiv (s_H - s_L)\left(\frac{1}{\gamma e(1-\theta_H)^2+\theta_H} - 1\right)$, *the equilibrium outcome of program (A.1) is*

$$\left(w^{post}, \theta^{post}, d^{post}\right) = \begin{cases} \left(s_H + \left(\frac{1}{\gamma e} - 1\right)\Delta(\theta_L), \theta_L, 0\right) & \text{if } g \geq \xi_2^{post}, I \geq \xi_5^{post}; \\ (s_H - \Delta(\theta_L), \theta_L, 1) & \text{if } I \geq \xi_1^{post}, g \leq \xi_2^{post}, I \geq \xi_6^{post}; \\ (s_H - \Delta(\theta_H), \theta_H, 1) & \text{if } I \leq \xi_1^{post}, g \leq \xi_4^{post}; \\ \left(s_H + \left\{\frac{1}{\gamma e(1-\theta_H)} - 1\right\}\Delta(\theta_H), \theta_H, 0\right) & \text{if } I \leq \xi_6^{post}, g \geq \xi_4^{post}, I \leq \xi_5^{post}. \end{cases} \tag{A.8}$$

Lemma 8 *When* $\left\{m_2 < m < \overline{m}, \gamma e < \dfrac{1}{2-\theta_H}\right\}$ *or* $\gamma e > \dfrac{1}{2-\theta_H}$, *the equilibrium outcome of program (A.1) is*

$$\left(w^{post}, \theta^{post}, d^{post}\right) = \begin{cases} \left(s_H + \left(\frac{1}{\gamma e} - 1\right)\Delta(\theta_L), \theta_L, 0\right) & \text{if } g \geq \xi_2^{post}, I \geq \xi_5^{post}; \\ (s_H - \Delta(\theta_L), \theta_L, 1) & \text{if } g \leq \xi_2^{post}, I \geq \xi_6^{post}; \\ \left(s_H + \left\{\frac{1}{\gamma e(1-\theta_H)} - 1\right\}\Delta(\theta_H), \theta_H, 0\right) & \text{if } I \leq \xi_6^{post}, I \leq \xi_5^{post}. \end{cases} \tag{A.9}$$

Proof of Proposition 3. (a) The supplier's profit is 0 under the inspection-alone strategy or the do-nothing strategy, $(1/(\gamma e)-1)\Delta(\theta_L)$ under the premium-alone strategy, and $\{1/(\gamma e(1-\theta_H))-1\}\Delta(\theta_H)$ under the premium & inspection strategy, where $(1/(\gamma e)-1)\Delta(\theta_L) > \{1/(\gamma e(1-\theta_H))-1\}\Delta(\theta_H)$. From Table 1, we can verify the change of the supplier's profit in each area. (b) The proof is provided in the main body. (c) This can be verified from the change of the manufacturer's strategy in each area shown in Table 2.1. ■

A.2 Examples of Companies Committing to No or Low-Level Inspection Efforts

- Caterpillar Inc.: "We do not currently verify our product supply chain or audit suppliers specifically to evaluate risks of human trafficking and slavery or require our direct suppliers to certify that materials incorporated into products comply with laws regarding slavery and human trafficking in the countries in which they are doing business." (Source: http://www.jointeamcaterpillar.com/cda/layout?m=470716&x=69, Accessed on January 31, 2015)

- Danaher Corp.: "Accordingly, Danaher Corporation does not verify its supply chains to evaluate the risks of human trafficking or slavery, audit its suppliers for compliance with or require suppliers to certify compliance with the laws regarding human trafficking and slavery" (Source: http://www.danaher.com/sites/default/files/California_Transparency_in_Supply_Chains_ Act_of_2010.pdf, Accessed on January 31, 2015)

- IDEX Corporation: "Although IDEX Units expect their suppliers to comply with applicable laws and frequently obtain agreements and certifications from their key suppliers relating to compliance with applicable laws in supplying products to them, IDEX Units do not verify their supply chain or audit suppliers specifically to evaluate risks of human trafficking and slavery or require their suppliers to certify specifically that products supplied to them were not produced with child labor or slave labor." (Source: http://idexcorp.com/terms/SupplyChainTransparency.asp, Accessed on January 31, 2015)

- Hyundai Motor America: "Hyundai has no policy regarding, and does not monitor, human

trafficking and slavery in its direct product supply chain." (Source: https://www.hyundaiusa.com/terms-conditions.aspx, Accessed on January 31, 2015)

- Krispy Kreme Doughnuts, Inc.: "We do not engage in verification of product supply chains to evaluate and address risks of human trafficking and slavery, nor conduct audits of suppliers to evaluate supplier compliance with company standards against trafficking and slavery in supply chains." (Source: http://www.krispykreme.com/SharedContent/Media/FormsLib/supply_chains_act.pdf, Accessed on January 31, 2015)

- Orora North America: "Orora has not (a) reviewed its product supply chains to evaluate and address risks of human trafficking and slavery or employed a third party to audit or evaluate such risks, (b) established entity standards on human trafficking and slavery and then conducted supplier audits to evaluate whether suppliers comply with standards." (Source: http://www.mppmfg.com/docs/CA-Transparency-Supply-Chain-Act.pdf, Accessed on January 31, 2015)

- Overhill Farms, Inc.: "Overhill Farms does not (1) engage in verification of product supply chains to determine and address risks of human trafficking and slavery; (2) conduct audits of suppliers to determine supplier compliance with company standards for human trafficking and slavery in supply chains;" (Source: http://www.overhillfarms.com/pdf/2012/Supply%20Chains%20Act%20Disclosure%20_OFI_%20Final1.pdf#zoom=80, Accessed on January 31, 2015)

- Valero Energy Corp.: "while Valero has not undertaken action with the intent of specifically addressing California Civil Code Sections 1714.43(c)(1) – (c)(5), Valero recognizes and respects all labor and employment laws, including those addressing slavery and human trafficking, wherever Valero operates." (Source: http://www.valero.com/businesspartners/valerosuppliers/pages/californiatransparencyinsupplychainsactdisclosure.aspx, Accessed on January 31, 2015)

A.3 Additional Results

1 Supplier's Employment Decision on a Mix of Child and Adult Laborers

In the base model, we assume that the child labor decision is binary: the supplier employs only child labor ($d = 1$) or only adult labor ($d = 0$). In practice, the supplier may employ a mix of both. In this case, the supplier decides a portion of child labor employed, d ($\in [0, 1]$). We present main results of this case in this section and provide a sketch of proof in Section D. Subscript "*cont*" is used to denote equilibrium in this section. The following corollary presents the equilibrium outcome in the pre-Act scenario.

Corollary 4 *When the supplier chooses a portion of child labor d ($\in [0, 1]$), the subgame-perfect equilibrium in the pre-Act scenario is:*

$$\{w_{cont}^{pre}, \theta_{cont}^{pre}, d_{cont}^{pre}\} =$$
$$\begin{cases} (s_H - \Delta(\theta_L), \theta_L, 1) & \text{if } I > \xi_{1,cont}^{pre} \text{ and } g \leq \xi_{2,cont}^{pre}; \\ (s_H - \Delta(\theta_H), \theta_H, 1) & \text{if } I \leq \xi_{1,cont}^{pre}, I \leq \xi_{3,cont}^{pre} \text{ and } I > \xi_{4,cont}^{pre}; \\ (s_H + \Delta(\theta_L)/\gamma e,\ \theta_L, 0) & \text{if } \{I > \xi_{1,cont}^{pre} \text{ and } g > \xi_{2,cont}^{pre}\} \\ & \text{or } \{I \leq \xi_{1,cont}^{pre}, I > \xi_{3,cont}^{pre} \text{ and } I > \xi_{5,cont}^{pre}\}; \\ \left(s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - 2d_I\Delta(\theta_H), \theta_H, d_I\right) & \text{otherwise;} \end{cases}$$

where $d_I = I/(e\theta_H g)$, $\xi_{1,cont}^{pre} = \xi_1^{pre}$, $\xi_{2,cont}^{pre} = \xi_2^{pre} + \Delta(\theta_L)/e$, $\xi_{3,cont}^{pre} = \xi_3^{pre} + \Delta(\theta_L)$, $\xi_{4,cont}^{pre} =$
$$\begin{cases} e\theta_H g\left(1 - \frac{\beta(\theta_H) - \Delta(\theta_H)}{e(1-\theta_H)g - 2\Delta(\theta_H)}\right), \text{ for } g \in \left(\frac{2\Delta(\theta_H)}{e(1-\theta_H)}, +\infty\right) \\ 0, \text{ for } g \in \left[0, \frac{2\Delta(\theta_H)}{e(1-\theta_H)}\right) \end{cases}, \text{ and}$$
$\xi_{5,cont}^{pre} = \dfrac{egm\theta_H^2}{\gamma e(1-\theta_H)[eg - 2\Delta(\theta_H)]}$.

By comparing $\{w_{cont}^{pre}, \theta_{cont}^{pre}, d_{cont}^{pre}\}$ presented in Corollary 4 with $\{w^{pre}, \theta^{pre}, d^{pre}\}$ presented in Proposition 1, we observe that the first two equilibrium outcomes correspond to the strategies of do-nothing and inspection-alone in the base model, respectively. The third equilibrium outcome is parallel to the premium-alone strategy, but with a higher wholesale price of $s_H + \Delta(\theta_L)/\gamma e$ than $s_H + (1/(\gamma e) - 1)\Delta(\theta_L)$ in Proposition 1. The reason is that it is more costly for the manufacturer to incentivize the supplier not to hire child labor when the supplier has more flexibility in choosing a level of child labor employment. Lastly, the fourth equilibrium outcome in

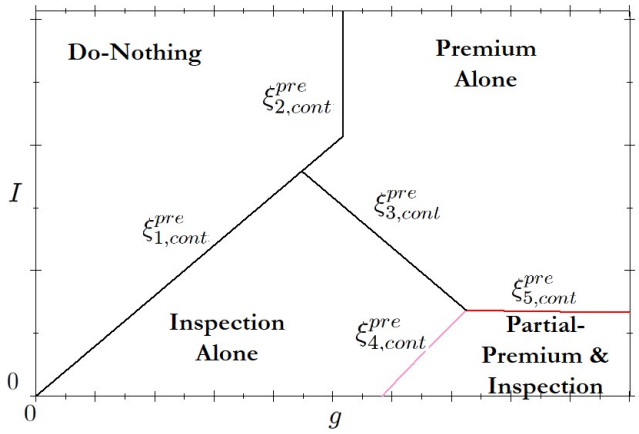

Figure A.31: Equilibrium Outcomes in the Pre-Act Scenario under a Continuous Child Labor Decision.

Corollary 4 indicates a new strategy, "partial-premium & inspection". Under this strategy, the manufacturer conducts internal inspections (i.e., $\theta^{pre}_{cont} = \theta_H$) and pays a partial premium (i.e., $w^{pre}_{cont} = s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - 2d_I\Delta(\theta_H)$) to induce the supplier to hire only a portion of workforce with child labor (i.e., $d^{pre}_{cont} = d_I \in (0,1)$). The expected amount of child labor used for production under the new strategy is $d_E = d_I(1-\theta_H) = I(1-\theta_H)/(ge\theta_H)$, which increases in I and decreases in g and e. Figure A.31 illustrates four strategies in equilibrium. It shows a structure similar to Figure 2.2 except the following two differences. First, the threshold lines $\xi^{pre}_{2,cont}$ and $\xi^{pre}_{3,cont}$ in Figure A.31 are higher than ξ^{pre}_2 and ξ^{pre}_3 in Figure 2.2, respectively, due to the higher wholesale price associated with the premium-alone strategy. Second, Figure A.31 shows a new area at the right bottom defined by thresholds $\xi^{pre}_{4,cont}$and $\xi^{pre}_{5,cont}$, representing the parameter space where the partial-premium & inspection strategy is adopted. The impact of goodwill cost g and inspection cost I on the manufacturer's strategy and the amount of child labor remains mostly unaffected. From Figure A.31, we can observe that as g increases, the expected amount of child labor d_E decreases – this is also true when the manufacturer's strategy changes from inspection-alone to partial-premium & inspection. When a lower inspection cost I induces the manufacturer to change her strategy from do-nothing to inspection-alone or from inspection-alone to partial-premium & inspection, the expected amount of child labor d_E decreases. On the other hand, a lower inspection cost I increases d_E when it induces the manufacturer to switch from premium-alone to inspection-alone or partial premium & inspection. Next, the following corollary presents the equilibrium outcome in

the post-Act scenario.

Corollary 5 *When the supplier chooses a portion of child labor* d $(\in [0,1])$*, the subgame-perfect equilibrium in the post-Act scenario is:*

$$\left\{w_{cont}^{post}, \theta_{cont}^{post}, d_{cont}^{post}\right\} =$$

$$\begin{cases} \left(s_H - \Delta\left(\theta_L\right), \theta_L, 1\right) & \text{if } I \geq \xi_{1,cont}^{post},\ g \leq \xi_{2,cont}^{post},\ I \geq \xi_{6,cont}^{post}; \\ \left(s_H - \Delta\left(\theta_H\right), \theta_H, 1\right) & \text{if } I \leq \xi_{1,cont}^{post},\ I \leq \xi_{3,cont}^{post},\ g \leq \xi_{4,cont}^{post}; \\ \left(s_H + \Delta\left(\theta_L\right)/\gamma e, \theta_L, 0\right) & \text{if } g \geq \xi_{2,cont}^{post},\ I \geq \xi_{3,cont}^{post},\ I \geq \xi_{5,cont}^{post}; \\ \left(s_H + \Delta\left(\theta_H\right)/\left(\gamma e\left(1-\theta_H\right)\right), \theta_H, 0\right) & \text{if } I \leq \xi_{6,cont}^{post},\ g \geq \xi_{4,cont}^{post},\ I \leq \xi_{5,cont}^{post}. \end{cases}$$

where $\xi_{1,cont}^{post} = \xi_1^{post}$, $\xi_{2,cont}^{post} = \xi_2^{post} + \Delta\left(\theta_L\right)/e$, $\xi_{3,cont}^{post} = \xi_3^{post} + \Delta\left(\theta_L\right)$,
$\xi_{4,cont}^{post} = \xi_4^{post} + \Delta\left(\theta_H\right)/\left(e\left(1-\theta_H\right)\right)$, $\xi_{5,cont}^{post} = \xi_5^{post} - \Delta\left(\theta_H\right) + \Delta\left(\theta_L\right)$, *and* $\xi_{6,cont}^{post} = \xi_6^{post} - \Delta\left(\theta_H\right)$.

Corollary 5 reveals that the manufacturer chooses one of the four equilibrium strategies as in the base model (cf. Proposition 2). Specifically, the first two strategies in Corollary 5 correspond to the strategies of do-nothing and inspection-alone, respectively, and the latter two strategies are parallel to the strategies of premium-alone and premium & inspection, respectively, with higher price premiums due to the same reason as in the pre-Act scenario. Therefore, one can infer that the structure of the equilibrium is also the same as that in the base model shown in Figure 2.3. The only difference is that the values of thresholds are changed as follows: $\xi_{2,cont}^{post} > \xi_2^{post}$, $\xi_{3,cont}^{post} > \xi_3^{post}$, $\xi_{4,cont}^{post} > \xi_4^{post}$, $\xi_{5,cont}^{post} > \xi_5^{post}$, and $\xi_{6,cont}^{post} < \xi_6^{post}$. Finally, we examine the effects of the Act by comparing the equilibrium outcomes in the pre-Act scenario (in Corollary 4) with those in the post-Act scenario (Corollary 5). Figure A.32 illustrates the equilibrium outcomes[1], and shows that there are five areas in which the equilibrium outcomes are different between the two scenarios: $R\left(a\right)$, $R\left(b\right)$, $R\left(c\right)$, $R\left(d\right)$ and $R\left(e\right)$. Table A.31 provides precise conditions for each of these five areas. We can observe that the four areas $R\left(a\right) - R\left(d\right)$ in Table A.31 are the same as their respective area in Table 2.1 in the base model. The only difference lies in the area $R\left(e\right)$, in which the manufacturer's strategy changes from partial-premium & inspection to premium & inspection, but we can easily show that its impacts on the manufacturer's profit, the supplier's profit, and the amount of child

[1]Similar to Figure 2.4, there are three possible comparison plots with different structures, but here we only show a representative one.

Table A.31: The Manufacturer's Equilibrium Strategy: the Pre-Act Scenario vs. the Post-Act Scenario under a Continuous Child Labor Decision

Area	Equilibrium Outcome (Pre-Act $\rightarrow$ Post-Act)	Conditions
$R(a)$	Inspection Alone $\rightarrow$ Do-Nothing	$\xi_{1,cont}^{post} \leq I \leq \xi_{1,cont}^{pre}$ and $I \leq \xi_{3,cont}^{pre}$
$R(b)$	Premium Alone $\rightarrow$ Do-Nothing	$\xi_{3,cont}^{pre} \leq I \leq \xi_{1,cont}^{pre}$ and $g \leq \xi_{2,cont}^{pre}$
$R(c)$	Inspection Alone $\rightarrow$ Premium & Inspection	$g \geq \xi_{4,cont}^{post}$ and $\xi_{4,cont}^{pre} \leq I \leq \xi_{3,cont}^{pre}$
$R(d)$	Premium Alone $\rightarrow$ Premium & Inspection	$\xi_{5,cont}^{pre} \leq I \leq \xi_{5,cont}^{post}$ and $I \geq \xi_{3,cont}^{pre}$
$R(e)$	Partial-Premium & Inspection $\rightarrow$ Premium & Inspection	$g \geq \xi_{4,cont}^{post}$, $I < \xi_{4,cont}^{pre}$ and $I < \xi_{5,cont}^{pre}$

labor are similar to those in $R(c)$. Thus, the Act may still backfire and introduce more child labor in supply chains. strategy.

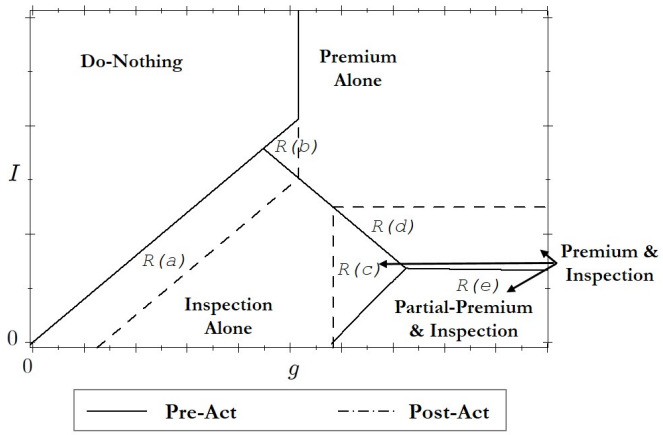

Figure A.32: Comparison of Equilibrium Outcomes between the Pre-Act Scenario and the Post-Act Scenario under a Continuous Child Labor Decision.

2 Mixed Strategy

In the base model we restrict our attentions on the case that the manufacturer and the supplier only choose a pure strategy on inspection level and child labor employment level. In this section, we analyze the case where the two players may choose a mixed strategy on the two decisions. The manufacturer's decision on inspection is characterized by an "inspection" probability $P_\theta \equiv Pr[\theta = \theta_H]$, and the supplier's decision on child labor employment is characterized by a probability of

employing child labor $P_d \equiv Pr[d=1]$. We present main results of this case in this section and provide a sketch of proof in Section D. Subscript "*mixed*" is used to denote equilibrium in this section. The following corollary presents the equilibrium outcome in the pre-Act scenario.

Corollary 6 *When the manufacturer and the supplier adopt mixed strategy on inspection and child labor employment, the subgame-perfect equilibrium in the pre-Act scenario is:*

$$\left\{w^{pre}_{mixed}, P^{pre}_{\theta,mixed}, P^{pre}_{d,mixed}\right\} =$$
$$\begin{cases} (s_H - \Delta(\theta_L), 0, 1) & \text{if } I > \xi^{pre}_{1,mixed} \text{ and } g \le \xi^{pre}_{2,mixed} \\ (s_H - \Delta(\theta_H), 1, 1) & \text{if } I \le \xi^{pre}_{1,mixed}, I \le \xi^{pre}_{3,mixed} \text{ and } I > \xi^{pre}_{4,mixed} \\ \left(s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L), 0, 0\right) & \text{if } I > \xi^{pre}_{1,mixed} \text{ and } g > \xi^{pre}_{2,mixed} \\ & \text{or if } I \le \xi^{pre}_{1,mixed}, I > \xi^{pre}_{3,mixed} \text{ and } I > \xi^{pre}_{5,mixed} \\ \left(s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H), 1, d_I\right) & o.w.; \end{cases}$$

where $d_I = I/(e\theta_H g)$, $\xi^{pre}_{1,mixed} = \xi^{pre}_1$, $\xi^{pre}_{2,mixed} = \xi^{pre}_2$, $\xi^{pre}_{3,mixed} = \xi^{pre}_3$,
$\xi^{pre}_{4,mixed} = eg\theta_H(1-\theta_H) - \theta_H\Delta(\theta_H)/(\gamma e(1-\theta_H))$, *and*
$\xi^{pre}_{5,mixed} = \theta_H\{(1/(\gamma e(1-\theta_L)) - 1)\Delta(\theta_L) - (1/(\gamma e(1-\theta_H)) - 1)\Delta(\theta_H)\}$.

By comparing $\left\{w^{pre}_{mixed}, P^{pre}_{\theta,mixed}, P^{pre}_{d,mixed}\right\}$ presented in Corollary 6 with $\{w^{pre}_{cont}, \theta^{pre}_{cont}, d^{pre}_{cont}\}$ presented in Corollary 4 in §C.1, we observe that the manufacturer chooses one of the four equilibrium strategies as in the case when the supplier may hire a portion of workforce with child labor (i.e., $d \in [0,1]$). Specifically, the first two strategies in $\left\{w^{pre}_{mixed}, P^{pre}_{\theta,mixed}, P^{pre}_{d,mixed}\right\}$ and $\{w^{pre}_{cont}, \theta^{pre}_{cont}, d^{pre}_{cont}\}$ are the same, corresponding to the strategies of do-nothing and inspection-alone in the base model. The third strategy is parallel to the premium-alone strategy, but with a wholesale price of $s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L)$, which is lower than $s_H + \frac{\Delta(\theta_L)}{\gamma e}$ in Corollary 4. The fourth strategy is parallel to the partial-premium & inspection strategy, but with a wholesale price of $s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H)$, which is independent of d_I, whereas in Corollary 4 the wholesale price $s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - 2d_I\Delta(\theta_H)$ decreases in d_I. The reason is as follows: when the supplier adopts a mixed strategy, the wholesale price always ensures that the supplier's profits from hiring child labor and not hiring child labor are the same (i.e., $\Pi(w,\theta,d=1) = \Pi(w,\theta,d=0)$); In contrast, when the supplier adopts a continuous portion of child labor, under a higher level of child labor employment, the supplier earns a lower expected profit margin from hiring child labor (i.e., $\frac{\partial\Pi(w,\theta,d)}{\partial d}$

decreases in d), and hence the manufacturer could pay a lower premium to incentivize the supplier not to hire child labor. The above comparison indicates that the structure of the subgame equilibrium presented in Corollary 6 is similar to that illustrated in Figure A.31. This suggests that when the the manufacturer and the supplier adopt a mixed strategy on inspection and child labor employment, the impact of goodwill cost g and inspection cost I on the manufacturer's strategy and the amount of child labor are mostly the same as in the case when the the supplier chooses a portion of child labor d ($\in [0,1]$). Next, the following corollary presents the equilibrium outcome in the post-Act scenario.

Corollary 7 *When the manufacturer and the supplier adopt mixed strategy on inspection and child labor employment, the subgame-perfect equilibrium in the post-Act scenario is the same as that in the base model given in (2.18).*

In the post-Act scenario, since the manufacturer discloses her inspection decision to the supplier in advance, only pure strategy arises and the subgame equilibrium outcome is the same as in the base model. Therefore, one can infer that the structure of the equilibrium is also the same as that in the base model shown in Figure 2.3. Finally, we examine the effects of the Act by comparing the equilibrium outcomes in the pre-Act scenario (in Corollary 6) with those in the post-Act scenario (Corollary 7). As suggested by the analysis of Corollary 6 and Corollary 7, the structure of the equilibrium in the pre-Act scenario is similar to that illustrated in Figure A.31 and the structure in the post-Act scenario is the same as presented in Figure 2.3. Therefore, the effects of the Act under the mixed strategy setting are similar to that illustrated in Figure A.32 and Table A.31. We can conclude that the Act may still backfire and introduce more child labor in supply chains under the mixed strategy setting.

3 Socially Conscious Consumers

With socially conscious consumers, the consumer valuation function is $v' := v + \delta\theta$ in the post-Act scenario. In this case, the manufacturer's profit function given in (2.1) is changed to

$$U_\delta(w,\theta,d) = v + \delta\theta - w - I(\theta) - deg(1-\theta). \tag{A.10}$$

The equilibrium outcomes in the post-Act scenarios are summarized in the following corollary.

Corollary 8 *With socially conscious consumers, the subgame-perfect equilibrium in the post-Act scenario is the same as that presented in Proposition 2 except the following changes in thresholds:* $\xi_{1,\delta}^{post} = \xi_1^{post} + \delta\theta_H$, $\xi_{3,\delta}^{post} = \xi_3^{post} + \delta\theta_H$, $\xi_{5,\delta}^{post} = \xi_5^{post} + \delta\theta_H$, *and* $\xi_{6,\delta}^{post} = \xi_6^{post} + \delta\theta_H$.

4 Decision of Compensation in Corrective Actions

When the manufacturer chooses the amount of compensation m between low m_L and high m_H, where $0 \leq m_L < m_H$, the subgame-perfect equilibria in the pre-Act and post-Act scenarios are given respectively as follows.

Corollary 9 *When the manufacturer chooses* $m \in \{m_L, m_H\}$ *endogenously, the subgame-perfect equilibrium in the pre-Act scenario is*

$$\{w_m^{pre}, m_m^{pre}, \theta_m^{pre}, d_m^{pre}\} =$$
$$\begin{cases} (s_H - \Delta(\theta_L, m_L), m_L, \theta_L, 1) & if\ I > \xi_{1,m}^{pre}\ and\ g \leq \xi_{2,m}^{pre}; \\ (s_H - \Delta(\theta_H, m_L), m_L, \theta_H, 1) & if\ I < \xi_{1,m}^{pre}\ and\ I \leq \xi_{3,m}^{pre}; \\ (s_H + \Delta(\theta_L, m_L)/\gamma e - \Delta(\theta_L, m_L), m_L, \theta_L, 0) & otherwise, \end{cases}$$

where $\xi_{1,m}^{pre} = e\theta_H g$, $\xi_{2,m}^{pre} = \Delta(\theta_L, m_L)/\gamma e^2$ *and* $\xi_{3,m}^{pre} = \Delta(\theta_L, m_L)/\gamma e + \Delta(\theta_H, m_L) - \Delta(\theta_L, m_L) - ge(1-\theta_H)$.

Corollary 10 *When the manufacturer chooses* $m \in \{m_L, m_H\}$ *endogenously, the subgame-perfect equilibrium in the post-Act scenario is*

$$\{w^{post}, m^{post}, \theta^{post}, d^{post}\} =$$
$$\begin{cases} (s_H - \Delta(\theta_L, m_L), m_L, \theta_L, 1) & if\ g \leq \xi_{2,m}^{post}, I \geq \xi_{1,m}^{post}, I \geq \xi_{6,m}^{post}; \\ (s_H - \Delta(\theta_H, m_L), m_L, \theta_H, 1) & if\ I \leq \xi_{1,m}^{post}, I \leq \xi_{3,m}^{post}, g \leq \xi_{4,m}^{post}; \\ \left(s_H + \dfrac{\Delta(\theta_L, m_L)}{\gamma e} - \Delta(\theta_L, m_L), m_L, \theta_L, 0\right) & if\ g \geq \xi_{2,m}^{post}, I \geq \xi_{3,m}^{post}, I \geq \xi_{5,m}^{post}; \\ \left(s_H + \dfrac{\Delta(\theta_H, m_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H, m_H), m_H, \theta_H, 0\right) & if\ g \geq \xi_{4,m}^{post}, I \leq \xi_{5,m}^{post}, I \leq \xi_{6,m}^{post}, \end{cases}$$

where $\xi_{1,m}^{post} = \Delta(\theta_H, m_L) - \Delta(\theta_L, m_L) + ge\theta_H$, $\xi_{2,m}^{post} = \dfrac{\Delta(\theta_L, m_L)}{\gamma e^2}$, $\xi_{3,m}^{post} = \Delta(\theta_H, m_L) + \dfrac{\Delta(\theta_L, m_L)}{\gamma e} - \Delta(\theta_L, m_L) - ge(1-\theta_H)$, $\xi_{4,m}^{post} = \dfrac{\Delta(\theta_H, m_H)}{\gamma e^2(1-\theta_H)^2} - \dfrac{\Delta(\theta_H, m_H)}{e(1-\theta_H)} + \dfrac{\Delta(\theta_H, m_L)}{e(1-\theta_H)}$, $\xi_{5,m}^{post} = \Delta(\theta_H, m_H) -$

$\Delta(\theta_L, m_L) - \dfrac{\Delta(\theta_H, m_H)}{\gamma e(1-\theta_H)} + \dfrac{\Delta(\theta_L, m_L)}{\gamma e}$, *and* $\xi_{6,m}^{post} = \Delta(\theta_H, m_H) + ge - \dfrac{\Delta(\theta_H, m_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_L, m_L)$.

5 Restricted Penalty Schemes

Under the penalty scheme that requires only corrective actions, the supplier's decision on child labor depends entirely on the expected labor cost saving from using child labor $\Delta(\theta)$ for $\theta \in \{\theta_L, \theta_H\}$ (see the definition of $\Delta(\theta)$ in §2.6.3).

Corollary 11 *Under the penalty scheme that requires only corrective actions, the following results hold:*

(a)If $\Delta(\theta_L) > \Delta(\theta_H) \geq 0$, *the supplier always employs child labor for any* $\theta \in \{\theta_L, \theta_H\}$.

(b)If $0 \geq \Delta(\theta_L) > \Delta(\theta_H)$, *the supplier always combats child labor for any* $\theta \in \{\theta_L, \theta_H\}$.

(c)If $\Delta(\theta_L) > 0 \geq \Delta(\theta_H)$, *the supplier employs child labor if* $\theta = \theta_L$, *and combats child labor otherwise.*

In Corollary 11, cases (a) and (b) are somewhat unrealistic because the supplier's decision on child labor is unaffected by the manufacturer's inspection decision on θ. In case (c), the manufacturer can easily control the supplier's use of child labor through internal inspections. In all cases, a wholesale price does not affect the supplier's decision. Under the zero-tolerance policy, the manufacturer terminates her contract with a supplier whenever child labor is found. In this case, the supplier's profit function becomes

$$\Pi_{ct}(w, \theta, d) = (1 - \gamma d(\theta + (1-\theta)e))(w - s_H + d\Delta(0)). \tag{A.11}$$

In (A.11), $w - s_H + d\Delta(0)$ represents the supplier's profit when there is no risk of child labor being detected during inspections.[2] The term $\gamma d(\theta + (1-\theta)e)(w - s_H + d\Delta(0))$ represents the supplier's expected opportunity cost of employing child labor, in which $\theta + (1-\theta)e$ is the probability that the manufacturer will discontinue her contract with the supplier. The manufacturer's profit function remains the same.

Corollary 12 *Under the zero-tolerance policy, the following results hold.*

[2]Since the supplier is not required to remove child labor detected in internal inspections, the supplier's labor cost is either s_H if $d = 0$ or s_L if $d = 1$, even under a high inspection level ($\theta = \theta_H$).

(a) The subgame-perfect equilibrium in the pre-Act scenario is

$$\{w_{ct}^{pre}, \theta_{ct}^{pre}, d_{ct}^{pre}\} = \begin{cases} (s_H - \Delta(0), \theta_L, 1) & if\ I > \xi_{1,ct}^{pre}\ and\ g \leq \xi_{2,ct}^{pre}; \\ (s_H - \Delta(0), \theta_H, 1) & if\ I < \xi_{1,ct}^{pre}\ and\ I \leq \xi_{3,ct}^{pre}; \\ (s_H + \Delta(0)/\gamma e - \Delta(0),\ \theta_L, 0) & otherwise, \end{cases}$$

where $\xi_{1,ct}^{pre} = e\theta_H g$, $\xi_{2,ct}^{pre} = \Delta(0)/\gamma e^2$ *and* $\xi_{3,ct}^{pre} = \Delta(0)/\gamma e - ge(1-\theta_H)$.

(b) The subgame-perfect equilibrium in the post-Act scenario is

$$\left\{w_{ct}^{post}, \theta_{ct}^{post}, d_{ct}^{post}\right\} =$$
$$\begin{cases} (s_H - \Delta(0), \theta_L, 1) & if\ I \geq \xi_{1,ct}^{post}, g \leq \xi_{2,ct}^{post}, I \geq \xi_{6,ct}^{post}; \\ (s_H - \Delta(0), \theta_H, 1) & if\ I \leq \xi_{1,ct}^{post}, I \leq \xi_{3,ct}^{post}, g \leq \xi_{4,ct}^{post}; \\ (s_H + (1/(\gamma e) - 1)\Delta(0), \theta_L, 0) & if\ g \geq \xi_{2,ct}^{post}, I \geq \xi_{3,ct}^{post}, I \geq \xi_{5,ct}^{post}; \\ (s_H + \{1/[\gamma(\theta_H + (1-\theta_H)e)] - 1\}\Delta(0), \theta_H, 0) & if\ g \geq \xi_{4,ct}^{post}, I \leq \xi_{5,ct}^{post}, I \leq \xi_{6,ct}^{post}, \end{cases}$$

where $\xi_{1,ct}^{post} = e\theta_H g$, $\xi_{2,ct}^{post} = \Delta(0)/\gamma e^2$, $\xi_{3,ct}^{post} = \Delta(0)/\gamma e - ge(1-\theta_H)$,
$\xi_{4,ct}^{post} = \frac{\Delta(0)}{e\gamma(1-\theta_H)(\theta_H+(1-\theta_H)e)}$, $\xi_{5,ct}^{post} = \frac{\Delta(0)}{\gamma e} - \frac{\Delta(\theta_L)}{\gamma(\theta_H+(1-\theta_H)e)}$, *and* $\xi_{6,ct}^{post} = ge - \frac{\Delta(0)}{\gamma(\theta_H+(1-\theta_H)e)}$.

A.4 Additional Proofs

Proof of Lemma 7. The first condition $\gamma e < \dfrac{1}{2-\theta_H}$ ensures that $m_1 < m_2$ and there exists $m \in [m_1, m_2]$; The second condition $m > m_1$ suggests $\xi_2^{post} > \xi_4^{post}$. Given that $\xi_2^{post} > \xi_4^{post}$, program (A.1) can be simplified as follows:

$$\max_{\theta \in \{\theta_L, \theta_H\}} \begin{cases} \{U(s_H - \Delta(\theta_L), \theta_L, 1), U(s_H - \Delta(\theta_H), \theta_H, 1)\} & \text{if } g \leq \xi_4^{post}; \\ \left\{U(s_H - \Delta(\theta_L), \theta_L, 1), U\left(s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H), \theta_H, 0\right)\right\} & \text{if } \xi_4^{post} < g \leq \xi_2^{post}; \\ \left\{U\left(s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L), \theta_L, 0\right), U\left(s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H), \theta_H, 0\right)\right\} & \text{if } g > \xi_2^{post}. \end{cases} \tag{A.12}$$

Therefore, we obtain the following results: (i) $(s_H - \Delta(\theta_L), \theta_L, 1)$ is the equilibrium outcome if and only if $\left\{g \leq \xi_2^{post}, I \geq \xi_1^{post}\right\}$ or

$\left\{\xi_4^{post} < g \leq \xi_2^{post}, I \geq \xi_6^{post}\right\}$, where

$$\begin{aligned} & U\left(s_H - \Delta\left(\theta_L\right), \theta_L, 1\right) \geq U\left(s_H + \tfrac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta\left(\theta_H\right), \theta_H, 0\right) \\ & \Leftrightarrow I \geq \Delta\left(\theta_H\right) - \tfrac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta\left(\theta_L\right) + eg := \xi_6^{post}. \end{aligned} \tag{A.13}$$

We can show that these conditions can be further simplified to $I \geq \xi_1^{post}, g \leq \xi_2^{post}, I \geq \xi_6^{post}$. (ii) $\left(s_H - \Delta\left(\theta_H\right), \theta_H, 1\right)$ is the equilibrium outcome if and only if $g \leq \xi_4^{post}$ and $I \geq \xi_1^{post}$. In addition, the condition $m \leq m_2$ ensures that the parametric space for $\left(s_H - \Delta\left(\theta_H\right), \theta_H, 1\right)$ to be in equilibrium is not null; i.e., $m \leq m_2 \Leftrightarrow \{(I,g) | I \leq \xi_1^{post}, g \leq \xi_4^{post}, I \geq 0, g \geq .0$ (iii) $\left(s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta\left(\theta_L\right), \theta_L, 0\right)$ is the equilibrium outcome if and only if $g > \xi_2^{post}$ and $I \geq \xi_5^{post}$. (iv) $\left(s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta\left(\theta_H\right), \theta_H, 0\right)$ is the equilibrium outcome if and only if $\left\{\xi_4^{post} < g \leq \xi_2^{post}, I \leq \xi_6^{post}\right\}$ or $\left\{g \geq \xi_2^{post}, I \leq \xi_5^{post}\right\}$. Similar to (ii), these conditions can be simplified to $\xi_4^{post} < g$, $I \leq \xi_6^{post}$ and $I \leq \xi_5^{post}$. ■ **Proof of Lemma 8.** Note that the equilibrium outcome presented in (A.9) in Lemma 8 is a special case of the outcome given in (A.8) presented in Lemma 7. To prove Lemma 8, we will show that under either condition (a) $\left\{m_2 < m < \overline{m}, \gamma e < \frac{1}{2-\theta_H}\right\}$ or (b) $\gamma e > \frac{1}{2-\theta_H}$, the program (A.1) can be simplified as (A.12), and the parameter space for the outcome $\left(s_H - \Delta\left(\theta_H\right), \theta_H, 1\right)$ is null. First, we show that either condition (a) $\left\{m_2 < m < \overline{m}, \gamma e < \frac{1}{2-\theta_H}\right\}$ or (b) $\gamma e > \frac{1}{2-\theta_H}$ suggests $m > m_1$ and $m > m_2$. Condition (a) suggests $\gamma e < \frac{1}{2-\theta_H}$, which indicates $m_2 > m_1$ (cf. the proof of Lemma 7); moreover, since $m > m_2$ holds under condition (a), it is indicated that under condition (a), $m > m_2 > m_1$. Under condition (b), $\gamma e > \frac{1}{2-\theta_H}$holds and indicates that $m_2 < m_1$ and $m_1 < \underline{m}$ (cf. the proof of Lemma 6); therefore, $\forall m \in [\underline{m}, \overline{m}]$, it satisfies $m > m_1 > m_2$. Second, we solve (A.12) as follows: given that $m > m_1$, the program (A.1) can be simplified as (A.12); moreover, when $m > m_2$ holds, the parameter space for the outcome $\left(s_H - \Delta\left(\theta_H\right), \theta_H, 1\right)$ is null (cf. the proof of Lemma 7). As a result, we can remove the outcome $\left(s_H - \Delta\left(\theta_H\right), \theta_H, 1\right)$ from the equilibrium presented in (A.8) as well as the corresponding conditions. Therefore, the equilibrium outcome in (A.8) reduces to (A.9). ■ **Proof of Corollary 4.** We prove the corollary via three steps: we first derive the supplier's best response $d_{cont}^{pre}(\theta, w)$ and the manufacturer's best response $\theta_{cont}^{pre}(d, w)$, then we show the Nash equilibrium $\left(\theta_{cont}^{pre}(w), d_{cont}^{pre}(w)\right)$ at the production stage, finally we derive the subgame-perfect equilibrium in the pre-Act scenario under the new setting. First, the following

lemma presents the supplier's best response $d_{cont}^{pre}(\theta, w)$.

Lemma 9 *(a) In the pre-Act scenario, for any given w, the supplier's best response $d_{cont}^{pre}(\theta, w) \in [0,1]$ to the manufacturer's decision on inspection level θ is:*

$$d_{cont}^{pre}(\theta, w) = \begin{cases} 1 & \text{if } w < s_H + \beta(\theta) - 2\Delta(\theta); \\ d_0(\theta, w) & \text{if } s_H + \beta(\theta) - 2\Delta(\theta) \leq w < s_H + \beta(\theta); \\ 0 & \text{if } w \geq s_H + \beta(\theta), \end{cases} \tag{A.14}$$

where $\beta(\theta) = \dfrac{\Delta(\theta)}{\gamma e(1-\theta)} = \dfrac{(1-\theta)(s_H - s_L) - m\theta}{\gamma e(1-\theta)}$, and $d_0(\theta, w) = \dfrac{1}{2\gamma e(1-\theta)} - \dfrac{w - s_H}{2\Delta(\theta)}$.
(b) $d_{cont}^{pre}(\theta, w)$ is monotonically decreasing in w.
(c) $d_{cont}^{pre}(\theta, w)$ is monotonically decreasing in θ under the following assumption ($A3^{cont}$):

$$2\gamma e \leq \frac{m}{(s_H - s_L + m)(1 - \theta_H)}. \tag{A.15}$$

Proof: (a) Since $\dfrac{\partial^2 \Pi}{\partial d^2} = -2\gamma e(1-\theta)\Delta(\theta) > 0$, $\Pi(w, \theta, d)$ is strictly concave in d for any given w and θ. Therefore, the supplier's best response on the child labor employment level can be derived through first-order conditions. We discuss the results based on there cases. Case 1: When $w < s_H + \beta(\theta) - 2\Delta(\theta)$, we have $d_0 > 1$ and hence $\dfrac{\partial \Pi}{\partial d} \geq 0$ for $d \in [0,1]$, so the best response $d_{cont}^{pre}(\theta, w) = 1$. Case 2: When $s_H + \beta(\theta) - 2\Delta(\theta) \leq w < s_H + \beta(\theta)$, we have $0 < d_0 \leq 1$, so the best response $d_{cont}^{pre}(\theta, w) = d_0$. Case 3: When $w \geq s_H + \beta(\theta)$, we have $d_0 \leq 0$ and hence $\dfrac{\partial \Pi}{\partial d} \leq 0$ for all $d \in [0,1]$, so the best response $d_{cont}^{pre}(\theta, w) = 0$. A combination of Case 1, 2 and 3 proves (A.14). (b) From (A.14), it is straightforward to find that for any θ, $d_{cont}^{pre}(\theta, w)$ is monotonically decreasing in w. (c) Assumption (A.15) suggests $2\gamma e \leq \dfrac{m}{(s_H - s_L + m)(1 - \theta_H)} \iff \beta(\theta_L) - 2\Delta(\theta_L) \geq \beta(\theta_H) - 2\Delta(\theta_H)$, which indicates $d_0(\theta_H, w) \leq d_0(\theta_L, w)$. Given that $d_0(\theta_H, w) \leq d_0(\theta_L, w)$, it is not difficult to verify that for any w, $d_{cont}^{pre}(\theta, w)$ is monotonically decreasing in θ. ■

Remark 3 *Assumption ($A3^{cont}$) given in (A.15) is a counter-assumption of Assumption (A3) under the base setting, and it ensures that the supplier' expected profit from hiring child labor is lower when the manufacturer conducts inspections than when no inspections are undertaken; otherwise the supplier's best response would be to hire more child labor when the manufacturer*

chooses a higher inspection level, which is unrealistic and thus not considered.

Assumption (A3^{cont}) in (A.15) holds if and only if $\theta_H \geq 1 - \dfrac{m}{2(s_H - s_L + m)\gamma e}$. *To ensure that there exists such a* θ_H *that satisfies both Assumption (A2) and Assumption (A3^{cont}), we need to have* $1 - \dfrac{m}{2(s_H - s_L + m)\gamma e} \leq \dfrac{s_H - s_L}{s_H - s_L + m}$, *which indicates another assumption under the setting of a continuous child labor decision,* $\gamma e \leq \dfrac{1}{2}$; *Otherwise if* $\gamma e > \dfrac{1}{2}$, *Assumption (A3^{cont}) will always be violated.*

The manufacturer's best response function $\theta_{cont}^{pre}(d, w)$ to the supplier's decision on d for any given w is:

$$\theta_{cont}^{pre}(d, w) = \begin{cases} \theta_H & \text{if } \ d \geq d_I; \\ \theta_L & \text{if } \ d < d_I. \end{cases} \tag{A.16}$$

Second, by examining $d_{cont}^{pre}(\theta, w)$ in (A.14) and $\theta_{cont}^{pre}(d, w)$ in (A.16), we obtain a fixed point $(\theta_{cont}^{pre}(w), d_{cont}^{pre}(w))$ that satisfies $\theta_{cont}^{pre}(w) = \theta_{cont}^{pre}(d_{cont}^{pre}(w), w)$ and $d_{cont}^{pre}(w) = d_{cont}^{pre}(\theta_{cont}^{pre}(w), w)$, which is shown in the following lemma.

Lemma 10 *When the supplier can choose a continuous level of child labor employment, for any given* w, *at the production stage the supplier's and the manufacturer's decision on child labor employment and inspection level* $(\theta_{cont}^{pre}(w), d_{cont}^{pre}(w))$ *is:*

(a) When $I > e\theta_H g$,

$$\begin{aligned} &(\theta_{cont}^{pre}(w), d_{cont}^{pre}(w)) = (\theta_L, d_{cont}^{pre}(\theta_L, w)) \\ &= \begin{cases} (\theta_L, 1) & \text{if } w < s_H + \beta(\theta_L) - 2\Delta(\theta_L); \\ (\theta_L, d_0(\theta_L, w)) & \text{if } s_H + \beta(\theta_L) - 2\Delta(\theta_L) \leq w < s_H + \beta(\theta_L); \\ (\theta_L, 0) & \text{if } w \geq s_H + \beta(\theta_L). \end{cases} \end{aligned} \tag{A.17}$$

(b) When $I \leq e\theta_H g$:

$$
\begin{aligned}
&(\theta^{pre}_{cont}(w), d^{pre}_{cont}(w)) = \\
&\begin{cases}
(\theta_H, 1) & \text{if } w \leq s_H + \beta(\theta_H) - 2\Delta(\theta_H); \\
(\theta_H, d_0(\theta_H, w)) & \text{if } s_H + \beta(\theta_H) - 2\Delta(\theta_H) \leq w \leq s_H + \beta(\theta_H) - 2d_I\Delta(\theta_H); \\
(\theta_L, d_0(\theta_L, w)) & \text{if } s_H + \beta(\theta_L) - 2d_I\Delta(\theta_L) \leq w \leq s_H + \beta(\theta_L); \\
(\theta_L, 0) & \text{if } w \geq s_H + \beta(\theta_L).
\end{cases}
\end{aligned} \tag{A.18}
$$

The proof of Lemma 10 is omitted and can be provided under request. The equilibrium outcome shown in (A.17) and (A.18) is unique because (i) Lemma 9 suggests that $d^{pre}_{cont}(\theta, w)$ is monotonically decreasing in θ and(ii) $\theta^{pre}_{cont}(d, w)$ given in (A.16) strictly increases in d. Finally, we substitute the unique equilibrium outcome $(\theta^{pre}(w), d^{pre}(w))$ given in Lemma 10 into the maximization program (2.12) to solve for the subgame-perfect equilibrium in the pre-Act scenario. We discuss two cases. Case 1: $I > e\theta_H g$. Substituting $(\theta^{pre}_{cont}(w), d^{pre}_{cont}(w))$ in (A.17) into program (2.12) yields:

$$
\max_w \begin{cases}
U(w, \theta_L, 1) & \text{if } w \in [s_H - \Delta(\theta_L), s_H + \beta(\theta_L) - 2\Delta(\theta_L)]; \\
U(w, \theta_L, d_0(\theta_L, w)) & \text{if } w \in [s_H + \beta(\theta_L) - 2\Delta(\theta_L), s_H + \beta(\theta_L)]; \\
U(w, \theta_L, 0) & \text{if } w \in [s_H + \beta(\theta_L), +\infty).
\end{cases}
$$

We further obtain the following subgame-perfect equilibrium when $I > e\theta_H g$:

$$
(w^{pre}_{cont}, \theta^{pre}_{cont}, d^{pre}_{cont}) = \begin{cases}
(s_H - \Delta(\theta_L), \theta_L, 1) & \text{if } I > \xi^{pre}_{1,cont} \text{ and } g \leq \xi^{pre}_{2,cont}; \\
(s_H + \Delta(\theta_L)/\gamma e, \theta_L, 0) & \text{if } I > \xi^{pre}_{1,cont} \text{ and } g > \xi^{pre}_{2,cont}.
\end{cases} \tag{A.19}
$$

Case 2: $I \leq e\theta_H g$. Substituting $(\theta^{pre}_{cont}(w), d^{pre}_{cont}(w))$ in (A.18) into program (2.12) yields:

$$
\max_w \begin{cases}
U(w, \theta_H, 1) & \text{if } w \in [s_H - \Delta(\theta_H), s_H + \beta(\theta_H) - 2\Delta(\theta_H)]; \\
U(w, \theta_H, d_0(\theta_H, w)) & \text{if } w \in [s_H + \beta(\theta_H) - 2\Delta(\theta_H), s_H + \beta(\theta_H) - 2d_I\Delta(\theta_H)]; \\
U(w, \theta_L, d_0(\theta_L, w)) & \text{if } w \in [s_H + \beta(\theta_L) - 2d_I\Delta(\theta_L), s_H + \beta(\theta_L)]; \\
U(w, \theta_L, 0) & \text{if } w \in [s_H + \beta(\theta_L), +\infty).
\end{cases}
$$

Further analysis indicates that the equilibrium wholesale price w_{cont}^{pre} is one of three options: $s_H - \Delta(\theta_H)$, $s_H + \beta(\theta_H) - 2d_I\Delta(\theta_H)$ and $s_H + \beta(\theta_L)$. By comparing the three options pair by pair, we obtain the following results: when $I \leq e\theta_H g$,

$$w_{cont}^{pre} = \begin{cases} s_H - \Delta(\theta_H) & \text{if } \xi_{4,cont}^{pre} \leq I \leq \xi_{3,cont}^{pre}; \\ s_H + \beta(\theta_L) & \text{if } I > \xi_{3,cont}^{pre}, I > \xi_{5,cont}^{pre}; \\ s_H + \beta(\theta_H) - 2d_I\Delta(\theta_H) & \text{if } I < \xi_{5,cont}^{pre}, I < \xi_{4,cont}^{pre}. \end{cases}$$

Therefore, for $I \leq e\theta_H g \equiv \xi_{1,cont}^{pre}$, the subgame-perfect equilibrium is

$$(w_{cont}^{pre}, \theta_{cont}^{pre}, d_{cont}^{pre}) = \begin{cases} (s_H - \Delta(\theta_H), \theta_H, 1) & \text{if } I \leq \xi_{1,cont}^{pre}, I \leq \xi_{3,cont}^{pre}, I > \xi_{4,cont}^{pre}; \\ (s_H + \Delta(\theta_L)/\gamma e, \theta_L, 0) & \text{if } I \leq \xi_{1,cont}^{pre}, I > \xi_{3,cont}^{pre}, I > \xi_{5,cont}^{pre}; \\ s_H + \beta(\theta_H) - 2d_I\Delta(\theta_H) & \text{if } I \leq \xi_{1,cont}^{pre}, I \leq \xi_{5,cont}^{pre}, I \leq \xi_{4,cont}^{pre}. \end{cases} \tag{A.20}$$

A combination of (A.19) and (A.20) proves the corollary. ■ **Proof of Corollary 5.** Similar to the proof of Proposition 2, in the following we first simplify the program (2.17), and then solve for the subgame-perfect equilibrium through three lemmas. First, substituting $d_{cont}^{post}(\theta, w) = d_{cont}^{pre}(\theta, w)$ given in (A.14) into (2.17) yields:

$$\max_{w,\theta} \begin{cases} U(w, \theta, 1) & \text{if } w \in [s_H - \Delta(\theta), s_H + \beta(\theta) - 2\Delta(\theta)); \\ U(w, \theta, d_0(\theta, w)) & \text{if } w \in [s_H + \beta(\theta) - 2\Delta(\theta), s_H + \beta(\theta)); \\ U(w, \theta, 0) & \text{if } w \in [s_H + \beta(\theta), +\infty). \end{cases}$$

We can further simplify the above program as:

$$\max_{\theta} \begin{cases} U(s_H - \Delta(\theta), \theta, 1) = v - (s_H - \Delta(\theta)) - I(\theta) - e(1-\theta)g; \\ U\left(s_H + \frac{\Delta(\theta)}{\gamma e(1-\theta)}, \theta, 0\right) = v - \left(s_H + \frac{\Delta(\theta)}{\gamma e(1-\theta)}\right) - I(\theta). \end{cases} \tag{A.21}$$

The analysis of (A.21) is quite similar to the analysis of (A.1). Similar to the proof of Proposition 2, in the rest of the proof, we provide the remaining conditions under three possible structures of the equilibrium outcome that depend on the value of m. Lemmas 11, 12 and 13 together prove Corollary

5. The proofs of the three lemma are omitted and can be provided under request. For convenience, we rewrite assumptions (A2) and (A3$^{\text{cont}}$), respectively, as $m \leq \overline{m} \equiv \dfrac{(1-\theta_H)(s_H - s_L)}{\theta_H}$ and $m \geq \underline{m}^{cont} \equiv \dfrac{2\gamma e(1-\theta_H)(s_H - s_L)}{1-2\gamma e(1-\theta_H)}$.

Lemma 11 *When the supplier can choose a continuous level of child labor employment, under the conditions* $\gamma e < \frac{-2+\theta_H+\sqrt{\theta_H^2-6\theta_H+6}}{2(1-\theta_H)}$ *and* $\underline{m}^{cont} \leq m < m_1^{cont} \equiv \frac{(1-\theta_H)(s_H-s_L)}{1+\gamma e(1-\theta_H)}$*, the equilibrium outcome of program (A.21) is*

$$\left\{w_{cont}^{post}, \theta_{cont}^{post}, d_{cont}^{post}\right\} =$$
$$\begin{cases} (s_H - \Delta(\theta_L), \theta_L, 1) & \text{if } I \geq \xi_{1,cont}^{post},\ g \leq \xi_{2,cont}^{post}; \\ (s_H - \Delta(\theta_H), \theta_H, 1) & \text{if } I \leq \xi_{1,cont}^{post},\ I \leq \xi_{3,cont}^{post},\ g \leq \xi_{4,cont}^{post}; \\ (s_H + \Delta(\theta_L)/(\gamma e), \theta_L, 0) & \text{if } g \geq \xi_{2,cont}^{post},\ I \geq \xi_{3,cont}^{post},\ I \geq \xi_{5,cont}^{post}; \\ (s_H + \Delta(\theta_H)/(\gamma e(1-\theta_H)), \theta_H, 0) & \text{if } g \geq \xi_{4,cont}^{post},\ I \leq \xi_{5,cont}^{post}. \end{cases}$$

Lemma 12 *When the supplier can choose a continuous level of child labor employment, under the conditions* $\gamma e < \frac{-2+\sqrt{1+2(1-\theta_H)^2}}{2(1-\theta_H)}$ *and* $m_1^{cont} \leq m \leq m_2^{cont} \equiv \frac{(s_H-s_L)(1-\theta_H)}{\gamma e(1-\theta_H)+\theta_H}$*, the equilibrium outcome of program (A.21) is*

$$\left\{w_{cont}^{post}, \theta_{cont}^{post}, d_{cont}^{post}\right\} =$$
$$\begin{cases} (s_H - \Delta(\theta_L), \theta_L, 1) & \text{if } I \geq \xi_{1,cont}^{post},\ g \leq \xi_{2,cont}^{post}, I \geq \xi_{6,cont}^{post}; \\ (s_H - \Delta(\theta_H), \theta_H, 1) & \text{if } I \leq \xi_{1,cont}^{post},\ g \leq \xi_{4,cont}^{post}; \\ (s_H + \Delta(\theta_L)/(\gamma e), \theta_L, 0) & \text{if } g \geq \xi_{2,cont}^{post},\ I \geq \xi_{5,cont}^{post}; \\ (s_H + \Delta(\theta_H)/(\gamma e(1-\theta_H)), \theta_H, 0) & \text{if } I \leq \xi_{6,cont}^{post},\ g \geq \xi_{4,cont}^{post},\ I \leq \xi_{5,cont}^{post}. \end{cases}$$

Lemma 13 *When the supplier can choose a continuous level of child labor employment, under the conditions* $m_2^{cont} < m < \overline{m}$*, the equilibrium outcome of program (A.21) is*

$$\left\{w_{cont}^{post}, \theta_{cont}^{post}, d_{cont}^{post}\right\} =$$
$$\begin{cases} (s_H - \Delta(\theta_L), \theta_L, 1) & \text{if } g \leq \xi_{2,cont}^{post}, I \geq \xi_{6,cont}^{post}; \\ (s_H + \Delta(\theta_L)/(\gamma e), \theta_L, 0) & \text{if } g \geq \xi_{2,cont}^{post},\ I \geq \xi_{5,cont}^{post}; \\ (s_H + \Delta(\theta_H)/(\gamma e(1-\theta_H)), \theta_H, 0) & \text{if } I \leq \xi_{6,cont}^{post},\ I \leq \xi_{5,cont}^{post}. \end{cases}$$

■

Proof of Corollary 6. We first derive the (mixed) Nash equilibrium $\left(P_\theta^{pre}(w), P_d^{pre}(w)\right)$ in the simultaneous game under a fixed wholesale price w, then derive the subgame perfect equilibrium. First, under any given wholesale price w, the following table presents the strategic form of the subgame. The top expression in each cell is the expected profit of the supplier derived based on (2.3), and the bottom expression is the profit of the manufacturer derived based on (2.1).

	$\theta = \theta_L$	$\theta = \theta_H$
$d = 0$	$(w - s_H, v - w)$	$(w - s_H, v - w - I)$
$d = 1$	$\begin{pmatrix} (1-\gamma e)(w - s_H + \Delta(\theta_L)), \\ v - w - eg \end{pmatrix}$	$\begin{pmatrix} (1-\gamma e(1-\theta_H))(w - s_H + \Delta(\theta_H)), \\ v - w - eg(1-\theta_H) - I \end{pmatrix}$

Based on the table above, it is not difficult to derive the supplier's and the manufacturer's best response functions. The supplier's best response is

$$P_d^{pre}(P_\theta, w) = \begin{cases} 1 & \text{if } P_\theta < f(w) \\ x \in [0,1] & \text{if } P_\theta = f(w) \\ 0 & \text{if } P_\theta > f(w) \end{cases},$$

where $f(w) \equiv \frac{\Delta(\theta_L) - \gamma e(w - s_H + \Delta(\theta_L))}{\Delta(\theta_L) - \gamma e(w - s_H + \Delta(\theta_L)) - (\Delta(\theta_H) - \gamma e(1-\theta_H)(w - s_H + \Delta(\theta_H)))}$, and x denotes any value between 0 and 1. The manufacturer's best response is:

$$P_\theta^{pre}(P_d, w) = \begin{cases} 1 & \text{if } P_d > d_I \\ x \in [0,1] & \text{if } P_d = d_I \\ 0 & \text{if } P_d < d_I. \end{cases}$$

Based on the two players' best responses, we obtain the Nash equilibrium $\left(P_\theta^{pre}(w), P_d^{pre}(w)\right)$ for any given wholesale price w in the simultaneous subgame:

$$\left(P_\theta^{pre}(w), P_d^{pre}(w)\right) = \begin{cases} (0,0) & \text{if } w > s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L) \\ (0, x \in [0, \min(d_I, 1)]) & \text{if } w = s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L) \ \ (\Leftrightarrow f(w) = 0) \\ (0,1) & \text{if } w < s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L) \text{ and } I > e\theta_H g \\ (f(w), d_I) & \text{if } s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H) < w < s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L) \text{ and } I < e\theta_H g \\ (x \in [0, f(w)], 1) & \text{if } s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H) < w < s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L) \text{ and } I = e\theta_H g \\ (1, x \in [d_I, 1]) & \text{if } w = s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H) \ \ (\Leftrightarrow f(w) = 1) \text{ and } I < e\theta_H g \\ (1,1) & \text{if } w < s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H) \text{ and } I < e\theta_H g \\ (x \in [0,1], 1) & \text{if } w \leq s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H) \text{ and } I = e\theta_H g \end{cases} \tag{A.22}$$

where x denotes any value within the given bound. To proceed, we need to refine the multiple equilibria $\left(P_\theta^{pre}(w), P_d^{pre}(w)\right)$ given in (A.22). We adopt the following rule: if one player are indifferent between multiple actions, the player would choose an action that maximizes the other player's profit. In particular, if the supplier can choose an arbitrary probability on employing child labor (e.g., $P_d^{pre}(w) = x \in [0, \min(d_I, 1)]$), the supplier is supposed not to hire child labor (e.g., $P_d^{pre}(w) = 0$), because zero child labor employment minimizes the manufacturer's goodwill loss. Or if the manufacturer can choose an arbitrary probability on undertaking inspections or not, the manufacturer would choose not to undertake inspections when the supplier hires child labor because the supplier's expected profit from hiring child labor is lower when the manufacturer conducts inspections than when no inspections are undertaken (cf. (2.8)). As a result, the mixed equilibrium $\left(P_\theta^{pre}(w), P_d^{pre}(w)\right)$ is refined as follows:

$$\left(P_\theta^{pre}(w), P_d^{pre}(w)\right) = \begin{cases} (0,0) & \text{if } w \geq s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L) \\ (0,1) & \text{if } w < s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L) \text{ and } I \geq e\theta_H g \\ (1,1) & \text{if } w < s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H) \text{ and } I < e\theta_H g \\ (f(w), d_I) & \text{if } s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H) \leq w < s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L) \text{ and } I < e\theta_H g \end{cases} \tag{A.23}$$

By comparing the mixed Nash equilibrium $\left(P_\theta^{pre}(w), P_d^{pre}(w)\right)$ given in (A.23) with $(\theta^{pre}(w), d^{pre}(w))$ given in (2.11), we can infer that the first three Nash equilibrium outcomes in (A.23) are the same as the three pure strategies in (2.11), respectively. Only the last outcome $(d_I, f(w))$ constitutes a mixed Nash equilibrium. Second, similar to the proof of Proposition 1, we substitute $\left(P_\theta^{pre}(w), P_d^{pre}(w)\right)$ given in (A.23) into a revised maximization program (2.12), in which $(\theta^{pre}(w), d^{pre}(w))$ is replaced with $\left(P_\theta^{pre}(w), P_d^{pre}(w)\right)$, the constraints (2.5)-(2.7) are revised accordingly, and the profit functions are also revised to fit in the mixed strategy settings. The revised program can be simplified as:

$$\max_w \begin{cases} U(w,\theta_L,0) = v - w \text{ where } w = s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L); & \\ U(w,\theta_L,1) = v - w - eg \text{ where } w = s_H - \Delta(\theta_L) & \text{if } I \geq \xi_{1,mixed}^{pre}; \\ U(w,\theta_H,1) = v - w - eg \text{ where } w = s_H - \Delta(\theta_H) & \text{if } I < \xi_{1,mixed}^{pre}; \\ d_I U(w,\theta_H,1) + (1-d_I) U(w,\theta_H,0) = v - w - \frac{I}{\theta_H} & \\ \qquad \text{where } w = s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H) & \text{if } I < \xi_{1,mixed}^{pre}. \end{cases}$$

The simplification for the first three pure strategies are the same as that given in the proof of Proposition 1. The only difference lies in the simplification for the mixed equilibrium $(d_I, f(w))$. In the simplification, $(d_I, f(w))$ is reduced to be $(d_I, 1)$ with a wholesale price $w = s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H)$. The reason is as follows. Under the mixed equilibrium $(d_I, f(w))$, the manufacturer's profit is

$$\begin{aligned} & f(w)\left(d_I U(w,\theta_H,1) + (1-d_I) U(w,\theta_H,0)\right) + (1-f(w))\left(d_I U(w,\theta_L,1) + (1-d_I) U(w,\theta_L,0)\right) \\ & = v - w - I \cdot f(w) - egd_I\left(1 - f(w)\theta_H\right) \overset{d_I = I/(e\theta_H g)}{=} v - w - I/\theta_H, \end{aligned}$$

which linearly decreases in w. Therefore, under the constraint $s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H) \leq w < s_H + \frac{\Delta(\theta_L)}{\gamma e} - \Delta(\theta_L)$ and a revised (2.5), the optimal wholesale price for the manufacturer to induce the mixed equilibrium $(d_I, f(w))$ is $w = s_H + \frac{\Delta(\theta_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H)$, which suggests that $P_\theta^{pre}(w) = 1$. Lastly, we solve for the equilibrium wholesale price w_{mixed}^{pre}. The procedure is similar to the proof of Proposition 1 and is omitted here. ■ **Proof of Corollary 7.** The proof is provided in the main body. **Proof of Corollary 8.** Corollary 8 can be proved by replicating the proof of proposition

2 with $U(w,\theta,d)$ replaced by $U_\delta(w,\theta,d)$ defined in (A.10). It is not difficult to verify that the equilibrium outcome would be affected by the new setting in that the thresholds would change as follows: $\xi_i^{\delta\text{-}post} \equiv \xi_i^{post}$ where $i = 2,4$ (see (A.2) and (A.3)); and $\xi_i^{\delta\text{-}post} \equiv \xi_i^{post} + \delta\theta_H$ where $i = 1,3,5,6$ (see (A.5), (A.6), (A.7) and (A.13)). ■ **Proof of Corollary 9.** Similar to the derivation of the subgame-perfect equilibrium in the pre-Act scenario in the base setting, we will first derive the two players' best response, and then characterize Nash equilibrium in the subgame and finally the subgame-perfect equilibrium. Before we proceed, we introduce a counter-assumption to Assumption (A2). We assume that the supplier could save his labor cost (in expectation) by hiring child labor even when the manufacturer has chosen to undertake internal inspections; i.e., $\Delta(\theta_H, m_L) \geq 0$ and $\Delta(\theta_H, m_H) \geq 0$. Since $m_L < m_H$, to ensure $\Delta(\theta_H, m_L) \geq \Delta(\theta_H, m_H)$, we have Assumption (A2'): $m_L < m_H \leq \overline{m} = \dfrac{(1-\theta_H)(s_H - s_L)}{\theta_H}$. First, under the new setting, for any given w, the supplier's best response child labor decision to the manufacturer's decision on θ and m is

$$
\begin{aligned}
&d^{pre}(\theta, m, w) = 1 \ \forall \theta \in \{\theta_L, \theta_H\}, \forall m \in \{\theta_L, \theta_H\} \text{ for } w \in [0, \beta(\theta_H, m_H));\\
&\begin{cases} d^{pre}(\theta_L, m_L, w) = d^{pre}(\theta_L, m_H, w) = 1 \\ d^{pre}(\theta_H, m_L, w) = 1, d^{pre}(\theta_H, m_H, w) = 0 \end{cases} \text{for } w \in [\beta(\theta_H, m_H), \beta(\theta_H, m_L));\\
&d^{pre}(\theta_L, m, w) = 1 \text{ and } d^{pre}(\theta_H, m, w) = 0 \ \forall m \in \{\theta_L, \theta_H\} \text{ for } w \in [\beta(\theta_H, m_L), \beta(\theta_L, m_L));\\
&d^{pre}(\theta, m, w) = 0 \ \forall \theta \in \{\theta_L, \theta_H\}, \forall m \in \{\theta_L, \theta_H\} \text{ for } w \in [\beta(\theta_L, m_L), +\infty),
\end{aligned} \tag{A.24}
$$

where $\beta(\theta, m) = s_H + \frac{\Delta(\theta,m)}{\gamma e(1-\theta)} - \Delta(\theta)$. Note that the above best response function holds under the assumption (A3'): $\beta(\theta_H, m_L) < \beta(\theta_L, m_L) \Leftrightarrow m_L > \underline{m} = \dfrac{\gamma e(1-\theta_H)(s_H - s_L)}{1 - \gamma e(1-\theta_H)}$, which is a counterpart assumption of Assumption (A3) under the base setting. Correspondingly, for any given w the manufacturer's best response to the supplier's decision d are:

$$
\begin{aligned}
&\{\theta^{pre}, m^{pre}\}(d=0, w) = \{\theta_L, m_L\};\\
&\{\theta^{pre}, m^{pre}\}(d=1, w) = \begin{cases} \{\theta_H, m_L\} \text{ or } \{\theta_H, m_H\} & \text{if } I < e\theta_H g; \\ \{\theta_L, m_L\} & \text{if } I \geq e\theta_H g. \end{cases}
\end{aligned}
$$

Second, by examining $d^{pre}(\theta, m, w)$ and $\{\theta^{pre}, m^{pre}\}(d, w)$, we derive a fixed point $(\theta^{pre}(w), m^{pre}(w), d^{pre}(w))$ that satisfies

$$\begin{aligned}\{\theta^{pre}(w), m^{pre}(w)\} &= \{\theta^{pre}, m^{pre}\}(d^{pre}(w), w)\\ d^{pre}(w) &= d^{pre}(\{\theta^{pre}(w), m^{pre}(w)\}, w).\end{aligned}$$

We obtain the following result:

$$(\theta^{pre}(w), m^{pre}(w), d^{pre}(w)) = \begin{cases} (\theta_L, m_L, 1) & \text{if } I \geq e\theta_H g \text{ and } w < \beta(\theta_L, m_L); \\ (\theta_H, m_L, 1) \text{ or } (\theta_H, m_H, 1) & \text{if } I < e\theta_H g \text{ and } w < \beta(\theta_H, m_H); \\ (\theta_H, m_L, 1) & \text{if } I < e\theta_H g \text{ and } \beta(\theta_H, m_H) \leq w < \beta(\theta_H, m_L); \\ (\theta_L, m_L, 0) & \text{if } w \geq \beta(\theta_L, m_L). \end{cases} \tag{A.25}$$

The equilibrium $(\theta^{pre}(w), m^{pre}(w), d^{pre}(w))$ given in (A.25) is not unique when $I < e\theta_H g$ and $w < \beta(\theta_H, m_H)$. However, later analysis would show that at the contract stage (S1) the manufacturer would prefer to offer a wholesale price to induce $\{\theta_H, m_L, 1\}$ rather than $\{\theta_H, m_H, 1\}$ because the wholesale price for $\{\theta_H, m_L, 1\}$ is lower. Finally, similar to the proof of Proposition 1, we substitute the equilibrium outcome $(\theta^{pre}(w), m^{pre}(w), d^{pre}(w))$ given in (A.25) into the revised maximization program (2.12), and then solve for the subgame-perfect equilibrium in the pre-Act scenario under the new setting. The derivation is omitted here as it is a replication of the proof of Proposition 1 with one only change that replaces $\Delta(\theta_L)$ with $\Delta(\theta_L, m_L)$ and $\Delta(\theta_H)$ with $\Delta(\theta_H, m_L)$. ■ **Proof of Corollary 10**. Similar to the proof of Proposition 2, We first simplify the program (2.17), and then solve for the subgame-perfect equilibrium through three lemmas. First, First, substituting $d^{post}(\theta, m, w) = d^{pre}(\theta, m, w)$ given in (A.24) into (2.17) yields:

$$\max_{\theta, m} \begin{cases} U(s_H - \Delta(\theta), \theta, m, 1) = v - (s_H - \Delta(\theta, m)) - I(\theta) - e(1-\theta)g, \\ U\left(s_H + \frac{\Delta(\theta)}{\gamma e(1-\theta)} - \Delta(\theta), \theta, m, 0\right) = v - \left(s_H + \frac{\Delta(\theta, m)}{\gamma e(1-\theta)} - \Delta(\theta, m)\right) - I(\theta). \end{cases} \tag{A.26}$$

Since $\theta \in \{\theta_L, \theta_H\}$ and $m \in \{m_L, m_H\}$, the manufacturer needs to compare two functions: $U(s_H - \Delta(\theta, m), \theta, 1)$ and $U\left(s_H + \frac{\Delta(\theta)}{\gamma e(1-\theta)} - \Delta(\theta), \theta, m, 0\right)$ under four possible combinations of

$(\theta, m) \in \{\theta_L, \theta_H\} \times \{m_L, m_H\}$. Further analysis indicates that there are several redundant options that can be removed. First, since for $\theta = \theta_L$, the manufacturer is indifferent between the two values m_L and m_H because when there are no internal inspections there are no chances that the supplier needs to pay the compensation. Here for ease of exposition, we adopt the tie-breaking rule under which the manufacturer would choose $\{\theta_L, m_L\}$ rather than $\{\theta_L, m_H\}$. Second, since $\Delta(\theta_H, m_L) > \Delta(\theta_H, m_H)$, it is not difficult to infer that for $\theta = \theta_H$,

$$U(s_H - \Delta(\theta_H, m_L), \theta_H, m_L, 1) > U(s_H - \Delta(\theta_H, m_H), \theta_H, m_H, 1);$$
$$U\left(s_H + \frac{\Delta(\theta_H, m_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H, m_H), \theta_H, m_H, 0\right) > U\left(s_H + \frac{\Delta(\theta_H, m_L)}{\gamma e(1-\theta_H)} - \Delta(\theta_H, m_L), \theta_H, m_L, 0\right).$$

As a result, the manufacturer only needs to compare four functions: $U(s_H - \Delta(\theta_L, m_L), \theta_L, m_L, 1)$, $U(s_H - \Delta(\theta_H, m_L), \theta_H, m_L, 1)$, $U\left(s_H + \frac{\Delta(\theta_L, m_L)}{\gamma e(1-\theta_L)} - \Delta(\theta_L, m_L), \theta_L, m_L, 0\right)$, and $U\left(s_H + \frac{\Delta(\theta_H, m_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H, m_H), \theta_H, m_H, 0\right)$. Similar to the proof of Proposition 2, in the rest of the proof, we provide the remaining conditions under three possible structures of the equilibrium outcome that depend on the value of m. Lemmas 14, 15 and 16 together prove Corollary 10. The proofs of the three lemma are omitted and can be provided under request. For convenience, we rewrite assumptions (A2') and (A3'), respectively, as $m_L < m_H \leq \overline{m} \equiv \frac{(1-\theta_H)(s_H - s_L)}{\theta_H}$ and $m \geq \underline{m} \equiv \frac{\gamma e(1-\theta_H)(s_H - s_L)}{1 - \gamma e(1-\theta_H)}$.

Lemma 14 *The equilibrium outcome of program (A.26) is*

$$\left\{w_m^{post}, \theta_m^{post}, m_m^{post}, d_m^{post}\right\} =$$
$$\begin{cases} (s_H - \Delta(\theta_L, m_L), \theta_L, m_L, 1) & \text{if } g \leq \xi_{2,m}^{post}, I \geq \xi_{1,m}^{post}; \\ (s_H - \Delta(\theta_H, m_L), \theta_H, m_L, 1) & \text{if } I \leq \xi_{1,m}^{post}, I \leq \xi_{3,m}^{post}, g \leq \xi_{4,m}^{post}; \\ \left(s_H + \frac{\Delta(\theta_L, m_L)}{\gamma e} - \Delta(\theta_L, m_L), \theta_L, m_L, 0\right) & \text{if } g \geq \xi_{2,m}^{post}, I \geq \xi_{3,m}^{post}, I \geq \xi_{5,m}^{post}; \\ \left(s_H + \frac{\Delta(\theta_H, m_H)}{\gamma e(1-\theta_H)} - \Delta(\theta_H, m_H), \theta_H, m_H, 0\right) & \text{if } g \geq \xi_{4,m}^{post}, I \leq \xi_{5,m}^{post}. \end{cases}$$

when $\gamma e < \frac{1}{2 - \theta_H}$ *and* $b_1^m(m_L, m_H) < 0$, *where*

$$b_1^m(m_L, m_H) := \gamma e(1-\theta_H) m_L + (1 - \gamma e(1-\theta_H)) m_H - (1-\theta_H)(s_H - s_L).$$

Lemma 15 *The equilibrium outcome of program (A.26) is*

$$\left\{w_m^{post}, \theta_m^{post}, m_m^{post}, d_m^{post}\right\} =$$
$$\begin{cases} \left(s_H - \Delta\left(\theta_L, m_L\right), \theta_L, m_L, 1\right) & \text{if } g \leq \xi_{2,m}^{post}, I \geq \xi_{1,m}^{post}, I \geq \xi_{6,m}^{post}; \\ \left(s_H - \Delta\left(\theta_H, m_L\right), \theta_H, m_L, 1\right) & \text{if } I \leq \xi_{1,m}^{post}, g \leq \xi_{4,m}^{post}; \\ \left(s_H + \dfrac{\Delta\left(\theta_L, m_L\right)}{\gamma e} - \Delta\left(\theta_L, m_L\right), \theta_L, m_L, 0\right) & \text{if } g \geq \xi_{2,m}^{post}, I \geq \xi_{5,m}^{post}; \\ \left(s_H + \dfrac{\Delta\left(\theta_H, m_H\right)}{\gamma e\left(1-\theta_H\right)} - \Delta\left(\theta_H, m_H\right), \theta_H, m_H, 0\right) & \text{if } g \geq \xi_{4,m}^{post}, I \leq \xi_{5,m}^{post}, I \leq \xi_{6,m}^{post}. \end{cases}$$

when $b_1^m\left(m_L, m_H\right) \geq 0$, $b_2^m\left(m_L, m_H\right) \leq 0$, *and* $\gamma e < \dfrac{1}{2-\theta_H}$, *where*

$$b_2^m\left(m_L, m_H\right) := \gamma e\left(1-\theta_H\right) m_L + \theta_H\left(1-\gamma e\left(1-\theta_H\right)\right) m_H - \left(s_H - s_L\right)\left(1-\gamma e\left(1-\theta_H\right)\right)\left(1-\theta_H\right).$$

Lemma 16 *The equilibrium outcome of program (A.26) is*

$$\left\{w_m^{post}, \theta_m^{post}, m_m^{post}, d_m^{post}\right\} =$$
$$\begin{cases} \left(s_H - \Delta\left(\theta_L, m_L\right), \theta_L, m_L, 1\right) & \text{if } g \leq \xi_{2,m}^{post}, I \geq \xi_{6,m}^{post}; \\ \left(s_H + \dfrac{\Delta\left(\theta_L, m_L\right)}{\gamma e} - \Delta\left(\theta_L, m_L\right), \theta_L, m_L, 0\right) & \text{if } g \geq \xi_{2,m}^{post}, I \geq \xi_{5,m}^{post}; \\ \left(s_H + \dfrac{\Delta\left(\theta_H, m_H\right)}{\gamma e\left(1-\theta_H\right)} - \Delta\left(\theta_H, m_H\right), \theta_H, m_H, 0\right) & \text{if } I \leq \xi_{5,m}^{post}, I \leq \xi_{6,m}^{post}. \end{cases}$$

when $b_1^m\left(m_L, m_H\right) \geq 0$ *and* $b_2^m\left(m_L, m_H\right) > 0$. ■

Proof of Corollary 11. The proof is straightforward and thus is omitted here. **Proof of Corollary 12(a).** Similar to the derivation of the subgame-perfect equilibrium in the pre-Act scenario in the base setting, we will first derive the two players' best response, and then characterize Nash equilibrium in the subgame and finally the subgame-perfect equilibrium. First, based on the supplier's new profit function defined in (A.11), for any given w and θ the supplier's best response is:

$$\begin{aligned} &d_{ct}^{pre}\left(\theta, w\right) = 1 \ \forall \theta \in \{\theta_L, \theta_H\} \text{ for } w \in \left[0, s_H + \tfrac{\Delta(0)}{\gamma(\theta_H + (1-\theta_H)e)} - \Delta\left(0\right)\right); \\ &d_{ct}^{pre}\left(\theta_L, w\right) = 1 \text{ and } d_{ct}^{pre}\left(\theta_H, w\right) = 0 \text{ for } w \in \left[s_H + \tfrac{\Delta(0)}{\gamma(\theta_H + (1-\theta_H)e)} - \Delta\left(0\right), s_H + \tfrac{\Delta(0)}{\gamma e} - \Delta\left(0\right)\right); \\ &d^{pre}\left(\theta, w\right) = 0 \ \forall \theta \in \{\theta_L, \theta_H\} \text{ for } w \in \left[s_H + \tfrac{\Delta(0)}{\gamma e} - \Delta\left(0\right), +\infty\right). \end{aligned} \tag{A.27}$$

Second, since the manufacturer's best response function is not affected under the new penalty scheme, we can derive the Nash equilibrium as follows:

$$\left(\theta_{ct}^{pre}(w), d_{ct}^{pre}(w)\right) = \begin{cases} (\theta_L, 1) & \text{if } I \geq e\theta_H g \text{ and } w < s_H + \frac{\Delta(0)}{\gamma e} - \Delta(0); \\ (\theta_L, 0) & \text{if } w \geq s_H + \frac{\Delta(0)}{\gamma e} - \Delta(0); \\ (\theta_H, 1) & \text{if } I < e\theta_H g \text{ and } w < s_H + \frac{\Delta(0)}{\gamma(\theta_H + (1-\theta_H)e)} - \Delta(0). \end{cases}$$

Lastly, we can substitute the equilibrium outcome $\left(\theta_{ct}^{pre}(w), d_{ct}^{pre}(w)\right)$ given above to simplify the maximization program (2.12) and obtain the subgame-perfect equilibrium. The derivation is similar to the proof of Proposition 1 and thus is omitted here. ■ **Proof of Corollary 12(b).** Similar to the proof of Proposition 2, In the following, we first simplify the program (2.17), then solve for the subgame-perfect equilibrium through two lemmas.. First, substituting $d_{ct}^{post}(\theta, w) = d_{ct}^{pre}(\theta, w)$ given in (A.27) into (2.17) yields:

$$\max_{\theta} \begin{cases} U(s_H - \Delta(0), \theta, 1) = v - (s_H - \Delta(\theta)) - I(\theta) - e(1-\theta)g; \\ U\left(s_H + \frac{\Delta(0)}{\gamma(\theta + (1-\theta)e)} - \Delta(0), \theta, 0\right) = v - \left(s_H + \frac{\Delta(0)}{\gamma(\theta + (1-\theta)e)} - \Delta(0)\right) - I(\theta). \end{cases}$$

Second, we analyze the above program similar to that of (A.1) in the proof of Proposition 2. The only difference is that in the new penalty scheme, there are only two possible structures of the equilibrium outcome. For the sake of space, here we omit all the details but only present the three lemmas that provide the remaining conditions under the two possible structures of the equilibrium outcome.

Lemma 17 *If the manufacturer only terminates its contract with the supplier when child labor is detected in both internal and external inspections, the equilibrium outcome in the post-Act scenario is*

$$\left\{w_{ct}^{post}, \theta_{ct}^{post}, d_{ct}^{post}\right\} =$$
$$\begin{cases} (s_H - \Delta(0), \theta_L, 1) & \text{if } I \geq \xi_{1,ct}^{post}, g \leq \xi_{2,ct}^{post}; \\ (s_H - \Delta(0), \theta_H, 1) & \text{if } I \leq \xi_{1,ct}^{post}, I \leq \xi_{3,ct}^{post}, g \leq \xi_{4,ct}^{post}; \\ (s_H + (1/(\gamma e) - 1)\Delta(0), \theta_L, 0) & \text{if } g \geq \xi_{2,ct}^{post}, I \geq \xi_{3,ct}^{post}, I \geq \xi_{5,ct}^{post}; \\ (s_H + \{1/[\gamma(\theta_H + (1-\theta_H)e)] - 1\}\Delta(0), \theta_H, 0) & \text{if } g \geq \xi_{4,ct}^{post}, I \leq \xi_{5,ct}^{post}. \end{cases}$$

when $e > \frac{1-\theta_H}{2-\theta_H}$.

Lemma 18 *If the manufacturer only terminates its contract with the supplier when child labor is detected in both internal and external inspections, the equilibrium outcome in the post-Act scenario is*

$$\left\{w_{ct}^{post}, \theta_{ct}^{post}, d_{ct}^{post}\right\} =$$
$$\begin{cases} (s_H - \Delta(0), \theta_L, 1) & \text{if } I \geq \xi_{1,ct}^{post}, g \leq \xi_{2,ct}^{post}, , I \geq \xi_{6,ct}^{post}; \\ (s_H - \Delta(0), \theta_H, 1) & \text{if } I \leq \xi_{1,ct}^{post}, g \leq \xi_{4,ct}^{post}; \\ (s_H + (1/(\gamma e) - 1)\Delta(0), \theta_L, 0) & \text{if } g \geq \xi_{2,ct}^{post}, I \geq \xi_{5,ct}^{post}; \\ (s_H + \{1/[\gamma(\theta_H + (1-\theta_H)e)] - 1\}\Delta(0), \theta_H, 0) & \text{if } I \leq \xi_{6,ct}^{post}, g \geq \xi_{4,ct}^{post}, I \leq \xi_{5,ct}^{post}. \end{cases}$$

when $e \leq \frac{1-\theta_H}{2-\theta_H}$. ■

Appendix B

Supplements for Chapter 3

B.1 Proofs of Analytical Results

Proof of Proposition 4: See Righter and Shanthikumar (1998), which has proved that under any non-priority and non-preemptive scheduling, if service times are more variable in the convex sense, the waiting times are stochastically larger. Note that our service differentiation with holding the aggregate average service time is a special case of adding variability in the convex sense. **Proof of proposition 6**: To study the effects of adding one more service grade to the current mixed service with K grades, we choose to divide the first service grade with rate μ_1 into two new grades with rate μ_{1a} and μ_{1b} with allocation probability p_{1a} and p_{1b}, and we keep the other $K-1$ grades and corresponding allocation rules unchanged. Similar to proposition 2, we assume $\mu_{1a} > \mu_1 > \mu_{1b}$, $p_{1a} + p_{1b} = p_1$, $\frac{p_1}{\mu_1} = \frac{p_{1a}}{\mu_{1a}} + \frac{p_{1b}}{\mu_{1b}}$ and $\rho_1 = \rho_{1a} + \rho_{1b}$. Thus the newly designed mixed service with $K+1$ service grades has the same aggregate average service time as the original mixed service with K service grades, just as the proposition claimed. First of all, we compute the first and second moments of the service time. For the original service with K service grades, the service time is distributed as: $X_{(K)} = \begin{cases} \frac{\mu}{\mu_1} X_o \text{ w.p. } p_1 \\ ... \\ \frac{\mu}{\mu_K} X_o \text{ w.p. } p_K \end{cases}$, so its first moment is $E[X_{(K)}] = \sum_{k=1}^{K} \frac{p_k}{\mu_k}$ and the second moment is

$$E[X_{(K)}^2] = \mu^2 E[X_o^2] \sum_{k=1}^{K} \frac{p_k}{\mu_k^2}. \tag{B.1}$$

Similarly, the service time of the newly designed K+1-grade service is $X_{(K+1)} = \begin{cases} \frac{\mu}{\mu_{1a}} X_o \text{ w.p. } p_{1a} \\ \frac{\mu}{\mu_{1b}} X_o \text{ w.p. } p_{1b} \\ ... \\ \frac{\mu}{\mu_K} X_o \text{ w.p. } p_K \end{cases}$,

so $E[X_{(K+1)}] = \frac{p_{1a}}{\mu_{1a}} + \frac{p_{1b}}{\mu_{1b}} + \sum_{k=2}^{K} \frac{p_k}{\mu_k} = \sum_{k=1}^{K} \frac{p_k}{\mu_k}$, which is just equal to that of service with K service grades. The second moment of $X_{(K+1)}$ is: $E[X^2_{(K+1)}] = \mu^2 E[X_o^2] \left(\sum_{k=1}^{K} \frac{p_k}{\mu_k^2} + \frac{p_{1a}}{\mu_{1a}^2} + \frac{p_{1b}}{\mu_{1b}^2} - \frac{p_1}{\mu_1^2} \right)$.

As we have: $\frac{p_{1a}}{\mu_{1a}^2} + \frac{p_{1b}}{\mu_{1b}^2} - \frac{p_1}{\mu_1^2} = \frac{p_{1a}}{\mu_{1a}^2} + \frac{p_{1b}}{\mu_{1b}^2} - \frac{1}{p_1} \left(\frac{p_{1a}}{\mu_{1a}} + \frac{p_{1b}}{\mu_{1b}} \right)^2 = \frac{p_{1a} p_{1b}}{p_1} \left(\frac{1}{\mu_{1b}} - \frac{1}{\mu_{1a}} \right)^2$, we can express the second moment of $X_{(K+1)}$ as $E[X^2_{(K+1)}] = \mu^2 E[X_o^2] \left(\sum_{k=1}^{K} \frac{p_k}{\mu_k^2} + \frac{p_{1a} p_{1b}}{p_1} \left(\frac{1}{\mu_{1b}} - \frac{1}{\mu_{1a}} \right)^2 \right)$.

For convenience, we let $A = \sum_{k=1}^{K} \frac{p_k}{\mu_k^2} + \frac{p_{1a} p_{1b}}{p_1} \left(\frac{1}{\mu_{1b}} - \frac{1}{\mu_{1a}} \right)^2$, thus

$$E[X^2_{(K+1)}] = \mu^2 E[X_o^2] A. \tag{B.2}$$

Recall that the average waiting time in the kth service grade under priority scheduling rule is $E[W_k] = \frac{\lambda E[X^2]}{2(1 - \sum_{i=1}^{k} \rho_i)(1 - \sum_{i=1}^{k-1} \rho_i)}$. For the original case with total K service grades, the average waiting time in the kth grade is $E[W_k]_{(K)} = \frac{\lambda E[X^2_{(K)}]}{2(1 - \sum_{i=1}^{k} \rho_i)(1 - \sum_{i=1}^{k-1} \rho_i)}$. Thus the aggregate average waiting time of the first case is:

$$E[W]_{(K)} = \sum_{k=1}^{K} p_k E[W_k]_{(K)} = \frac{\lambda E[X^2_{(K)}]}{2} \sum_{k=1}^{K} \frac{p_k}{(1 - \sum_{i=1}^{k} \rho_i)(1 - \sum_{i=1}^{k-1} \rho_i)}.$$

For convenience, we let $B = \sum_{k=1}^{K} \frac{p_k}{(1 - \sum_{i=1}^{k} \rho_i)(1 - \sum_{i=1}^{k-1} \rho_i)}$ and we can find that B is a constant once the original K service grades are given. Finally we have

$$E[W]_{(K)} = \frac{\lambda E[X^2_{(K)}]}{2} B. \tag{B.3}$$

For the new service with $K+1$ service grades, the average waiting time in each service grade is distributed as:

$$
\begin{aligned}
k = 1a, E[w_{1a}]_{(K+1)} &= \frac{\lambda E[X^2_{(K+1)}]}{2(1-\rho_{1a})}; \\
k = 1b, E[w_{1b}]_{(K+1)} &= \frac{\lambda E[X^2_{(K+1)}]}{2(1-\rho_{1a})(1-\rho_1)}; \\
k \geq 2, E[w_k]_{(K+1)} &= \frac{\lambda E[X^2_{(K+1)}]}{2(1-\sum_{i=1}^{k}\rho_i)(1-\sum_{i=1}^{k-1}\rho_i)}.
\end{aligned}
$$

Thus the aggregate average waiting time with $K+1$ service grades is

$$
\begin{aligned}
E[W]_{(K+1)} &= p_{1a}E[W_{1a}]_{(K+1)} + p_{1b}E[W_{1b}]_{(K+1)} + \sum_{k=2}^{K} p_k E[W_k]_{(K+1)} \\
&= \frac{\lambda E[X^2_{(K+1)}]}{2}\left(\frac{p_{1a}}{1-\rho_{1a}} + \frac{p_{1b}}{(1-\rho_{1a})(1-\rho_1)} - \frac{p_1}{1-\rho_1} + B\right).
\end{aligned}
$$

As $\dfrac{p_{1a}}{1-\rho_{1a}} + \dfrac{p_{1b}}{(1-\rho_{1a})(1-\rho_1)} - \dfrac{p_1}{1-\rho_1} = \dfrac{-p_{1a}p_{1b}\lambda\left(\dfrac{1}{\mu_{1b}} - \dfrac{1}{\mu_{1a}}\right)}{(1-\rho_{1a})(1-\rho_1)}$, finally we have:

$$
E[W]_{(K+1)} = \frac{-\lambda E[X^2_{(K+1)}]}{2}\frac{p_{1a}p_{1b}\lambda\left(\dfrac{1}{\mu_{1b}} - \dfrac{1}{\mu_{1a}}\right)}{(1-\rho_{1a})(1-\rho_1)} + \frac{\lambda E[X^2_{(K+1)}]}{2}B. \tag{B.4}
$$

To compare the average waiting time of newly designed mixed service with $K+1$ service grades with that of K service grades, we get the difference between them as:

$$
E[W]_{(K+1)} - E[W]_{(K)} = \frac{\lambda\left(E[X^2_{(K+1)}] - E[X^2_{(K)}]\right)}{2}B - \frac{\lambda E[X^2_{(K+1)}]}{2}\frac{p_{1a}p_{1b}\lambda\left(\dfrac{1}{\mu_{1b}} - \dfrac{1}{\mu_{1a}}\right)}{(1-\rho_{1a})(1-\rho_1)}. \tag{B.5}
$$

Based on (B.1) and (B.2), we substitute $E[X^2_{(K+1)}]$ and $E[X^2_{(K)}]$ into (B.5) and we have:

$$E[W]_{(K+1)} - E[W]_{(K)} = \frac{\lambda}{2}\left(\mu^2 E[X_o^2]\left(A - \sum_{k=1}^{K}\frac{p_k}{\mu_k^2}\right)\right)B - \frac{\lambda\mu^2 E[X_o^2]A}{2}\frac{p_{1a}p_{1b}\lambda\left(\frac{1}{\mu_{1b}} - \frac{1}{\mu_{1a}}\right)}{(1-\rho_{1a})(1-\rho_1)}$$
$$= \frac{\lambda}{2}\mu^2 E[X_o^2]p_{1a}p_{1b}\left(\frac{1}{\mu_{1b}} - \frac{1}{\mu_{1a}}\right)\left(\frac{B}{p_1}\left(\frac{1}{\mu_{1b}} - \frac{1}{\mu_{1a}}\right) - \frac{A\lambda}{(1-\rho_{1a})(1-\rho_1)}\right).$$

As we have

$$\frac{\lambda}{2}\mu^2 E[X_o^2]p_{1a}p_{1b}\left(\frac{1}{\mu_{1b}} - \frac{1}{\mu_{1a}}\right) > 0,$$

we can claim that the mixed service with $K+1$ service grades generates a lower waiting time than the mixed service with K grades if and only if the newly differentiated service rates μ_{1a} and μ_{1b} satisfies the below inequality:

$$\frac{B}{p_1}\left(\frac{1}{\mu_{1b}} - \frac{1}{\mu_{1a}}\right) < \frac{A\lambda}{(1-\rho_{1a})(1-\rho_1)}, \tag{B.6}$$

where $A = \left(\sum_{k=1}^{K}\frac{p_k}{\mu_k^2} + \frac{p_{1a}p_{1b}}{p_1}\left(\frac{1}{\mu_{1b}} - \frac{1}{\mu_{1a}}\right)^2\right)$ and $B = \sum_{k=1}^{K}\frac{p_k}{(1-\sum_{i=1}^{k}\rho_i)(1-\sum_{i=1}^{k-1}\rho_i)}$. The inequality (B.6) gives the sufficient and necessary condition for Proposition 6, which also suggests the following sufficient condition as:

$$\frac{\mu_{1a}}{\mu_{1b}} < 1 + \frac{\rho_1}{1-\rho_1}\Big/B\,. \tag{B.7}$$

Next we prove that (B.7) is a sufficient condition for (B.6). As $\frac{\rho_1}{1-\rho_1} = \frac{\lambda p_1}{(1-\rho_1)\mu_1}$ and $\frac{\mu_{1a}}{\mu_{1b}} - 1 = \mu_{1a}\left(\frac{1}{\mu_{1b}} - \frac{1}{\mu_{1a}}\right)$, we can transform the necessary condition $(B.7)$ as:

$$(B.7) \Leftrightarrow \mu_{1a}\left(\frac{1}{\mu_{1b}} - \frac{1}{\mu_{1a}}\right) < \frac{\lambda p_1}{(1-\rho_1)\mu_1}\Big/B$$
$$\Leftrightarrow \frac{B}{p_1}\left(\frac{1}{\mu_{1b}} - \frac{1}{\mu_{1a}}\right) < \frac{\lambda}{(1-\rho_1)\mu_1\mu_{1a}}.$$

Based on $0 < 1-\rho_{1a} < 1 \Rightarrow 1 < \frac{1}{1-\rho_{1a}}$ and $\mu_{1a} > \mu_1 \Rightarrow \frac{1}{\mu_{1a}} < \frac{1}{\mu_1}$, we can increase the positive right hand side of the above inequality as $\frac{\lambda}{(1-\rho_1)\mu_1\mu_{1a}} < \frac{\lambda}{(1-\rho_1)\mu_1\mu_{1a}}\frac{1}{1-\rho_{1a}} <$

$\frac{\lambda}{(1-\rho_{1a})(1-\rho_1)}\frac{1}{\mu_1^2}$. Now we have:

$$(B.7) \Rightarrow \frac{B}{p_1}\left(\frac{1}{\mu_{1b}}-\frac{1}{\mu_{1a}}\right) < \frac{\lambda}{(1-\rho_{1a})(1-\rho_1)}\frac{1}{\mu_1^2}.$$

Furthermore as $\mu_1 > \mu_2 > ... > \mu_K > 0 \Rightarrow \frac{1}{\mu_1^2} < \frac{1}{\mu_2^2} < ... < \frac{1}{\mu_K^2}$, and $\sum_{k=1}^K p_k = 1$, we have $\frac{1}{\mu_1^2} = \sum_{k=1}^K \frac{p_k}{\mu_1^2} < \sum_{k=1}^K \frac{p_k}{\mu_k^2}$. So:

$$(B.7) \Rightarrow \frac{B}{p_1}\left(\frac{1}{\mu_{1b}}-\frac{1}{\mu_{1a}}\right) < \frac{\lambda}{(1-\rho_{1a})(1-\rho_1)}\left(\sum_{k=1}^K \frac{p_k}{\mu_k^2}\right).$$

As $\frac{p_{1a}p_{1b}}{p_1}\left(\frac{1}{\mu_{1b}}-\frac{1}{\mu_{1a}}\right)^2 > 0$, it is easy to derive that $\sum_{k=1}^K \frac{p_k}{\mu_k^2} < A$. Finally we get:

$$(B.7) \Rightarrow \frac{B}{p_1}\left(\frac{1}{\mu_{1b}}-\frac{1}{\mu_{1a}}\right) < \frac{\lambda}{(1-\rho_{1a})(1-\rho_1)}A,$$

indicating that (B.7) is a sufficient condition for Proposition 6. ■ **Proof of Lemma 2**: There are multiple methods to solve the problem (3.19). One straightforward approach applies the Cauchy–Schwarz inequality, which provides a lower bound of the objective function in (3.19) under the condition that $0 < p_k < 1$ for all $k = 1, ...K$:

$$\begin{aligned}
&\left(\sum_{k=1}^K \frac{p_k}{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}\right)\left(\sum_{k=1}^K \frac{\rho_k^2}{p_k}\right) \\
&= \left(\sum_{k=1}^K \left(\sqrt{\frac{p_k}{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}}\right)^2\right)\left(\sum_{k=1}^K \left(\sqrt{\frac{\rho_k^2}{p_k}}\right)^2\right) \\
&\geq \left(\sum_{k=1}^K \sqrt{\frac{p_k}{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}}\sqrt{\frac{\rho_k^2}{p_k}}\right)^2 \qquad \text{(B.8)} \\
&= \left(\sum_{k=1}^K \frac{\rho_k}{\sqrt{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}}\right)^2 \qquad \text{(B.9)}
\end{aligned}$$

Based on the Cauchy–Schwarz inequality, the inequality (B.8) holds on the condition that:

$$\sqrt{\frac{p_k}{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}} : \sqrt{\frac{\rho_k^2}{p_k}} = \text{constant}$$

$$\Leftrightarrow \frac{p_k}{\rho_k\sqrt{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}} = \text{constant} \quad \text{for } k = 1, ...K. \tag{B.10}$$

Note that the lower bound in (B.8) is a function which only contains the decision variable $\overrightarrow{\rho}$, in other words, a constant term independent of the decision variables $\overrightarrow{p}$. This means that (B.9) is the minimal value that the objective function can achieve for any positive vector $\overrightarrow{p}$. Once there exists a vector $\overrightarrow{p}$ with positive elements which satisfies both the constraints (B.10) and $\sum_{k=1}^{K} p_k = 1$, such a vector $\overrightarrow{p}$ is the optimal solution of the minimization problem. It is easy to verify that the optimal $\overrightarrow{p}$ is

$$p_k = \frac{\rho_k\sqrt{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}}{\sum_{k=1}^{K}\rho_k\sqrt{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}} \quad \text{for } k = 1, ...K \tag{B.11}$$

because the vector $\overrightarrow{p}$ defined in (B.11) satisfies:

$$\frac{p_k}{\rho_k\sqrt{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}} = \sum_{k=1}^{K}\rho_k\sqrt{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)} \quad \text{for } k = 1, ...K,$$

and

$$\sum_{k=1}^{K} p_k = \sum_{k=1}^{K} \frac{\rho_k\sqrt{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}}{\sum_{k=1}^{K}\rho_k\sqrt{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}} = 1,$$

where the first equation validates the condition (B.10) and the second satisfies the constraint $\sum_{k=1}^{K} p_k = 1$. ■ **Proof of Corollary 2**: To prove that $p_k E[W_k] = p_{k+1}E[W_{k+1}]$ for $k = 1, ...K-1$, referring to (3.4), it is equivalent to prove that

$$\frac{p_k}{(1-\sum_{i=1}^{k}\rho_i)(1-\sum_{i=1}^{k-1}\rho_i)} = \frac{p_{k+1}}{(1-\sum_{i=1}^{k+1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}.$$

Together with (B.11), the objective can be written in terms of $\{\rho_k\}$ as:

$$\frac{\rho_k}{\sqrt{\left(1-\sum_{i=1}^{k-1}\rho_i\right)\left(1-\sum_{i=1}^{k}\rho_i\right)}} = \frac{\rho_{k+1}}{\sqrt{\left(1-\sum_{i=1}^{k}\rho_i\right)\left(1-\sum_{i=1}^{k+1}\rho_i\right)}}. \tag{B.12}$$

The chain rule given in Lemma 3 suggests that

$$\rho_{k+1} = \frac{\left(1-\sum_{i=1}^{k}\rho_i\right)\rho_k}{1-\sum_{i=1}^{k-1}\rho_i}$$

which is equivalent to

$$\frac{\rho_{k+1}}{\sqrt{\left(1-\sum_{i=1}^{k}\rho_i\right)\left(1-\sum_{i=1}^{k+1}\rho_i\right)}} = \frac{\rho_k}{\sqrt{\left(1-\sum_{i=1}^{k-1}\rho_i\right)\left(1-\sum_{i=1}^{k}\rho_i\right)}}.$$

■ **Proof of Corollary 3**: First, by induction on k, we prove that based on the chain rule given in Lemma 3 all the grades' loads $\{\rho_1, \rho_2, ...\rho_K\}$ can be derived in terms of ρ_1:

$$\rho_k = (1-\rho_1)^{k-1}\rho_1 \quad \text{for } k = 1, ...K.$$

When $k = 1$, we have:

$$(1-\rho_1)^2 = 1-\rho_1-\rho_2$$
$$\Leftrightarrow \rho_2 = (1-\rho_1) - (1-\rho_1)^2 = \rho_1(1-\rho_1).$$

When it is k, as

$$\sum_{i=1}^{k}\rho_i = \sum_{i=1}^{k}(1-\rho_1)^{i-1}\rho_1 = \rho_1\frac{1-(1-\rho_1)^k}{1-(1-\rho_1)} = 1-(1-\rho_1)^k, \tag{B.13}$$

with the chain rule for k:

$$\begin{aligned}\left(1-\sum_{i=1}^{k}\rho_i\right)^2 &= \left(1-\sum_{i=1}^{k+1}\rho_i\right)\left(1-\sum_{i=1}^{k-1}\rho_i\right)\\ \Rightarrow (1-\rho_1)^{2k} &= \left((1-\rho_1)^k-\rho_{k+1}\right)(1-\rho_1)^{k-1}\\ \Rightarrow \rho_{k+1} &= (1-\rho_1)^k-(1-\rho_1)^{k+1}=(1-\rho_1)^k\rho_1. \end{aligned} \tag{B.14}$$

Secondly we express ρ_1 with ρ together with the summation constraint:

$$\sum_{i=1}^{K}\rho_i = 1-(1-\rho_1)^K=\rho \Leftrightarrow \rho_1 = 1-(1-\rho)^{\frac{1}{K}}. \tag{B.15}$$

A substitution of (B.15) into (B.14) proves our objective. ■ **Proof of Proposition 7**: From Corollary 3 we have derived the optimal load vector $\overrightarrow{\rho}$, with which we can prove the geometric structure of $\{\rho_k\}$ and then extend it to both series $\{p_k\}$ and $\{\mu_k\}$. Now we have

$$\rho_k=(1-\rho_1)^{k-1}\rho_1=(1-\rho)^{\frac{k-1}{K}}\left(1-(1-\rho)^{\frac{1}{K}}\right) \quad \text{for } k=1,...K \tag{B.16}$$

which indicates the geometric structure of $\{\rho_k\}$:

$$\frac{\rho_{k+1}}{\rho_k}=\frac{(1-\rho)^{\frac{k}{K}}\left(1-(1-\rho)^{\frac{1}{K}}\right)}{(1-\rho)^{\frac{k-1}{K}}\left(1-(1-\rho)^{\frac{1}{K}}\right)}=(1-\rho)^{\frac{1}{K}} \quad \text{for } k=1,...K.$$

Next we show the geometric structure of $\{p_k\}$ based on ρ_k given in (B.16) and p_k in (B.11):

$$\begin{aligned}\frac{p_{k+1}}{p_k} &= \frac{\rho_{k+1}\sqrt{(1-\sum_{i=1}^{k}\rho_i)(1-\sum_{i=1}^{k+1}\rho_i)}}{\rho_k\sqrt{(1-\sum_{i=1}^{k-1}\rho_i)(1-\sum_{i=1}^{k}\rho_i)}}\\ &= (1-\rho)^{\frac{1}{K}}\sqrt{\frac{(1-\rho_1)^{k+1}}{(1-\rho_1)^{k-1}}} \quad \text{(with (B.13))}\\ &= (1-\rho)^{\frac{2}{K}} \qquad \text{for } k=1,...K.\end{aligned}$$

Finally with the definition of $\mu_k = \lambda p_k/\rho_k$, it is trivial to prove the geometric structure of $\{\mu_k\}$:

$$\frac{\mu_{k+1}}{\mu_k} = \frac{\lambda p_{k+1}/\rho_{k+1}}{\lambda p_k/\rho_k} = \frac{p_{k+1}}{p_k}\frac{\rho_k}{\rho_{k+1}} = (1-\rho)^{\frac{2}{K}}(1-\rho)^{-\frac{1}{K}} = (1-\rho)^{\frac{1}{K}} \qquad \text{for } k = 1, ...K.$$

■ **Proof of Proposition 8**: For convenience we define the first order derivative of the function (3.26) to ρ as a function $\theta(\rho) : (0,1) \to \mathbb{R}$ where

$$\begin{aligned}\theta(\rho) = \frac{\partial \pi}{\partial \rho} &= u - CK^2\left(\frac{1}{K}(1-\rho)^{-\frac{1}{K}-1} - \frac{1}{K}(1-\rho)^{\frac{1}{K}-1}\right) \\ &= u + CK\left((1-\rho)^{\frac{1}{K}-1} - (1-\rho)^{-\frac{1}{K}-1}\right).\end{aligned}$$

Next we will prove that ρ^* is the optimal solution of the problem (3.26) if and only if it satisfies $\theta(\rho^*) = 0$ by checking the second-order condition of the objective function:

$$\begin{aligned}\frac{\partial^2 \pi}{\partial \rho^2} = \frac{\partial \theta(\rho)}{\partial \rho} &= CK\left(\left(-\frac{1}{K} - 1\right)(1-\rho)^{-\frac{1}{K}-2} - \left(\frac{1}{K} - 1\right)(1-\rho)^{\frac{1}{K}-2}\right) \\ &= -C(K+1)(1-\rho)^{-\frac{1}{K}-2} - C(K-1)(1-\rho)^{\frac{1}{K}-2}.\end{aligned}$$

As $K \geq 1$ and $C > 0$, the second derivative $\dfrac{\partial^2 \pi}{\partial \rho^2} < 0$ holds for any $0 < \rho < 1$. After the optimality is proved, it is left to show that such ρ^* must exist and is unique in the interval $(0,1)$. As we have proved, $\theta(\rho)$ is a strictly decreasing continuous function of ρ. Moreover, we have $\theta(0) = u > 0$ and $\lim_{\rho \to 1}\theta(\rho) < 0$ as

$$\begin{aligned}\lim_{\rho\to 1}\theta(\rho) &= u + CK\lim_{\rho\to 1}\left((1-\rho)^{\frac{1}{K}-1} - (1-\rho)^{-\frac{1}{K}-1}\right) \\ &= u + CK\lim_{\rho\to 1}(1-\rho)^{-\frac{1}{K}-1}\left((1-\rho)^{\frac{2}{K}} - 1\right)\end{aligned}$$

and

$$(1-\rho)^{\frac{2}{K}} - 1 < 0 \text{ but } \lim_{\rho\to 1}(1-\rho)^{-\frac{1}{K}-1} \to \infty.$$

Finally the existence and uniqueness of β^* is proved based on the intermediate value theorem (Rutin 1964).■ **Proof of the claim in Section 3.4.4 that the index $\widetilde{R_V}$ depends on the external parameters u, h and $\mu_o^2 E[X_o^2]$ only through the term $\tau := \dfrac{u}{h\mu_o^2 E[X_o^2]}$.** Since $\widetilde{R_V} = \widetilde{V}/V_{(1)}$,

according to (3.31) and (3.28),

$$\widetilde{R_V} = \frac{2\tau\widetilde{\rho} - (\ln(1-\widetilde{\rho}))^2}{2\tau\rho_{(1)} - \left((1-\rho_{(1)})^{-1} - \rho_{(1)} - 1\right)}, \tag{B.17}$$

where $\tau := \dfrac{u}{h\mu_o^2 E[X_o^2]}$, $\widetilde{\rho}$ is defined in (3.29), and $\rho_{(1)}$ is the optimal system load with $K = 1$ satisfying

$$u(1-\rho) + \frac{h\mu_o^2 E[X_o^2]}{2}\left(1 - \rho - \frac{1}{1-\rho}\right) = 0. \tag{B.18}$$

(3.29) and (B.18) suggest that $\widetilde{\rho}$ and $\rho_{(1)}$ are determined by the value of $\dfrac{u}{h\mu_o^2 E[X_o^2]} = \tau$. Therefore we claim that the asymptotic dominance index $\widetilde{R_V}$ depends on τ alone. Furthermore, based on the envelope theorem, we can derive the first order condition of $\widetilde{R_V}$ with respect to τ:

$$\frac{d\widetilde{R_V}}{d\tau} = 0 \Leftrightarrow \widetilde{\rho}V_{(1)} = \rho_{(1)}\widetilde{V},$$

which has a unique positive root, as shown by numerical experiments. ■ **Proof of the claim in Section 3.5 that the grade performance increases with the grade's priority decreasing**: Before analyzing the grade performance, let's first derive the close forms of the grade metrics: p_k, ρ_k and μ_k. According to Corollary (3),

$$\rho_k = (1-\rho)^{\frac{k-1}{K}}\left(1 - (1-\rho)^{\frac{1}{K}}\right).$$

Based on the geometric sequence structure of p_k and the constraint that $\sum_{k=1}^{K} p_k = 1$, it is easy to derive that

$$p_k = \frac{(1-\rho)^{\frac{2(k-1)}{K}}\left(1 - (1-\rho)^{\frac{2}{K}}\right)}{\left(1 - (1-\rho)^{\frac{2(K+1)}{K}}\right)}$$

and correspondingly

$$\begin{aligned}\mu_k &= \frac{\lambda p_k}{\rho_k} = \lambda\frac{(1-\rho)^{\frac{2(k-1)}{K}}\left(1 - (1-\rho)^{\frac{2}{K}}\right)}{(1-\rho)^{\frac{k-1}{K}}\left(1 - (1-\rho)^{\frac{2(K+1)}{K}}\right)\left(1 - (1-\rho)^{\frac{1}{K}}\right)} \\ &= \lambda\frac{(1-\rho)^{\frac{k-1}{K}}\left(1 + (1-\rho)^{\frac{1}{K}}\right)}{1 - (1-\rho)^{\frac{2(K+1)}{K}}}.\end{aligned}$$

It is trivial to verify that

$$\frac{p_k}{\mu_k^2} = \frac{\left(1+(1-\rho)^{\frac{1}{K}}\right)\left(1-(1-\rho)^{\frac{2(K+1)}{K}}\right)}{\lambda^2\left(1+(1-\rho)^{\frac{1}{K}}\right)}$$

which is a constant and independent of k; and

$$\begin{aligned}
1-\sum_{i=1}^{k}\rho_i &= 1-\rho_1\frac{1-(1-\rho)^{\frac{k}{K}}}{1-(1-\rho)^{\frac{1}{K}}} \\
&= 1-\left(1-(1-\rho)^{\frac{1}{K}}\right)\frac{1-(1-\rho)^{\frac{k}{K}}}{1-(1-\rho)^{\frac{1}{K}}} = (1-\rho)^{\frac{k}{K}}
\end{aligned}$$

Based on the above formulae, together with the grade average waiting time defined in (3.4), now we can derive the performance of grade k

$$\begin{aligned}
V_k &= \frac{u}{\mu_k} - hw_k \\
&= \frac{u}{\mu_k} - \frac{h\lambda\mu_o^2 E[X_o^2]\left(\sum_{k=1}^{K}\frac{p_k}{\mu_k^2}\right)}{2(1-\sum_{i=1}^{k}\rho_i)(1-\sum_{i=1}^{k-1}\rho_i)} \\
&= \frac{u}{\mu_k} - \frac{h\mu_o^2 E[X_o^2]}{2}\frac{K\frac{\lambda p_k}{\mu_k^2}}{(1-\sum_{i=1}^{k}\rho_i)(1-\sum_{i=1}^{k-1}\rho_i)} \\
&= \frac{1}{\mu_k}\left(u - \frac{h\mu_o^2 E[X_o^2]}{2}K\rho_k\,(1-\rho)^{-\frac{2k-1}{K}}\right) \\
&= \frac{1}{\mu_k}\left(u - \frac{h\mu_o^2 E[X_o^2]}{2}K\,(1-\rho)^{-\frac{k}{K}}\left(1-(1-\rho)^{\frac{1}{K}}\right)\right).
\end{aligned}$$

Similarly,

$$V_{k+1} = \frac{1}{\mu_{k+1}}\left(u - \frac{h\mu_o^2 E[X_o^2]}{2}K\,(1-\rho)^{-\frac{k+1}{K}}\left(1-(1-\rho)^{\frac{1}{K}}\right)\right).$$

Thus

$$\begin{aligned}
\frac{V_{k+1}}{V_k} &= \frac{\mu_k}{\mu_{k+1}} \frac{\frac{2u}{h\mu_o^2 E[X_o^2]} - K(1-\rho)^{-\frac{k}{K}}\left(1-(1-\rho)^{\frac{1}{K}}\right)}{\frac{2u}{h\mu_o^2 E[X_o^2]} - K(1-\rho)^{-\frac{k+1}{K}}\left(1-(1-\rho)^{\frac{1}{K}}\right)} \\
&= (1-\rho)^{-\frac{1}{K}} \frac{\frac{2u}{h\mu_o^2 E[X_o^2]} - K(1-\rho)^{-\frac{k}{K}}\left(1-(1-\rho)^{\frac{1}{K}}\right)}{\frac{2u}{h\mu_o^2 E[X_o^2]} - K(1-\rho)^{-\frac{k+1}{K}}\left(1-(1-\rho)^{\frac{1}{K}}\right)} \\
&= \frac{\frac{2u(1-\rho)^{-\frac{1}{K}}}{h\mu_o^2 E[X_o^2]} - K(1-\rho)^{-\frac{k+1}{K}}\left(1-(1-\rho)^{\frac{1}{K}}\right)}{\frac{2u}{h\mu_o^2 E[X_o^2]} - K(1-\rho)^{-\frac{k+1}{K}}\left(1-(1-\rho)^{\frac{1}{K}}\right)}.
\end{aligned}$$

Since $(1-\rho)^{-\frac{1}{K}} > 1$, and the denominator and the numerator are both positive, thus $V_{k+1} > V_k$ is proved. ■

B.2 Variance Benefits Illustration

This Section prevents several examples to illustrate the claims and/or statements in Section 3.6.1

Example I: An illustration of information value, variance effect and variance benefit. Table B.21 presents examples of three types of services: uniform service, a two-grade mixed service that does not use any service rate information and applies FCFS scheduling, and a second two-grade mixed service that uses information and applies SEPT. The three service polices have the same aggregate average service time and thus the same service value, so we can compare their performances directly through the average waiting time. First, we compare uniform service with the first mixed service (the second and the third rows), both of which make no use of service rate information. We can observe that the waiting time of mixed service, 0.52, is larger than that of uniform service, 0.5, due to an increment of service time variance (from 0.25 to 0.27). The increased waiting of 0.52-0.5=0.02 represents the variance effect. Secondly, by comparing the two mixed services (the third and fourth rows), we can observe that despite the same service rates and probabilities, the SEPT rule help reduce the waiting from 0.52 to 0.4875, generating a reduction of 0.0325, representing the information value. With an increment of 0.02 and a reduction of 0.0325, finally we obtain a surplus, of 0.0325-0.02=0.0125, which implies a 2.5% reduction of the original

waiting time 0.5. This is the benefit of added variance, the net of the information value and the variance effect. **Example II: An illustration of the claim that information value may**

Table B.21: An illustration of information value, variance effect and variance benefit

	Service rates	Allocation probability	Average service time	Service time variance	Average waiting time
Uniform Service	μ_o= 2	$p = 1$	0.5	0.25	0.5
Mixed Service Under FCFS	μ_1= 2.5 μ_2= 1.67	p_1= 0.5 p_2= 0.5	0.5	0.27	0.52
Mixed Service Under SEPT	μ_1= 2.5 μ_2= 1.67	p_1= 0.5 p_2= 0.5	0.5	0.27	0.4875

increase or decrease with the number of service grades with the same service time variance. Table B.22 compares a two-grade service (Mixed service I) with two types of three-grade service (Mixed service II and III), which have the same service time variance but generates more (as 0.0682>0.0532) and less (as 0.0491<0.0532) information value than two-grade service, respectively. **Example III: An illustration of the claim that information value may**

Table B.22: A comparison of information value with respect to different number of service grades with the same service time variance

	Service rates	Allocation probability	Average service time	Service time variance	Average waiting time without SEPT	Average waiting time with SEPT	Informatio value
Mixed Service I with two grades	μ_1= 2.5 μ_2= 1.36	p_1= 0.7 p_2= 0.3	0.5	0.297	0.5467	0.4935	0.0532
Mixed Service II with three grades	μ_1= 3.3 μ_2= 2 μ_3= 1.43	p_1= 0.3 p_2= 0.4 p_3= 0.3	0.5	0.297	0.5467	0.4785	0.0682
Mixed Service III with three grades	μ_1= 2.6 μ_2= 2 μ_3= 1.26	p_1= 0.49 p_2= 0.31 p_3= 0.20	0.5	0.297	0.5467	0.4976	0.0491

increase or decrease with service time variance. Table B.23 compares a two-grade service with a three-grade service, showing that though the latter has a higher service time variance, it generates less information value than the former one.

Table B.23: A comparison of information value with respect to different number of service grades with the same service time variance

	Service rates	Allocation probability	Average service time	Service time variance	Average waiting time without SEPT	Average waiting time with SEPT	Informatic value
Mixed Service with two grades	$\mu_1 = 2.15$ $\mu_2 = 1.82$	$p_1 = 0.58$ $p_2 = 0.342$	0.5	0.253	0.503	0.489	0.014
Mixed Service with three grades	$\mu_1 = 2.76$ $\mu_2 = 2$ $\mu_3 = 1.57$	$p_1 = 0.3$ $p_2 = 0.4$ $p_3 = 0.3$	0.5	0.273	0.523	0.512	0.011

B.3 Detailed Extensions and Future Work

This section describes details of the extensions discussed in Section 3.6.2.

1 Nonlinear service value and delay cost

Though we adopt linear functions in our model to represent service value and delay cost, these two functions may not in actuality be linear. If we generalize these definitions to permit either concavity or convexity, we are interested in two questions: whether the dominance of mixed service still holds, and if it does how this dominance compares with that under linear forms. One critical difference between linear functions and non-linear ones is that under non-linear settings the expected service value cannot be computed through the expected service time, instead it is a function of the distribution of service time X, i.e. $E[G(X)] \neq G[E(X)]$. Likewise the expected delay cost depends on the distribution of average waiting time. The former difference poses only a slight problem, as using the distribution of the differentiated service time shown in (3.1), it is trivial to compute $E[G(X)]$. However, in general, the distribution of the random waiting time is intractable, so extension to nonlinear waiting cost is complex. Therefore we focus our discussions on nonlinear service value and linear waiting cost. Under linear waiting cost, the SEPT scheduling rule maintains its optimality, and the conclusion in Corollary 1 remains true. Thus we can conclude that under this new setting mixed service can have a lower waiting cost than pure service when both have the same aggregate average service time. We still need to compare the service value generated by both systems with the same aggregate average service time. It is not difficult to prove that for a convex service value function, the expected service value of mixed service exceeds that of pure service under the same aggregate average service time, due to the

definition of convexity. Thus we know that it is more beneficial to introduce service rate differentiation under convex service value functions than under linear or concave ones. In contrast, concave service value functions may generate both lower service value and smaller waiting cost, which complicates our analysis. We can show the existence of the dominance of mixed service over pure service through an example as shown in Table B.34, in which the service value function is $V(X) = \sqrt{X}$ and the original job type is exponentially distributed, i.e. $X_o \sim \exp(\mu_o)$. Table B.34 shows that the optimal mixed service with two differentiated service rates generates a higher system utility than the optimal pure service as it is shown in the row for "System utility" that $2.6434 > 2.6377$. In conclusion, dominance can be extended to convex value and linear waiting costs; concave value, or nonlinear waiting costs require more study.

Table B.34: An illustration of the dominance of mixed service over pure service under a concave service value function $V(X) = \sqrt{X}$ (Other parameters: $\lambda = 1$, $u/h = 5$, and the original job type is exponentially distributed)

	Pure service	Mixed service
Optimal service rates	$\mu_p = 1.973$	$\mu_m^1 = 2.07 \quad \mu_m^2 = 1.681$ $p_m^1 = 0.71 \quad p_m^2 = 0.29$
System utility	2.6337	2.6434
Service value	3.1547	3.1776
Delay cost	0.521	0.5342

2 Preemptive scheduling

In our analysis we constrain our optimal scheduling rule to be non-preemptive, but it may also be possible to adopt a preemptive scheduling rule, which is more flexible, and particularly useful when dealing with jobs with high variance. Basic queueing theory tells us that the optimal preemptive scheduling rule is always to process a job with the shortest expected remaining processing time first (SERPT) (Harchol-Balter 2013). However, the SERPT rule is far too complicated to study in the general case. If we constrain our assumption to exponential service times and allow the possibility of resuming a preempted job instead of restarting it, we can conclude that the preemptive SERPT rule acts the same as the preemptive SEPT rule, which makes the problem tractable. With the preemptive SEPT rule, jobs that are assigned to the service grade with higher service rate experience even greater processing priorities: when a high priority job arrives, the server, though it may be processing a job in lower priority, will immediately stop the current work and begin serving the new arrival. Thus the preemptive rule saves more average waiting time for the smaller jobs. We conjecture that the dominance of mixed service over pure service continues to hold in general under some mild conditions; we can prove that for exponentially distributed jobs, any service rate differentiation with two service grades can

always generate a shorter average waiting time than pure service with the same aggregate average service time, which is summarized in the following proposition.

Proposition 10 *Under the preemptive SERPT scheduling rule, for any given arrival rate λ and any originally exponentially distributed jobs, a mixed service policy with two service grades can always generate a shorter average waiting time than pure service with the same aggregate average service time*

Proof. Proof: Because we assume mixed service has the same expected service time as pure service, the following equation holds:

$$\frac{1}{\mu} = \frac{p_1}{\mu_1} + \frac{p_2}{\mu_2}. \tag{B.19}$$

First we compute the average waiting time for pure service, and then for mixed service. Finally we prove the dominance of mixed service by comparing these two average waiting times. As both μ and λ are fixed, the system load ρ is also fixed as $\rho = \lambda/\mu$. When $K = 1$, the service time r.v. $X_{(1)} = X_o$, thus $E[X_{(1)}^2] = E[X_o^2]$ and $\rho_{(1)} = \rho$. So we have

$$E[w]_{(1)} = \frac{\lambda E[X_{(1)}^2]}{2(1-\rho_{(1)})} = \frac{\lambda E[X_o^2]}{2(1-\rho)}.$$

When $K = 2$, the service time random variable of the kth service grade $X_{(2)k} \sim \frac{\mu}{\mu_k} X_o$ w.p. p_k, where $k = 1, 2$. Thus $E[X_{(2)k}^2] = \frac{\mu^2}{\mu_k^2} E[X_o^2]$. The definition of the average waiting time of the kth grade with the preemptive priority-based scheduling rule is (Harchol-Balter (2011)):

$$E[w_k] = \frac{E[X_k] \sum_{i=1}^{k-1} \rho_i}{1 - \sum_{i=1}^{k-1} \rho_i} + \frac{\lambda \sum_{i=1}^{k} p_i E[X_i^2]}{2(1 - \sum_{i=1}^{k-1} \rho_i)(1 - \sum_{i=1}^{k} \rho_i)}.$$

As we assume that $\mu_1 > \mu_2$, based on the algorithm of the preemptive SEPT scheduling rule, grade 1 is always served before grade 2. We have

$$E[w_{(2)1}] = \frac{\lambda p_1 E[X_{(2)1}^2]}{2(1-\rho_1)} = \frac{\lambda E[X_o^2]\mu^2}{2(1-\rho_1)} \frac{p_1}{\mu_1^2};$$

$$\begin{aligned} E[w_{(2)2}] &= \frac{\rho_1 E[X_{(2)2}]}{1-\rho_1} + \frac{\lambda \sum_{i=1}^{2} p_i E[X_{(2)i}^2]}{2(1-\rho_1)(1-\rho)} \\ &= \frac{\rho_1}{1-\rho_1} \frac{\mu}{\mu_2} E[X_o] \ + \frac{\lambda \left(p_1 E[X_1^2] + p_2 E[X_2^2]\right)}{2(1-\rho_1)(1-\rho)} \\ &= \frac{1}{1-\rho_1} \frac{\lambda p_1}{\mu_1 \mu_2} \ + \frac{\lambda E[X_o^2]\mu^2}{2(1-\rho_1)(1-\rho)} \left(\frac{p_1}{\mu_1^2} + \frac{p_2}{\mu_2^2}\right). \end{aligned}$$

Thus the average waiting time of mixed service is

$$\begin{aligned}
E[w]_{(2)} &= p_1 E[w_{(2)1}] + p_2 E[w_{(2)2}] \\
&= \frac{\lambda E[X_o^2]\mu^2}{2(1-\rho_1)}\frac{p_1^2}{\mu_1^2} + \frac{1}{1-\rho_1}\frac{\lambda p_1 p_2}{\mu_1\mu_2} + \frac{\lambda E[X_o^2]\mu^2 p_2}{2(1-\rho_1)(1-\rho)}\left(\frac{p_1}{\mu_1^2}+\frac{p_2}{\mu_2^2}\right) \\
&= \frac{\lambda p_1 p_2}{(1-\rho_1)\,\mu_1\mu_2} + \frac{\lambda E[X_o^2]\mu^2}{2(1-\rho_1)(1-\rho)}\left(\frac{p_1^2}{\mu_1^2}(1-\rho)+\frac{p_1 p_2}{\mu_1^2}+\frac{p_2^2}{\mu_2^2}\right).
\end{aligned}$$

Now we compute $\dfrac{E[w]_{(2)}}{E[w]_{(1)}}$ to compare the two variables $E[w]_{(1)}$ and $E[w]_{(2)}$.

$$\begin{aligned}
\frac{E[w]_{(2)}}{E[w]_{(1)}} &= \frac{\dfrac{\lambda p_1 p_2}{(1-\rho_1)\,\mu_1\mu_2} + \dfrac{\lambda E[X_o^2]\mu^2}{2(1-\rho_1)(1-\rho)}\left(\dfrac{p_1^2}{\mu_1^2}(1-\rho)+\dfrac{p_1 p_2}{\mu_1^2}+\dfrac{p_2^2}{\mu_2^2}\right)}{\dfrac{\lambda E[X_o^2]}{2(1-\rho)}} \\
&= \frac{2p_1 p_2\,(1-\rho)}{(1-\rho_1)\,\mu_1\mu_2 E[X_o^2]} + \frac{\mu^2}{1-\rho_1}\left(\frac{p_1}{\mu_1^2}(1-p_1\rho)+\frac{p_2^2}{\mu_2^2}\right). \qquad \text{(B.20)}
\end{aligned}$$

Expression (B.20) contains the metrics of the jobs' original distribution, meaning that such a characteristic makes sense for our final result. For convenience, here we evaluate a specific example of the exponential distribution to check whether $(B.20) < 1$ holds. With an exponential distribution, we have $E[X_o^2] = \dfrac{2}{\mu^2}$. Substituting the second moment into (B.20), we have:

$$\begin{aligned}
(B.20) - 1 &= \frac{\mu^2 p_1 p_2\,(1-\rho)}{(1-\rho_1)\,\mu_1\mu_2} + \frac{\mu^2}{1-\rho_1}\left(\frac{p_1}{\mu_1^2}(1-p_1\rho)+\frac{p_2^2}{\mu_2^2}\right) - 1 \\
&= \frac{\mu^2}{1-\rho_1}\left(\frac{p_1 p_2}{\mu_1\mu_2} - \frac{p_1 p_2}{\mu_1\mu_2}\rho + \frac{p_1}{\mu_1^2} - \frac{p_1^2}{\mu_1^2}\rho + \frac{p_2^2}{\mu_2^2}\right) - 1 \\
&= \frac{\mu^2}{1-\rho_1}\left(\frac{p_2}{\mu_2}\left(\frac{p_1}{\mu_1}+\frac{p_2}{\mu_2}\right) - \frac{p_1\rho}{\mu_1}\left(\frac{p_2}{\mu_2}+\frac{p_1}{\mu_1}\right) + \frac{p_1}{\mu_1^2}\right) - 1 \\
&= \frac{\mu^2}{1-\rho_1}\left(\frac{p_2}{\mu_2}\frac{1}{\mu} - \frac{p_1\rho}{\mu_1}\frac{1}{\mu} + \frac{p_1}{\mu_1^2}\right) - 1 \\
&= \frac{1}{1-\rho_1}\left(\frac{p_2\mu}{\mu_2} - \frac{p_1\lambda}{\mu_1} + \frac{p_1\mu^2}{\mu_1^2} - 1 + \rho_1\right) \\
&= \frac{1}{1-\rho_1}\left(\left(\frac{1}{\mu}-\frac{p_1}{\mu_1}\right)\mu - \rho_1 + \frac{p_1\mu^2}{\mu_1^2} - 1 + \rho_1\right) \\
&= \frac{1}{1-\rho_1}\frac{p_1\mu^2}{\mu_1}\left(-\frac{1}{\mu}+\frac{1}{\mu_1}\right) \\
&= \frac{1}{1-\rho_1}\frac{p_1\mu^2}{\mu_1}\left(\frac{1}{\mu_1}-\frac{p_1}{\mu_1}-\frac{p_2}{\mu_2}\right) \\
&= \frac{1}{1-\rho_1}\frac{p_1\mu^2}{\mu_1}\left(\frac{p_2}{\mu_1}-\frac{p_2}{\mu_2}\right) \\
&= \frac{1}{1-\rho_1}\frac{p_1 p_2\mu^2}{\mu_1}\left(\frac{1}{\mu_1}-\frac{1}{\mu_2}\right). \qquad \text{(B.21)}
\end{aligned}$$

The above equation $(B.21) < 0$ always holds as $\frac{1}{1-\rho_1}\frac{p_1 p_2 \mu^2}{\mu_1} > 0$ and $\mu_1 > \mu_2 > 0$. This shows that with a preemptive scheduling rule, service rate differentiation *always* benefits the system if the original distribution is exponential. ■However, due to the difficulty in characterizing the average waiting time under a preemptive priority-based scheduling rule, the general condition for the dominance of mixed service is extremely complicated to derive. But it would be worthwhile to check the optimal conditions for other specific distributions.

3 Dynamic policy

In the main body of the paper, we propose a static service rate control rule, in which multiple service rates are determined and assigned to jobs independent of system state. It is also interesting to consider a dynamic policy, which adjusts the service rate according to system state. Different from static service differentiation that can be analyzed for an arbitrary service distribution, a dynamic policy is tractable only for exponential service times, based on which a Markov Chain or Markov Decision Process can be formulated. Then a dynamic policy can be derived based on the current system state. There are two kinds of dynamic policies: the service rate is either dynamically determined from a continuum of choices, or it is chosen from a pre-determined menu of service rates (referred to as "dynamic allocation"). We will discuss the two models separately. When service rates are dynamically determined, the policy is equivalent to a pure state-dependent service rate control policy, which has been extensively studied in prior research (see Stidham and Weber 1989, George and Harrison 2001, Hopp *et al.* 2007.) The basic methodology is to formulate the problem as an MDP and to derive the optimal service rates as a function of system state and/or other parameters. Interested readers are referred to Hopp *et al.* 2007 for more details. Under dynamic allocation, each time a new job starts, the server chooses a service rate from a set of candidate rates based on the current system state. In this case it can be proved that the optimal policy is a threshold policy that assigns service rates based on queue length: The longer the queue, the faster the service rate allocated. (This is similar to the TP control proposed by Armony and Gurvich, 2010.) Based on this structure, we can model a continuous time Markov process that includes two states: the number of customers in system and the number of customers in system when the current service starts, the latter determining the service rate of the job in service. We can solve for the stationary distribution of the process and then the steady-state average waiting time and service time. Though it is too complicated to derive the closed-form of the optimal threshold, we numerically

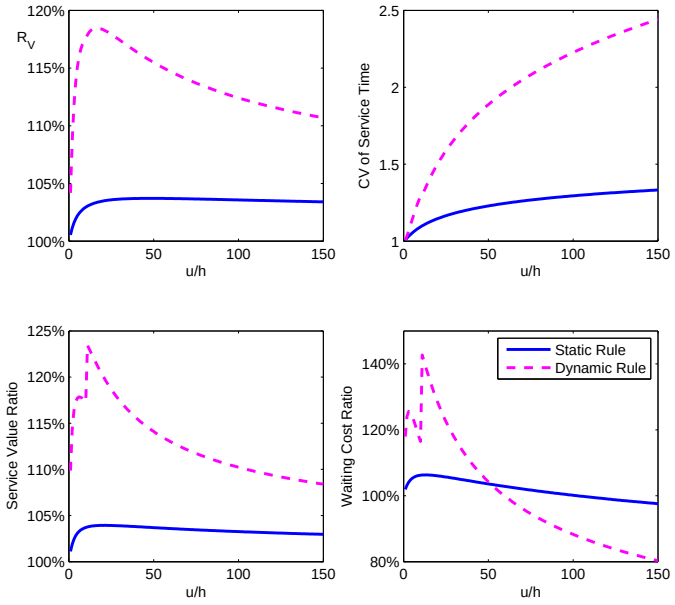

Figure B.31: How R_V, coefficient of variation of service time, service value and average waiting cost change with u/h under static and dynamic allocation rule, respectively

compute the optimal dynamic policy of a two-grade system and demonstrate the benefits of the dynamic allocation rule by comparing it with a static one; facing the same arrival rate and the same vector of service rates, each system chooses its own optimal allocation rules. In Figure B.31 we compare metrics including dominance index R_V, coefficient of variation of the optimal service time, service value and average waiting cost under different settings of u/h. The upper two plots in Figure B.31 suggests that a dynamic allocation rule introduces more service time variability and generates up to 18% more system value. The two graphs at the bottom demonstrate the ratio of service value and waiting cost under the dynamic rule to these of pure service, respectively. We can see that both ratios have phase changes with respect to the value of u/h.

Appendix C

supplements for Chapter 4

Bibliography

[1] Acaroglu H, Dagdemir O (2010) The effects of globalization on child labor in developing countries. *Business and Economic Horizons* 2:37-47.

[2] Babich V., Tang C (2012) Managing opportunistic supplier product adulteration: deferred payments, inspection, and combined mechanisms. *Manufacturing & Service Operations Management* 14(2): 301-314.

[3] Baiman S, Fischer PE, Rajan MV (2000) Information, contracting, and quality costs. *Management Science* 46(6):776-789.

[4] Balachandran KR, Radhakrishnan S (2005) Quality implications of warranties in a supply chain. *Management Science* 51(8):1266-1277.

[5] Bartlett CA, Dessain V, Sjőman A (2006) IKEA's global sourcing challenge: Indian rugs and child labor(A). Harvard Business School, Case 906-414.

[6] Basu K (1999) International labor standards and child labor. *Challenge* 42(5):80-93.

[7] Basu K, Zarghamee H (2009) Is product boycott a good idea for controlling child labor? A theoretical investigation. *Journal of Development Economics* 88(2):217-220.

[8] Basu K, Das S, Dutta B (2010) Child labor and household wealth: theory and empirical evidence of an inverted-U. *Journal of Development Economics* 91(1):8-14.

[9] BBC (2014) Samsung re-uses 'child labour' firm but cuts business by 30%. *BBC News.* August 5, 2014.

[10] Business & Human Rights Resource Centre (2014) 85 firms still "silent" on California Transparency in Supply Chains Act. Available at www.csrwire.com.

[11] Chen YJ, Deng M (2013) Supplier certification and quality investment in supply chains. *Naval Research Logistics* 60(3):175-189.

[12] Chen L, Lee H (2014) Mitigate supplier responsibility risk in emerging economies: an ethical sourcing framework. Working paper, Duke University.

[13] D'Innocenzio A (2013) Walmart adopts 'zero tolerance policy' on subcontracting after Bangladesh factor fire. *The Huffington Post.* January 22, 2013.

[14] Edmonds E, Pavcnik N (2005) Child labor in the global economy. *The Journal of Economic Perspective* 19(1):199-220.

[15] Fyfe A, Jankanish M (1997) Trade unions and child labour. International Labour Organization, Geneva.

[16] Guo R, Lee HL, Swinney R (2014) Responsible sourcing in supply chains. Working paper, Stanford University.

[17] Hsieh CC, Liu YT (2010) Quality investment and inspection policy in a supplier-manufacturer supply chain. *European Journal of Operational Research* 202(3):717-729.

[18] Hwang I, Radhakrishnan S, Su LN (2006) Vendor certification and appraisal: implications for supplier quality. *Management Science* 52(10):1472-1482.

[19] ILO (International Labour Organization) (2007) Toolkit for mainstreaming employment and decent work.

[20] ILO (International Labour Organization) (2010) Facts on child labour 2010.

[21] ILO (International Labour Organization) (2013) Marking progress against child labour: global estimates and trends 2000-2012.

[22] International Finance Corporation (2002) Good practice note: addressing child labor in the workplace and supply chain.

[23] Iram U, Fatima A (2008) International trade, foreign direct investment and the phenomenon of child labor: the case of Pakistan. *International Journal of Social Economics* 35(11):809-822.

[24] Kim SH (2014) Time to come clean? Disclosure and inspection policies for green production. *Operations Research*, forthcoming.

[25] Kolk A, van Tulder R (2002a) Child labor and multinational conduct: a comparison of international business and stakeholder codes. *Journal of Business Ethics* 36(3):291-301.

[26] Kolk A, van Tulder R (2002b) The effectiveness of self-regulation: corporate codes of conduct and child labour. *European Management Journal* 20(3):260-271.

[27] Krishna A, Rajan U (2009) Cause marketing: spillover effects of cause-related products in a product portfolio. *Management Science* 55(9):1469-1485.

[28] Lim SJ, Phillips J (2008) Embedding CSR values: the global footwear industry's evolving governance structure. *Journal of Business Ethics* 81(1):143-156.

[29] Locke RM (2003) The promise and perils of globalization: the case of Nike. In *Management: Inventing and Delivering Its Future*, Kochan T, Schmalensee R (eds). MIT Press: Cambridge, MA.

[30] Lund-Thomsen P (2008) The global sourcing and codes of conduct debate: five myths and five recommendations. *Development and Change* 39(6):1005-1018.

[31] Plambeck E, Taylor T (2014) Supplier evasion of a buyer's audit: implications for motivating supplier social and environmental responsibility. Working paper, Stanford University.

[32] Smith NC (2003) Corporate social responsibility: whether or how? *California Management Review* 45(4):52-76.

[33] Subramanian S (2013) Bayer Cropscience in India (A): against child labor. Richard Ivey School of Business, Case, 9B10M061.

[34] U.S. Department of Labor (2000) By the sweat and toil of children: an economic consideration of child labor (Vol. VI).

[35] U.S. Department of Labor (2014) List of goods produced by child labor or forced labor.

[36] Anand, KS, F Pac, S Veeraraghavan. (2011) Quality-Speed Conundrum: Trade-offs in Customer-Intensive Services. *Management Sci.* **57** 40-56.

[37] Armony, M, A Manderbaum. 2011. Routing and Staffing in Large-Scale Service Systems: The Case of Homogeneous Impatient Customers and Heterogeneous Servers. *Oper. Res.* **59**(1) 50-65.

[38] Ata, B, S Shneorson. 2006. Dynamic Control of an M/M/1 Service System with Adjustable Arrival and Service Rates. *Management Sci.* **52**(11) 1778-1791.

[39] Crabill, TB. 1972. Optimal Control of a Service Facility with Variable Exponential Service Times and Constant Arrival Rate. *Management Sci.* **18** 560-566.

[40] Crabill, TB. 1974. Optimal Control of a Maintenance system with Variable Service Rates. *Oper. Res.* **22** 736-745.

[41] Debo, LG, L. B. Toktay, L. N. Van Wassenhove. 2008. Queuing for Expert Services. *Management Sci.* **54**(8) 1497-1512.

[42] de Vericourt, F., P. Sun. 2009. Judgment Accuracy under Congestion in Service Systems. Working paper, Duke University.

[43] George, J. M., J. M. Harrison. 2001. Dynamic Control of a Queue with Adjustable Service Rate. *Oper. Res.* **49**(5) 720-731.

[44] Gurvich, I., M. Armony, A. Mandelbaum. 2008. Service-Level Differentiation in Call Centers with Fully Flexibile Servers. *Management Sci.* **54**(2) 279-294

[45] Gurvich, I., W. Whitt. 2010. Service-Level Differentiation in Many-Server Service Systems via Queue-Ratio Routing. *Oper. Res.* **58**(2) 316-328.

[46] Harchol-Balter, Mor. 2011. Performance Analysis and Design of Computer Systems. In progress.

[47] Harchol-Balter, Mor., C. Li, T. Osogami, A. Scheller-Wolf and M. S. Squillante. 2003. Cycle Stealing under Immediate Dispatch Task Assignment. *Proceedings of the Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA)(June 2003)*. pp. 274–285.

[48] Heyman, D. P. 1977. The T-Policy for the M/G/1 Queue. *Management Sci.* **23**(7) 775-778.

[49] Hopp, W. J., S. M. R. Iravani, G. Y. Yuen. 2007. Operations Systems with Discretionary Task Completion. *Management Sci.* **53**(1) 61-77.

[50] Kostami, V., S. Rajagopalan. 2009. Speed Quality Tradeoffs in a Dynamic Model. Working paper, University of Southern California.

[51] Lovejoy, W. S., K. Sethuraman. 2000. Congestion and Complexity Costs in a Plant with Fixed Resources that Strives to Make Schedule. *Manufacturing and Service Oper. Management* **2**(3) 221-239.

[52] Mendelson, H. and S. Whang. 1990. Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue. *Oper. Res.* **38**(5).

[53] Moreno, P. 2009. A Discrete-Time Single-Server Queueing System Under Multiple Vacations and Setup-Closedown Times. *Stochastic Analysis and Applications* **27**(2) 221-239.

[54] Rao, S., E. R. Petersen. 1998. Optimal Pricing of Priority Services. *Oper. Res.* **46**(1).

[55] Ren, Z. J., Y. Zhou. 2008. Call Center Outsourcing: Coordinating Staffing Level and Service Quality. *Management Sci.* **54**: 369-383.

[56] Stidham, S. Jr., R. R. Weber. 1989. Monotonic and Insensitive Optimal Policies for Control of Queues with Undiscounted Costs. *Oper. Res.* **87**(4) 611–625.

[57] van Mieghem, J. A. 2000. Price and Service Discrimination in Queuing Systems: Incentive Compatibility of Gcμ Scheduling. *Management Sci.* **46**(9).

[58] Wang, K., Yang D., W. L. Pearn. 2010a. Comparison of Two Randomized Policy M/G/1 Queues with Second Optional Service, Server Breakdown and Startup. *Journal of Computational and Applied Mathematics* **234**(3) 812-824.

[59] Wang, X., L. G. Debo, A. Scheller-Wolf, S. Smith. 2007. Service Design at Diagnostic Service Centers. Working paper, Tepper School of Business, Carnegie Mellon University, Pittsburgh.

[60] Wang, X., L. G. Debo, A. Scheller-Wolf, S. Smith. 2010b. Design and Analysis of Diagnostic Service Centers. *Management Sci.* **56** 1873-1890.

[61] Wang, X., N. Policella, S. Smith, A. Oddi. 2011. Cosntraint-based Methods for Scheduling Discretionary Services. *AI Communications* **24** 51-73.

[62] Wolff, R. W. 1989. Stochastic Modeling and the Theory of Queues. Prentice-Hall, Englewood Cliffs, N.J.

[63] Mailath GJ, Samuelson L. 2006. Repeated games and reputations (Vol. 2). Oxford: Oxford university press.

[64] Puterman, ML. 1994. Discrete Stochastic Dynamic Programming, John Wiley and Sons.

[65] Ross, SM. 1983. Introduction to stochastic dynamic programming. Academic press.

[66] Benjaafar, S, Elahi, E, Donohue, KL (2007). Outsourcing via Service Competition. *Management Sci.* **53**(2) 241–259.

[67] Wright, B. 2014. Contact Center Gamification: Competition or Collaboration? September 28, 2014. Source: http://www.customerexperiencereport.com/tactics-and-operations/contact-center-gamification-competition-collaboration/.

[68] Cachon, G. P., Zhang, F. (2007). Obtaining Fast Service in a Queueing System via Performance-Based Allocation of Demand. *Management Sci.* **53**(3) 408–420.

[69] C. Conti. 2014. How to improve cusoter satisfaction in your medical call center. January 27, 2014. Source: http://www.the-connection.com/how-to-improve-customer-satisfaction-in-your-medical-call-center/.

[70] Elahi E, Benjaafar S, Donohue, K (2012) Optimal Service-Based Competition with Heterogeneous Suppliers. Working paper. University of Massachusetts.

[71] Gilbert, S. M., Weng ZK (1998) Incentive effects favor nonconsolidating queues in a service system: The principal–agent perspective. *Management Sci.* **44**(12-Part-1) 1662-1669.

[72] He QM, Neuts MF (2002). Two M/M/1 queues with transfers of customers. *Queueing Systems* 42(4), 377–400.

[73] Hu B, Duenyas I, Beil DR (2013) Does Pooling Purchases Lead to Higher Profits? *Management Sci.* 59(7), 1576–1593.

[74] Anily S, Haviv M (2010) Cooperation in Service Systems. *Oper. Res.* 58(3), 660–673.

[75] Yu Y, Benjaafar S, Gerchak Y (2015) Capacity sharing and cost allocation among independent firms with congestion. *Production and Oper. Management.* forthcoming.

[76] Clover C (2014) Apple suppliers Pegatron and Foxconn ramp up hiring. *Finantial Times.* June 23, 2014

[77] Levina N, Su N (2008) Global multisourcing strategy: The emergence of a supplier portfolio in services offshoring. *Decision Sci.* **39** (3) 541-570.

[78] de Vericourt, F, YP. Zhou. 2005. Managing response time in a call-routing problem with service failure. *Oper. Res.* **53**(6) 968981.