1-1-2014

Essays on Service Operations Management

Jaelynn Oh
University of Pennsylvania, jaelynnoh@gmail.com

Essays on Service Operations Management

Abstract
This dissertation studies three different problems service firms can face. The first chapter looks at the optimal way to price reservations and services when customers make reservations in advance, while they are uncertain about the future value of service, to avoid waiting on the day of service. We show that charging customers the full price as non-refundable deposit when they make reservations and charging zero for service when they show up to claim their reservations is optimal for the firm. When the firm faces very large potential market, then it is better for the firm to not take reservations and accept only walk-ins. The second chapter looks at a problem of how to mitigate worker demotivations due to fairness concerns, when workers have intrinsic difference in quality, and higher quality server tends to be overcrowded by customers willing to receive higher quality service. We suggest distributing workload fairly between workers and compensating workers per workload as potential remedies and show which remedy works well under what operational conditions. We show that compensating workers per customer they serve results in high customer expected utility and expected quality. However, when customers also care about fairness and dislike receiving inferior service compared to other customers, then there does not exist a single remedy that results in both high customer expected utilization and high expected quality. In the third chapter, we study how a service firm should choose its advertising strategy when the service quality is not perfectly known to the customers. We model customers' learning process using a Markov chain, and show that when customers do not perfectly learn the quality of service from advertisements, then the firm is better off by advertising actively when customers' initial belief about service quality is low. Oppositely, when customers initially believe the service quality to be high, then it is better for the firm to stay silent and not use advertisement to signal its quality. In all three chapters, we use game theory to model the interactions among the participants of the problem and find the equilibrium outcomes.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Operations & Information Management

First Advisor
Xuanming Su

Keywords
Game Theory, Pricing, Service Operations Management

Subject Categories
Business

ESSAYS ON SERVICE OPERATIONS MANAGEMENT

Jaelynn Oh

A DISSERTATION

in

Operations and Information Management

For the Graduate Group in Managerial Science and Applied Economics
Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

2014

Supervisor of Dissertation

_____

Xuanming Su,
Associate Professor of Operations and Information Management

Graduate Group Chairperson

_____

Eric Bradlow,
K.P. Chao Professor, Professor of Marketing, Statistics and Education,
Vice-Dean and Director of Wharton Doctoral Programs

Dissertation Committee:

Morris Cohen,
Panasonic Professor of Manufacturing and Logistics,
Professor of Operations and Information Management

Senthil Veeraraghavan,
Associate Professor of Operations and Information Management

Acknowledgements

I would like to express my special appreciation to my advisor, Professor Xuanming Su, who has been a great mentor, friend and a supporter throughout my Ph.D. studies. My non-native English cannot express my thanks to Xuanming enough, but I do not think my native Korean can do either. I feel extremely grateful to have an advisor who became my role model both academically and personally. I would also like to thank professor Morris Cohen and professor Senthil Veeraraghavan for serving as my committee members and for the brilliant comments and suggestions. I also thank the faculties in the OPIM department I had interactions with. I felt lucky to be at the same place as the great scholars in operations management with outstanding scholarly achievements but who still kept humble attitudes towards life and research.

I thank my friends at Wharton, Hengchen, Jun, Vibhanshu, Hessam, Pnina, Necati, Alessandro, Santiago, Bob, Joel, Fazil, and John who became my brothers and sisters. I enjoyed their presence throughout the program, and will truly miss the good times we spent together as Ph.D. students. I also give thanks to my church friends, Daehwan, Jungmin and Helen who kept me in their prayers when I was going through hardships of life. I give special thanks to Myung, who gave me unconditional support and volunteered to be called as my "crew" to make my first year in Salt Lake City a wonderful one.

I give special thanks to my family. The prayer of my mother and father is what sustained me thus far. I was never alone because of their prayers even though I was physically apart

from them studying overseas. I cannot help but mention my mother's effort to send mom-made Korean food to the U.S. to feed me well. I also thank Yonggwan, who gave me the energy to finish up the dissertation.

Lastly, I thank God, who made all the goodness in my life happen.

"I am not saying this because I am in need, for I have learned to be content whatever the circumstances. I know what it is to be in need, and I know what it is to have plenty. I have learned the secret of being content in any and every situation, whether well fed or hungry, whether living in plenty or in want. I can do everything through him who gives me strength." <Philippians 4:11-13>

ABSTRACT

ESSAYS ON SERVICE OPERATIONS MANAGEMENT

Jaelynn Oh

Xuanming Su

This dissertation studies three different problems service firms can face. The first chapter looks at the optimal way to price reservations and services when customers make reservations in advance, while they are uncertain about the future value of service, to avoid waiting on the day of service. We show that charging customers the full price as non-refundable deposit when they make reservations and charging zero for service when they show up to claim their reservations is optimal for the firm. When the firm faces very large potential market, then it is better for the firm to not take reservations and accept only walk-ins. The second chapter looks at a problem of how to mitigate worker demotivations due to fairness concerns, when workers have intrinsic difference in quality, and higher quality server tends to be overcrowded by customers willing to receive higher quality service. We suggest distributing workload fairly between workers and compensating workers per workload as potential remedies and show which remedy works well under what operational conditions. We show that compensating workers per customer they serve results in high customer expected utility and expected quality. However, when customers also care about fairness and dislike receiving inferior service compared to other customers, then there does not exist a single remedy that results in both high customer expected utilization and high expected quality. In the third chapter, we study how a service firm should choose its advertising strategy when the service quality is not perfectly known to the customers. We model customers' learning process using a Markov chain, and show that when customers do not perfectly learn the quality of service from advertisements, then the firm is better off by advertising actively when customers' initial belief about service quality is low. Oppositely,

when customers initially believe the service quality to be high, then it is better for the firm to stay silent and not use advertisement to signal its quality. In all three chapters, we use game theory to model the interactions among the participants of the problem and find the equilibrium outcomes.

Contents

List of Figures

Chapter 1

Introduction

My dissertation thesis is on theoretically modeling the decision makers' behavior in service operation management problems. The thesis is comprised of three projects that focus on problems service firms face.

The first paper, "Optimal Pricing of Reservations and Services in Queues", looks at what the optimal pricing policy of reservations for services is. We model the service firm as a queue and frame reservations as an option selling strategy where strategic customers pay a non-refundable deposit while they are uncertain about their future valuation of service in return for a no-wait guarantee. We characterize the rational expectations equilibrium and offer recommendations on pricing, which defines the reservation policy to be one of three regimes: fully-prepaid reservations, partially-prepaid reservations, and no reservations. The result of our model suggests that service firms should sell fully-prepaid reservations by charging the full price of service for the non-refundable deposit and offering the service for free if a reservation holder shows up when customers hold homogeneous expectation about the valuation of service. The firm should have walk-in customers pay a higher price for service than what the reservation customers pay if it decides to both take reservations and allow walk-ins. When the potential market size becomes very large, then

the firm is better off by not offering any reservation and spot selling all its capacity even though the customers may have to wait on the day of service. We also show that it is optimal to sell partially-prepaid reservations when customers hold heterogenous expectations about the service valuation or the firm allows overbooking. A modeling contribution we make to the literature is that we model reservations as means to avoid waiting in a queue, while other existing works see it as a way to secure capacity.

The second paper, "Workload Inequality and Fairness Concerns in Service Firms", looks at a problem where service firms hiring multiple workers to supply necessary capacity, and difference in server ability leading to variance in quality. When customers can observe the difference in server quality, they tend to crowd the higher quality sever, and the better workers end up being overloaded with more customers. This may trigger the good workers' fairness concerns and lead them to work less hard. We consider two ways to deal with worker demotivation due to fairness concerns on workload: eliminating and compensating. A firm can eliminate inequality by distributing customers fairly between workers, or it can let the inequality in workload persist, but compensate the workers for having more work by paying workers piece rate. We compare the two remedies using two metrics, expected customer utility and expected quality. Our results show that eliminating inequality in workload helps fix the situation and provides higher expected customer utility and average quality, but compensating workers per customers they serve works even better. However, when customers also care for equality in service quality, congestion externality for the higher quality server exacerbates, and a remedy that leads to both high customer expected utility and average quality no longer exists. In this case, quality-utility trade-off exists, and paying workers piece rate leads to high expected quality but low customer expected utility by increasing service quality difference. Distributing workload fairly between workers then emerges as an option, which provides reasonable quality without sacrificing customer utility too much. This paper contributes to the literature in that, as far as I am

aware of, it is the first paper to model fairness concerns of workers in service firms, while most papers that consider fairness in congestion prone environment talks about customer fairness concerns about service speed or priority in service.

In the third paper, "Signaling Service Quality through Advertising: A Model of Consumer Memory", we try to answer how intensively service firms with certain service quality should advertise when their service qualities are not exactly known to the customers. We look at a model where the service value, either high or low, is known only to the firm and customers are only aware of the distribution of the service value. Given its service value, the firm chooses the frequency of advertising messages it releases to maximize profit. Each customer then encounters and forgets the messages stochastically, and forms heterogeneity in the number of advertisement messages they remember. Based on the number of messages a customer remembers, she updates her belief about the firm's service quality and strategically decides whether to purchase the service or not by comparing her updated expected value and the expected cost from crowding at the server that will be incurred upon purchase. We model the interaction between the firm and the customers using a signaling game. By analyzing the separating equilibrium, we first show that when customers' initial belief about service quality is low, service firms use separating strategy, and they "actively" use advertisement to attract more customers. Even when the firm chooses a separating equilibrium, because there is randomness in the way customers encounter and process the advertisement messages, advertisements cannot signal service quality perfectly. Therefore, not only a high-quality firm but also a low-quality firm benefits from advertisement for there will be some customers who may think that the firm is of high type after seeing the ad. We also find that when customers' initial expected service value is high, then not only a low-quality firm but also a high-quality firm will prefer not to signal its quality through advertisement by choosing a pooling strategy. In this case, the high-quality firm prefers to make advertisement obsolete by choosing the same advertisement intensity

as the low-quality firm so that the customers cannot update their initial belief about the service value.

Chapter 2

Optimal Pricing of Reservations and Services in Queues

2.1 Introduction

Reservations for services allow customers to receive immediate service without waiting and thus, help firms attract delay-sensitive customers. Because of this additional utility reservations bring to customers, there can be opportunity for higher profit to the firm when reservations are priced carefully. When choosing the price of reservations, firms should be aware of customers' uncertainty about future plans that lead to no-shows and trade-offs between taking reservations and having walk-in customers wait longer.

As an effort to benefit from the pros and fight the cons, service firms adopt various reservation policies with different pricing schemes. Some firms that have enough demand walking in even without reservations choose not to take any reservation. Popular restaurants in New York City such as Grimaldi's Pizzeria and Roberto are well known for not taking reservations and having their customers queue to dine at their restaurants. Disneyland sells non-refundable, full-priced tickets online to customers who want to avoid long

waits on the spot, selling fully-prepaid reservations. Some restaurants and doctor's offices sell partially-prepaid reservations by asking for non-refundable deposits to customers (or patients) making reservations or by taking their credit card information so as to charge a fee in case of a no-show.

In this paper, we are interested in finding the optimal way to price reservations. Especially, we intend to answer three questions. First, we study how reservations should be priced. In order to answer this question we find the optimal non-refundable deposit a firm should ask for when taking reservations. Depending on the amount of deposit the customers should pay, reservations can be fully-prepaid or partially-prepaid. Second, we address how service should be priced with or without reservation. To answer this question, we solve the optimal prices of service with and without reservation and compare the two. Third, we ask whether service firms should take any reservation at all and provide conditions under which reservations are profitable. By answering the above research questions, we can eventually find which pricing regime among fully-prepaid reservation, partially-prepaid reservation, and no reservation should be adopted under what operational circumstance.

We build an analytic model where a firm is a monopoly that takes walk-in customers on the day of service but sets a booking limit in advance to take reservations. When the number of customers requesting reservations exceeds the booking limit, those excess customers cannot secure a reservation. Reservation holders receive priority in service and need not wait in line before being served when they show up for the service. However, when reservations are made, customers need to pay a non-refundable deposit in addition to the price of service they pay on the day of service when they show up. At the moment customers make a reservation they are uncertain about their future valuation of service, so that they bear the risk of losing the deposit if they are unable to keep the reservation. On the other hand, to customers who were not able to make a reservation, the service firm is a queue where customers face an expected waiting time before being served. We adopt the concept of rational

expectations (RE) and assume that although customers do not know the exact waiting time they will face in case they walk in without reservations, they form rational expectations of waiting time, and make their decisions to reserve in advance or walk in on the spot, accordingly. Customers are strategic in that they anticipate future outcomes and make a reservation only if the *ex ante* expected payoff from reservation is higher than that from the outside option of walking in. In equilibrium, customers make reservation decisions given their expectations about the wait time and the firm's price and reservation policy, and the restaurant determines its optimal deposit and price of service given customers' behavior.

By solving for the equilibrium non-refundable reservation deposit and service prices, we first find that it is optimal to sell fully-prepaid reservations to customers that have homogeneous *ex ante* valuation of service when the firm does not overbook. That is, the firm should charge customers the full price of service the moment they make reservations and provide service for free when reservation holders show up on the day of service. We also find that when the firm decides to take both reservations and walk-ins, then walk-in customers should be charged a higher price despite the disutility they incur from waiting. Because the firm can better match capacity with demand from reservation holders, marginal revenue is higher for unit capacity reserved than the capacity waiting for walk-ins, and this leads the firm to give discounts to reservation customers. However, when a firm has a very large potential market size, then it is better that the firm does not take any reservation. Since reservations are made before customers observe their valuation of service, reservation price can only be decided based on customers' *ex ante* belief about their future valuation, whereas walk-in prices can be determined based on the realized valuation of customers. Thus, when the market size is large, there can be large number of customers with high realized valuation who are willing to pay high price.

We also show in the extension that partially-prepaid reservations are optimal either when the firm overbooks or when customers have heterogeneous *ex ante* valuation of ser-

vice. When capacity is overbooked, the firm can serve other customers with capacity originally reserved but unclaimed by no-shows. Therefore, charging non-zero price for service can increase profit for the firm. When customers are heterogeneous in their valuation of service, then the firm charges low reservation deposit and high service price to sell reservation to customers even with low *ex ante* valuation but induce only the ones with high realized valuation to keep the reservation.

The reminder of the paper is structured as follows. Section 2 provides literature review. Section 3 describes the model fundamentals. Section 4 provides the equilibrium analysis. Section 5 discuss extensions of the model, and Section 6 concludes.

2.2 Literature Review

This paper is closely related to literature on advance selling in which purchase and consumption are temporally separate, and customers are uncertain about the consumption valuation and the availability of the capacity at the point of purchase. Shugan and Xie (2000) show how customers make a purchasing decision when they do not know their valuation of consumption at the time they purchase the product. They show that service providers can earn profit up to the level of first degree price discrimination through advance selling strategy. DeGraba (1995) shows that capacity constraints and the threat of subsequent unavailability induce customers to buy in advance and allow prices to be set above market-clearing levels. Shugan and Xie (2005) show how advance selling can increase profitability of the seller under competition. Reservations are an example of advance selling in that customers make a reservation when there may still be uncertainty over their preferences at the time the service will take place.

Some papers in the advance selling literature further develop models of customer valuation uncertainty. Yu et al. (2008) look at the pricing and capacity decision of a firm in

advance selling when customers' valuations are correlated. Fay and Xie (2008) and Jerath et al. (2010) study opaque selling where the uncertainty is in "which product" instead of the consumption valuation of a certain product. In other words, customers make a purchasing decision without knowing exactly what product or service they will receive from a seller that sells multiple distinct items. Nasiry and Popescu (2012) look at how anticipated regret can affect customers' decision and firms' profit in an advance selling context. In our paper, restaurant patrons face two types of uncertainty: an independent idiosyncratic shock on the day of consumption (reflecting changes in their schedule, mood, or health condition), and the waiting cost they may incur when walking in without any reservation.

Another stream of research considers advance selling as means to acquire demand information for a later selling period. Fisher and Raman (1996) look at a quick response system for fashion products where a greater portion of production is scheduled in response to initial demand. Tang et al. (2004) suggest an advance booking discount program that attracts customers to commit to their orders prior to selling season. The authors evaluate the benefit of the program and characterize the optimal price when the seller uses the sales information collected from the program to update demand forecast for the selling season. In a similar context, Boyacı and Özer (2010) study a seller's capacity decision problem with exogenously given prices and Prasad et al. (2011) consider the price and inventory decision when the seller is a newsvendor. Li and Zhang (2013) study preorders as means to acquire demand information of the regular sales period. They show that accurate demand information can increase the availability of the product and thus, undermine the seller's ability to charge a high preorder price. Chu and Zhang (2011) consider information that flows from the seller to the customers. They study how much information about the product the seller should release in the preorder stage in order to maximize profit. Although some restaurant managers use reservations to forecast demand and staff accordingly, information is not our focus in this paper. We assume that the potential demand and the product characteristics

are known both to the restaurant and the customers.

There are papers on the specific topic of reservation systems as a form of advance selling in the presence of strategic customers. Png (1989) introduces a model where reservations act as an insurance mechanism that allows customers to show up only when their realized valuation is high. The author shows that when a seller with limited capacity sells its capacity to risk-averse customers who are uncertain about both the valuation of consumption and the availability of capacity, then the most profitable pricing strategy takes the form of a reservation system. Our paper is closely related to Gallego and Şahin (2006) in that they compare the revenue under call options (or an advance selling strategy that charges customers partially refundable fees) to those under pure advance selling, pure spot selling, and low-to-high pricing to find the best pricing policy. They show that call options result in significantly higher revenue than all other alternatives for wide range of capacity levels. They allow the firm to overbook the option capacity, and thus, revenue comes from the value of capacity itself and the value of option from overbooking. This result is similar to what we show in the overbooking extension in our paper although our model is fundamentally different from theirs in that customers who buy on the spot in our model has to join a queue and wait before receiving the service. Alexandrov and Lariviere (2012) look at a restaurant with fixed capacity facing uncertain demand. Customers *a priori* incur travel costs to visit the restaurant and incur additional costs when the restaurant is full and they cannot receive the service. The authors show that reservation is valuable on nights with small demand because it reduces uncertainty for customers and increases demand, while reservation is costly on busy nights due to no-shows. Cil and Lariviere (2013) assume that advance demand customers who request a reservation are more profitable to the service provider than customers who walk-in. The service provider then decides how much of a limited capacity to allocate to advance customers. Georgiadis and Tang (2012) determine a service firm's optimal reservation policy, which consists of the price of service, a non-

refundable deposit to avoid loss from no-shows, and a booking limit. The customers are heterogeneous and divided into four pools: valuation high or low and show-up probability high or low. They show that when the capacity and the demand are large so that they can be approximated by continuous values, then the firm discriminates customers based on their valuation, whereas when the demand and the capacity are finite, then the customers are discriminated based on their show-up probability.

As reviewed above, there is a vast body of literature on advance selling, and our work borrows the concepts introduced in the literature. Note that the three pricing regime that we consider come from the advance selling literature, where fully-prepaid reservations follow the form of advance selling, partially-prepaid reservations are call options, and no reservation corresponds to pure spot selling. However, our work uniquely applies the concept to a service setting and looks at advance selling as means to avoid waiting in a queue rather than ways to secure capacity. In addition, or work can be distinguished from the previous research in that we build one model that can result in one of three different regimes as the optimal pricing mechanism.

The last stream of literature our work is related to is the queueing literature that study pricing of services in a congested server. Naor (1969a) studied how a social planner should price a service when customers who arrive to the service center and queue to obtain the service cause negative congestion externalities to other customers. Customers have homogeneous valuation of the service, incur cost of waiting when standing in the queue, and decide whether to join the queue upon observing the queue length. The queue in our model resembles that of Mendelson and Whang (1990a) since customers do not observe the exact queue length before deciding to join the queue. Customers make the queueing decision based on their expected waiting cost. For an extensive review on queueing systems, we refer the readers to Hassin and Haviv (2003). In most of the queueing literature, customers' decisions are "to join or not" or "which queue to join." However, we apply the queueing

model to the service operations context, where the corresponding decisions are "whether to show up after making a reservation," and whether to "walk in on the spot or reserve in advance".

2.3 Model

2.3.1 Model Fundamentals

The Service Firm: The service firm is a monopoly with a fixed market size equal to Λ and capacity equal to μ. The market size and the capacity are fixed and exogenously given to the firm. For service that takes place in period 2, the firm takes reservations in period 1. It reserves a proportion α of its capacity to take reservations. Throughout the paper, we use the term "booking limit" to denote this proportion of capacity reserved. On the day of service, or in period 2, the firm also takes walk-in customers using its capacity that is either not booked by reservation customers or was originally reserved but unclaimed by reservation holders. In our base model we assume that α is exogenous to the firm. Later we consider the case where the decision of the booking limit, α, is endogenized.

Reservations can be thought of as real call options where the customer pays the price of reservation, ϕ, in period 1 for the right to purchase the service in period 2 at service price, p_r. For customers who make a reservation the firm provides service immediately after their arrival so that reservation holders do not wait before being served when they show up. When more customers try to make reservations than the capacity reserved to serve reservation customers, then the firm only takes reservations up to the capacity set aside for reservation customers.

Customers who were willing but not able to secure a reservation stay in the market to see whether they are still interested to walk into the firm on the day of service. Customers

walking in without reservations may have to wait in line in order to be served upon arriving to the service firm and are charged p_w for the service they receive. Thus, to the walk-in customers the service process is a queue with an effective arrival rate, $\lambda_w \in [0, \Lambda]$, determined in equilibrium, and the capacity, μ_w, that is used to serve walk-in customers, which consists of the capacity that was originally set aside to receive walk-ins on the spot and capacity that was originally booked but is unclaimed. In equilibrium, the expected waiting time for a customer who walks into the firm without reservation is given as $w(\lambda_w, \mu_w)$. We do not assume any specific form of queue for the firm. However, the expected waiting time function, $w(\lambda, \mu)$, satisfies general convexity conditions: 1) $w(\lambda, \mu)$ is convex and increasing in λ. 2) It is decreasing in μ. 3) Pooling the capacity is beneficial, i.e., $w(n\lambda, n\mu) < w(\lambda, \mu)$ for $n > 1$.

The firm's decision is to choose the price of service charged to walk-in customers, p_w, price of service charged to reservation customers, p_r, and the price of making a reservation, ϕ, that maximize its profit. Note that the price of service charged to reservation customers need not be equal to the price charged to walk-in customers, i.e., $p_r \neq p_w$ is possible. We interchangeably use the term option price and strike price for ϕ and p_r, respectively, throughout this paper.

The Customers: Customers are strategic utility maximizers. They are *ex ante* homogeneous but *ex post* differentiated in their valuation of service. A customer's valuation, v, consists of two parts, $v = v_0 + \varepsilon$. v_0 is the average customer valuation, which is known both to the firm and the customers in period 1. We assume that v_0 is exogenously given, equal for all customers and remains constant. Later in the paper, we relax the assumption of all customers having the same expected valuation, v_0, and consider a model with heterogeneous customers. On the day of service, customers face a random shock, ε, in their valuation of service. The shock may be due to medical emergency, change of outside option, or simple change of mind. The shock, ε, follows a continuous distribution, G, with

density, g, and support, $(-\infty, \infty)$ and is i.i.d. over all customers. For notational simplicity, we define a distribution for v to be F with density, f, so that $F(x) = G(x - v_0)$. We use $\bar{F}$ and $\bar{G}$ to denote the complement of F and G, respectively, i.e., $\bar{F} = 1 - F$ and $\bar{G} = 1 - G$. We assume that ε has mean 0 so that the end valuation of service, v, is in expectation the same as the initial perception of value, v_0.

Customers do not observe the queue length at the server before they show up. However, each customer possesses a belief about how long she will have to wait to be served when she walks into the firm without making a reservation in advance. We denote the customers' belief about the expected waiting time using $\hat{w}$. For example, customers usually have a good idea about what the average waiting time at their favorite restaurant is on Saturday nights and take this knowledge into account when they make their plans. In equilibrium, each customer's belief about waiting time is accurate. Based on this belief and the prices the service firm charges, customers' decisions are first, whether to make a reservation or to wait based on their expected valuation. On the day of service, contingent on the realized valuation, reservation holders decide whether to exercise the option or not, and customers without reservations decide whether to walk in or not.

Sequence of Events: Our model is a two-period model. In period 1 the firm chooses the price of service charged to walk-in customers, p_w, price of service charged to reservation customers, p_r, and the non-refundable deposit, ϕ. Given the prices, customers choose whether to make a reservation or not without knowing their valuation of service. At the beginning of period 2, the shock, ε, realizes, and customers observe their valuation of service, $v = v_0 + \varepsilon$. Reservation holders first decide whether to show-up or not. Customers who were not able to secure a reservation in period 1 then decide whether to walk in or stay home based on their realized valuation and the belief about the expected waiting time. We use the term "advance period" interchangeably with period 1 and "spot period" or "the day of service" interchangeably with period 2. The sequence of events is shown in Figure 2.1.

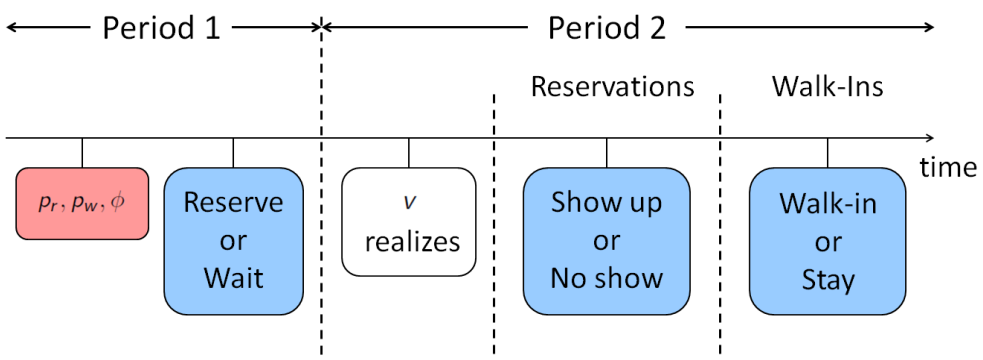

Figure 2.1: *Sequence of Events*

2.3.2 Problem Formulation

In this section we characterize the Rational Expectations (RE) equilibrium resulting from our model. Given the belief about the expected waiting time, $\hat{w}$, and the price vector, (p_w, p_r, ϕ), chosen by the firm, customers' decisions are summarized in Figure 2.2. In the advance period, a customer decides whether to make a reservation or not without knowing her valuation of service exactly. In the spot period, both reservation holders and walk-in customers decide whether to purchase the service or not. Given the customers' strategy, the firm solves its profit maximization problem to compute the equilibrium price. Customers' belief, $\hat{w}$, on the expected waiting time then should be consistent with the equilibrium outcome, i.e., $\hat{w} = w(\lambda_w, \mu_w)$.

The Reservation Market: We first explore a reservation holder's decision to exercise the option or not and the resulting profit for the firm from the reservation market. On the day of service, a customer who has made a reservation observes her valuation shock, ε, and decides to exercise the option when her realized valuation $v = v + \varepsilon$ is greater than the price of service for reservation holders, p_r. Recall that reservation holders need not wait before being served when they show up. When customers follow the above strategy, the

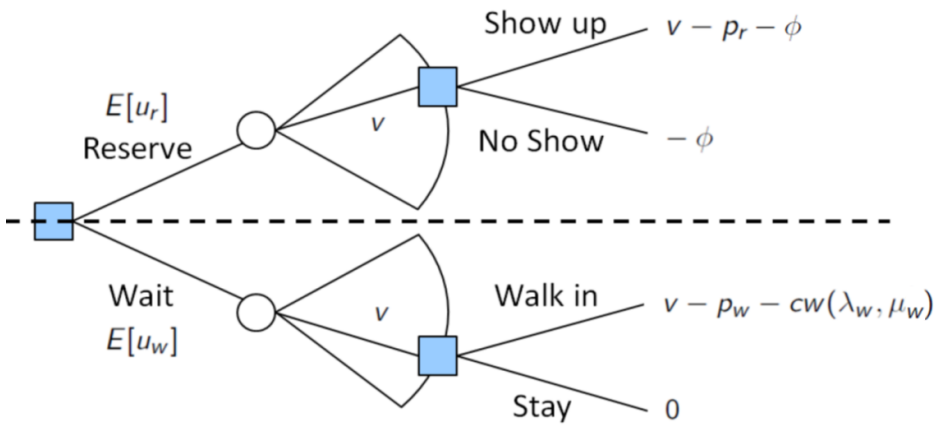

Figure 2.2: *Customers' Decision*

expected utility gained by making a reservation is given as

$$E_v[u_r] = -\phi + E_v[v - p_r]^+ = -\phi + \int_{p_r}^{\infty} (v - p_r) f(v) dv.$$

Moreover, given that making a reservation is an attractive option for customers, the equilibrium in-flow of reservation holders who exercise the option and show up on the day of service, λ_r, is given as

$$\lambda_r = \min\{\Lambda, \alpha\mu\}\, \bar{F}(p_r).$$

Note that since customers are *ex ante* homogeneous, if making a reservation is attractive, then all customers in the potential market will try to make a reservation. However, since the firm's capacity to take reservations is limited to $\alpha\mu$, it only takes $\min\{\Lambda, \alpha\mu\}$ reservations. Hence, given that the price of service charged to reservation customers, p_r, and the price of reservation, ϕ, are chosen so that reservation is an attractive option over waiting, the expected profit the firm earns from the reservation market is determined as

$$\pi_r(p_r, \phi) = \min\{\Lambda, \alpha\mu\} \left\{\phi + p_r \bar{F}(p_r)\right\}.$$

The Walk-in Market: Given the belief, $\hat{w}$, about the expected waiting time, a customer who has not made a reservation in advance anticipates to incur a total cost of $p_w + c\hat{w}$ if she walks in for the service on the day of service. Here, p_w is the price of service charged to walk-in customers, and c is the cost of waiting per unit time. Then, a customer with realized valuation, v, and without reservation will walk into the firm if and only if her belief about the net utility of walking in, $v - p_w - c\hat{w}$, is greater than the outside option that is normalized to zero. The decision of a walk-in customer on the day of service can then be summarized as

$$\begin{cases} \text{walk in} & \text{if } \ v \geq p_w + c\hat{w}, \\ \text{stay home} & \text{if } \ v < p_w + c\hat{w}. \end{cases}$$

The equilibrium walk-in arrival rate, $\lambda_w(p_w)$, is then given as

$$\lambda_w(p_w) = \max\{\Lambda - \alpha\mu, 0\}\bar{F}(p_w + c\hat{w}), \tag{2.1}$$

where the walk-in traffic consists of customers who were not able to secure a reservation in advance, $\max\{\Lambda - \alpha\mu, 0\}$, and among those who have realized valuation greater than the total cost the customers believe to incur, $\bar{F}(p_w + c\hat{w})$. The expected utility earned from walking in on the day of service is then given as

$$E_v[u_w] = E_v[v - p_w - c\hat{w}]^+ = \int_{p_w + c\hat{w}}^{\infty} (v - p_w - c\hat{w}) f(v) dv.$$

Given that the customers behave according to the above equilibrium, the firm's profit earned from the walk-in market can be written as

$$\pi_w(p_w) = p_w \lambda_w(p_w).$$

Firm's Profit maximization Problem: We now formulate the firm's profit maximiza-

tion problem, which solves for the optimal prices. First, note that customers will make a reservation in the advance period only when the expected utility gained by doing so, $E_v[u_r]$, is greater than the expected utility gained by waiting, $E_v[u_w]$. This condition becomes an individual rationality (IR) constraint for the firm's problem. Note that by imposing the (IR) constraint to the firm's profit maximization problem, we are implicitly assuming that taking reservations is profitable to the firm, which may not necessarily be true in all cases. We will revisit this matter when we compare the profitability reservations and walk-ins. Note that when the (IR) constraint does not satisfy, customers will not make any reservation, and the firm will become a pure walk-in service process. For now, we focus on the case where the firm charges prices, (p_w, p_r, ϕ) so that making reservations is attractive to customers. In this case, the firm's profit maximization problem is given as

$$\begin{aligned} \max_{p_w, p_r, \phi} \quad & \pi_r(p_r, \phi) + \pi_w(p_w) \\ \text{s.t.} \quad & E_v[u_r] \geq E_v[u_w] \qquad (IR) \end{aligned} \tag{2.2}$$

Rational Expectations Equilibrium: We use rational expectations equilibrium as a solution approach, originated in Muth (1961) and was developed in Stokey (1981), and define it as below.

Definition 1 *A rational expectations equilibrium* $(p_w^*, p_r^*, \phi^*, \lambda_w^*, \lambda_r^*, \hat{w}^*)$ *satisfies the following:*

(i) Customers' strategy: Given $\hat{w}^*$*,* p_w^**,* p_r^**, and* ϕ^**,*

$$\begin{cases} \lambda_w^* = \max\{\Lambda - \alpha\mu, 0\}\bar{F}(p_w^* + c\hat{w}^*), \\ \lambda_r^* = \min\{\Lambda, \alpha\mu\}\bar{F}(p_r^*). \end{cases} \tag{2.3}$$

(ii) Firm's strategy: Given $\hat{w}^*$, λ_w^*, and λ_r^*,

$$(p_w^*, p_r^*, \phi^*) = \arg \max_{\substack{p_w, p_r, \phi: \\ E[u_r] \geq E[u_w]}} \min\{\Lambda, \alpha\mu\}\, \phi + p_r \lambda_r^* + p_w \lambda_w^*.$$

(iii) Consistent belief:

$$\hat{w}^* = w(\lambda_w^*, \mu_w^*),$$

where $\mu_w^* = \mu - \lambda_r^*$.

Conditions (i) and (ii) assert that under expectations $\hat{w}^*$, customers and the firm will rationally choose the appropriate utility-maximizing actions. Condition (iii) assures that the belief, $\hat{w}^*$, should coincide with the expected waiting time resulting from the equilibrium outcome.

2.4 Analysis

Recall that our goal is to answer the following three questions: 1) How should service firms price reservations? 2) How should services be priced with or without reservation? 3) Should firms offer reservations or not?

2.4.1 Pricing of Reservations

To answer our first question of what the profit maximizing pricing policy for reservations is, we focus on the equilibrium non-refundable deposit, ϕ^* (the price of option), and the price of service for reservation customers, p_r^* (the strike price), in this section. We first formally state the result of the firm's profit maximization problem that tells us p_r^* and ϕ^*.

Proposition 2 (Optimal Reservation Price) *For any given $\alpha \in (0, 1)$, the RE equilibrium*

price vector, (p_w^*, p_r^*, ϕ^*)*, satisfies*

$$
\begin{aligned}
p_r^* &= 0 \\
\phi^* &= \int_0^\infty v f(v) dv - \int_{p_w^* + c\hat{w}^*}^\infty (v - p_w^* - c\hat{w}^*) f(v) dv,
\end{aligned} \tag{2.4}
$$

where $\hat{w}^*$ *is the equilibrium belief given as in Definition 1.*

Proposition 2 suggests that the best way to price reservations is to sell fully-prepaid reservations by setting the price of option, ϕ, as high as possible and charging 0 for the price of service when reservation holders show up to exercise the option. This is equivalent to advance selling the capacity to customers when they have uncertainty in their valuation of service. This is interesting in that the firm chooses to advance sell the capacity for reservation customers, while it has freedom to choose any other option prices, (p_r, ϕ). Therefore, a reservation holder with non-negative *ex post* valuation will show up to claim the valuation. This outcome is socially efficient because when a customer with non-negative realized valuation receives the service, it adds welfare to the whole society, while the prices are only economic transfers within the society.

Although selling fully-prepaid reservations seems to be a rare practice in reality, there are some restaurants that sell tickets for their prix-fixe menu for certain date and time. If a customer does not show up after purchasing a ticket, then no refund is given, and the ticket can be thought of as an option with zero strike price.

2.4.2 Pricing of Services

In order to answer our second question, we compare the equilibrium walk-in price, p_w^*, and the equilibrium reservation price, (ϕ^*, p_r^*) in this section. For a proper comparison, we assume that the market size is large enough, $\Lambda > \mu$, so that there are enough customers who are willing to wait and walk-in on the day of service. In this section, in order to properly

compare the two prices for reservations and walk-ins, we only consider the case where it is profitable for the firm to both take reservations and reserve some capacity to serve walk-in customers. Thus, we endogenize the booking limits for reservations, α, as a decision to the firm's revised profit maximization problem defined as the following:

$$\begin{aligned} \max_{p_w, p_r, \phi, \alpha,} \quad & \pi_r(p_r, \phi) + \pi_w(p_w) \qquad (2.5) \\ s.t. \quad & \int_{p_r - \phi}^{\infty} (v - p_r) f(v) dv - \phi F(p_r - \phi) - \int_{p_w + c\hat{w}} (v - p_w - c\hat{w}) f(v) dv \geq 0, \ \text{ if } \alpha > 0, \\ & \alpha \in [0, 1], \end{aligned}$$

and compare the prices when the optimal capacity reserved for reservations, α^*, is an interior solution with $\alpha^* \in (0, 1)$.

When the firm optimally chooses to reserve positive capacity for both reservation customers and walk-in customers, we show that the equilibrium price a reservation holder pays, ϕ^*, is strictly less than the equilibrium price of service for walk-in customers, p_w^*. The following proposition formally states this.

Proposition 3 (Price of Service) *Assume that $\Lambda > \mu$ and the optimal proportion of capacity reserved to take reservations, α^*, satisfies $\alpha^* \in (0, 1)$. Then, the equilibrium price vector, (p_w^*, p_r^*, ϕ^*) satisfies*

$$\begin{cases} p_r^* = 0, \\ \phi^* < p_w^*. \end{cases}$$

Hereby, we answer our second question of how the firm should price its service. This result states that if a firm can charge a different price for service to its reservation holders compared to that charged to customers walking in, it should give a discount to customers who pay money to buy an option before realizing their valuation of service and bear the risk of paying a price when they have chance not to claim the option.

To understand why the price of reservation ϕ^* is strictly less than p_w^*, let us assume that the two prices are equally set to p so that $p_w = \phi = p$. Then, for each unit of capacity set aside to serve reservation customers the firm earns price, p, independent of the reservation holder's choice to show up or not. However, for a unit of capacity set aside to serve walk-in customers the firm earns strictly less than p since a queueing system needs slack capacity to prevent the queue from exploding to infinity. Therefore, if the price of reservation were either greater than or equal to the price of service charged to walk-in customers, it would be more profitable for the firm to reserve its entire capacity to take reservations. This contradicts our assumption that the firm is optimally reserving positive capacity to walk-in customers, and shows that in equilibrium $\phi^* < p_w^*$ when $\alpha^* \in (0, 1)$.

Now, a natural question that follows is when does the firm indeed want to take both reservations and walk-ins. We answer this question by solving for the optimal proportion of capacity that should be reserved to take reservations.

2.4.3 Setting Booking Limits for Reservations

In this section we answer our last question: When is it profitable for the firm to take reservations? To answer this question, we solve for the optimal proportion of capacity that should be reserved for appointments, (2.5), and let $\alpha^*(\Lambda)$ denote the optimal booking limit for reservations as a function of market size, Λ, when the capacity is fixed to μ.

We begin our analysis by comparing the firm's profit under two special case models, pure reservation ($\alpha = 1$) and pure walk-in ($\alpha = 0$), under the same market size and capacity. Under pure reservation system, the entire capacity is sold as an option, and customers cannot walk into the service process on the day of service. The firm decides the equilibrium price of option, $\tilde{\phi}^*$, and the equilibrium strike price, $\tilde{p}_r^*$. Throughout, we use the "tilde" symbol $(\tilde{\cdot})$ to denote equilibrium measures of a pure system (i.e. either a pure walk-in system or a pure reservation system). The profit maximization problem for the firm,

(2.2), reduces to

$$\begin{aligned} \max_{p_r,\phi} \quad & \mu\left\{\phi + p_r \bar{F}(p_r)\right\} \\ \text{s.t.} \quad & E_v[u_r] \geq 0. \end{aligned}$$

Solving the above optimization problem gives the profit maximizing prices, $(\tilde{p}_r^*, \tilde{\phi}^*)$, characterized as

$$\begin{cases} \tilde{p}_r^* = 0, \\ \tilde{\phi}^* = \int_0^\infty v f(v) dv. \end{cases}$$

On the other hand, under pure walk-in system, the customers cannot make a reservation in advance, and the firm decides the price, $\tilde{p}_w^*$. In other words, pure walk-in system can be considered as the firm allowing spot selling only. When reservation is not offered, the firm's profit maximization becomes

$$\max_{p_w} \quad p_w \lambda_w,$$

where $\tilde{\lambda}_w^* = \Lambda \bar{F}(\tilde{p}_w^* + c\hat{w}^*)$, and $\hat{w}^* = w(\tilde{\lambda}_w^*, \mu)$ in equilibrium.

Now, we compare the equilibrium profits under the two special cases and address that a firm with a sufficiently small market size is better off under pure reservation system, whereas a firm with a sufficiently large potential market performs better under pure walk-in system.

Proposition 4 (Market Size) *Let $\tilde{\pi}_w^*(\Lambda)$ and $\tilde{\pi}_r^*(\Lambda)$ be the equilibrium profit for the firm as a function of Λ, under pure walk-in system and that under pure reservation system, respectively. Then given a fixed μ and the valuation distribution, F, there exists $\tilde{\Lambda}$ such that*

(i) $\tilde{\pi}_w^*(\Lambda) \leq \tilde{\pi}_r^*(\Lambda)$, for all $\Lambda \leq \tilde{\Lambda}$,

(ii) $\tilde{\pi}_w^*(\Lambda) > \tilde{\pi}_r^*(\Lambda)$*, for all* $\Lambda > \tilde{\Lambda}$.

Note that customers' purchasing decisions under pure reservation system are made *ex ante* so that the firm's price decision, $(\tilde{p}_r^*, \tilde{\phi}^*)$, depends purely on the customers' valuation distribution, F, and not on the market size, Λ. Therefore, the price and the profit the firm can earn under pure reservation system do not change as the market size increases. Nevertheless, since the walk-in customers' purchasing decisions are made after they realize their valuations, the firm can find more customers with higher valuation as the market size increases. To illustrate, let us assume that the valuation of consumption, v, follows a uniform distribution, $U(0, \$100)$. If there are 100 customers in the market, the realized valuations of top 75 customers are likely to be greater than \$25, while if the market size is 200, those 75 customers would have valuations greater than \$62.5. Hence, the firm can charge higher price when the potential market is larger, and earn more profit. This explains why pure walk-in system performs better in larger markets, while pure reservation system is more profitable in smaller markets.

Popular restaurants in New York City such as Grimaldi's Pizzeria and Roberto are well known for not taking reservations and have their customers queue up to dine at their restaurants. There are currently more restaurants that are adopting the strategy to not take any reservation and operate as a pure walk-in restaurant. There are many reasons for them to switch to a pure walk-in system, but one thing the restaurant managers agree is that they need high traffic for the pure walk-in system to work (Collins, 2010).

Next, we consider how scaling up the demand and the capacity changes a firm's equilibrium profit under pure walk-in system and that under pure reservation system.

Proposition 5 (Economies of Scale) *Let* $\tilde{\pi}_w^*(\Lambda, \mu)$ *and* $\tilde{\pi}_r^*(\Lambda, \mu)$ *be the optimal profit for the firm as a function of parameters,* Λ *and* μ*, under pure walk-in system and that under pure reservation system, respectively. For any given* Λ*,* μ*, and* $n > 1$*,*

(1) $\tilde{\pi}_w^*(n\Lambda, n\mu) > n\tilde{\pi}_w^*(\Lambda, \mu)$*, (increasing returns to scale)*

(2) $\tilde{\pi}_r^*(n\Lambda, n\mu) = n\tilde{\pi}_r^*(\Lambda, \mu)$*. (constant returns to scale)*

We find that scaling the market size and capacity by the same constant, $n > 1$, scales the walk-in profit by greater than n, while the same scaling in potential market size and capacity only scales the pure reservation profit exactly by n. This means that pure walk-in system has increasing returns to scale, while pure reservation system has constant returns to scale.

With the intuition developed above, we now return to the "hybrid" system and discuss the result of the capacity allocation problem, (2.5), which is summarized in the following proposition:

Proposition 6 (Optimal Booking Limit) *Let* $\Lambda' > \Lambda > \mu$*. Then, corresponding optimal booking limit for reservations,* $\alpha^*(\Lambda')$ *and* $\alpha^*(\Lambda')$ *satisfy:*

$$\alpha^*(\Lambda') \leq \alpha^*(\Lambda),$$

and

$$\lim_{\Lambda \to \infty} \alpha^*(\Lambda) = 0.$$

This result conforms to our previous finding in Proposition 4. In Proposition 4 we stated that for a service firm with greater market size, pure walk-in system brings higher profit than pure reservation system. This is also true in the hybrid system, so that the capacity that should be reserved decreases as the market size increases. Moreover, when the market size increases above a certain value, it shows that operating as a pure walk-in service process performs better than accepting any reservations. In other words, when potential market size is large, the service firm should spot sell all its capacity. Note that this is similar to setting a protection limit in a revenue management problem. Our result suggests that when potential

market size is large, service firms should set the protection limit to its entire capacity.

2.5 Extensions

2.5.1 Heterogeneous Customers

In this section, we relax the assumption of all customers having equal *ex ante* valuation and analyze the firm's equilibrium reservation pricing strategy when the customers have heterogeneous expected valuation, v_0. We assume that a customer is either of two types, high or low. For mathematical simplicity we consider a pure reservation system in this section. We denote $i \in \{H, L\}$ to be a superscript that represents the type of a customer. Customer's *ex ante* expected valuation is v_0^H if the customer is of high-type and v_0^L if she is of low-type, where $v_0^H > v_0^L$. $q \in [0, 1]$ is the proportion of high-type customers in the potential market, which is exogenously given and fixed. Type-i customer's *ex post* valuation of service is given as $v^i = v_0^i + \varepsilon$ with ε being her valuation shock, which realizes on the day of service. We denote the equilibrium price vector of the firm with heterogeneous customers as (p_r^{h*}, ϕ^{h*}), where the superscript, h, represents "heterogeneity" in customer types. Given a price vector, (p_r^h, ϕ^h), type-i customers' expected utility of making a reservation and that of walking-in on the spot period are given as

$$E_\varepsilon[u_r^i] = E_\varepsilon[v_0^i + \varepsilon - p_r^h]^+ = -\phi + \int_{p_r^h - v_0^i}^{\infty} (v_0^i + \varepsilon - p_r^h) g(\varepsilon) d\varepsilon.$$

Type-i customers then make a reservation when $E_\varepsilon[u_r^i]$ is no less than the outside option, which is normalized to 0.

With heterogeneous customer types, there are two possible strategies for the firm: (i) Charge the price of reservation and the strike price of service to the level that only high-type customers make reservations. (ii) Charge the prices so that both high-type and low-type

customers purchase reservations. Note that high-type customers gain higher utility from making reservations than low-type customers, and the case where only low-type customers making reservations is infeasible. The sets of incentive compatibility constraints for the firm then become

$E_\varepsilon[u_r^H] \geq 0$ (IC-H)	$E_\varepsilon[u_r^H] \geq 0$ (IC-H)
$E_\varepsilon[u_r^L] < 0$ (IC-H)	$E_\varepsilon[u_r^L] \geq 0$ (IC-H)
if strategy (i) is adopted.	if strategy (ii) is adopted.

For each strategy, the firm solves its profit maximization problem given as (2.2) subject to corresponding (IC) constraints. After comparing the maximum profits that can be earned from the two strategies, the firm decides its price vector, (p_r^h, ϕ^h). By solving this profit maximization problem, we can show the following:

Proposition 7 (Prices with Heterogeneous Customers) *Let (p_r^{h*}, ϕ^{h*}) denote the equilibrium price vector for the firm. When there are two types of customers in the potential market, for fixed Λ and μ there exists $\tilde{q} \in (0,1)$, such that*
(i) for $q \geq \tilde{q}$, only high-type customers make a reservation, and

$$\begin{cases} p_r^{h*} = 0, \\ \phi^{h*} = \int_{-v_0^H}^{\infty} (v_0^H + \varepsilon) g(\varepsilon) d\varepsilon, \end{cases}$$

(ii) for $q < \tilde{q}$, both high-type and low-type customers make reservations, and

$$\begin{cases} p_r^{h*} > 0, \\ \phi^{h*} = \int_{-v_0^L + p_r^{h*}}^{\infty} (v_0^H + \varepsilon - p_r^{h*}) g(\varepsilon) d\varepsilon. \end{cases}$$

The above proposition shows that when there are enough proportion of customers that have high expected valuation, the firm should have only the high-type customers make reservations. In this case, the firm sells fully-prepaid reservations. The price of reservation is equal to the expected gain, while the strike price is equal to zero. However, as more customers have low expected valuation of service, the firm is better off by selling partially-prepaid reservations than by selling fully-prepaid reservations. This is because when both types of customers purchase reservations, then the firm can no longer extract all surplus from the high-type customers, and selling options works similarly to price discrimination between customers with high *ex post* valuation and those with lower *ex post* service valuation.

2.5.2 Overbooking

In this section, we consider the option to overbook under pure reservation system. Note that overbooking affects the firm's operation only when the market size is greater than the capacity, and thus, we will focus on the case, where $\Lambda > \mu$, for this section. When the market size is smaller than or equal to the capacity, the analysis is equivalent to that under pure reservation system without overbooking. Now, let (p_r^{o*}, ϕ^{o*}) denote the equilibrium reservation price vector when the firm overbooks its capacity, where the superscript o represents "overbooking".

Given any price vector, (p_r^o, ϕ^o), the firm anticipates that fraction $F(p_r^o)$ of customers who have purchased a reservation will not exercise the reservation on the day of service. Thus, the firm can, in expectation, sell up to $\mu/\bar{F}(p_r^o)$ reservations without having to reject any reservation-holding customers due to shortage of capacity on the day of service. The firm's profit under overbooking is then given as $\pi_r^o(p_r^o, \phi^o) = \min\left\{\Lambda, \mu/\bar{F}(p_r^o)\right\}\left\{\phi^o + p_r^o\bar{F}(p_r^o)\right\}$. This expression shows that when the potential market size Λ is smaller than the amount of reservation the firm can handle, the firm takes Λ reservations. The firm's profit maximiza-

tion problem is then given as

$$\begin{aligned} \max_{p_r^o, \phi^o} \quad & \min\left\{\Lambda, \mu/\bar{F}(p_r^o)\right\}\left\{\phi^o + p_r^o \bar{F}(p_r^o)\right\} \\ s.t. \quad & -\phi^o + \int_{p_r^o}^{\infty} (v - p_r^o) f(v) dv \geq 0. \end{aligned}$$

Solving the above optimization problem gives the equilibrium price policy, (p_r^{o*}, ϕ^{o*}), as the following proposition suggests.

Proposition 8 (Overbooking Prices) *For fixed Λ and μ, with $\Lambda > \mu$, the equilibrium price vector, (p_r^{o*}, ϕ^{o*}), is given as the following:*

(i) If $\Lambda \leq \mu/\bar{F}(0)$:

$$\begin{cases} p_r^{o*} = 0, \\ \phi^{o*} = \int_0^{\infty} v f(v) dv. \end{cases}$$

(ii) If $\Lambda > \mu/\bar{F}(0)$:

$$\begin{cases} p_r^{o*} = \bar{F}^{-1}\left(\frac{\mu}{\Lambda}\right) > 0, \\ \phi^{o*} = \int_{p_r^{o*}}^{\infty} (v - p_r^{o*}) f(v) dv. \end{cases}$$

The result states that when the market size is small so that there is enough capacity to serve everybody that are willing to receive the service when it is given for free, then the firm should sell fully-prepaid reservations, or in other words, reservation should take the form of advance selling. As in the case without overbooking, the strike price (or the price of service) charged to reservation holders should be zero, and the price customers pay to make a reservation should equal the expected gain from reservation.

On the other hand, when the capacity is not as large as to cover the entire market when the service is provided for free, then the firm should sell partially-prepaid reservations. In other words, customers should pay $\phi^{o*} > 0$ in period 1 for the right to purchase the service in period 2 at price $p_r^{o*} > 0$. The firm sets the price of option, $\phi^{o*} > 0$, so that it can sell

the option to the entire market, Λ, while it sets the strike price, $p_r^{o*} > 0$. so that the capacity can, in expectation, serve all customers who exercise the option.

Note that when we allow overbooking, capacity that was reserved but was unclaimed by a customer has additional value to the value of capacity, ϕ^{o*}. Because there are more than one customer who booked the capacity, there will be customers who are willing to pay the strike price, p_r^{o*}, and claim the option. This model with overbooking is equivalent to the model Gallego and Şahin (2006) use in their paper. Allowing for overbooking, they show that the use of a call option results in higher profit than the forward selling does unless capacity is large enough to satisfy demand when the product is free, which is in line with the result we just showed.

However, note that the above result showing overbooking as a profitable policy to the firm holds when the profits were computed *ex ante*. While overbooking seems to benefit the firm's profit, there is always a chance that all reservation holders will show up *ex post* to exercise the option. Although this may be a rarity for large firms that are protected by the law of large numbers, smaller firms should exercise caution when adopting overbooking policies.

2.6 Conclusion

In this paper, we consider the optimal reservation pricing policy that results in the highest revenue. We build an analytic framework where a service firm is modeled as a queue, and reservations are modeled as call-options, i.e., customers pay a non-refundable deposit when they are uncertain about the valuation of consumption in return for a no-wait guarantee. We assume that customers are endowed with rational expectations about the waiting time they will face if they walk in to receive the service on spot, and solve for the RE equilibrium price of service and the deposit for reservation.

We offer three main recommendations to service firms. First, service firms should advance sell their reservations by charging the full price of service as the non-refundable deposit and charging zero strike price. This results in an socially efficient outcome in that customers with non-negative *ex post* valuation will show up and receive the service. However, such high non-refundable deposit may not be warranted when customers have heterogeneous expected valuation, in which case the firm should sell partially-prepaid reservations. Second, we show that reservation customers should be given discounts compared to the price of service charged to walk-in customers to encourage them to purchase the option despite having uncertainty about their future valuation at the point of purchase. Third, by solving for the optimal booking limit for reservations, we show that it is optimal to book less capacity for reservations as the potential market size increases. In particular, when the market size becomes greater than a certain threshold, it is optimal for the service firm to operate as a pure walk-in system, and thus, sell all its capacity on the spot.

The recommendations made above have solid theoretical support, but there are additional considerations when it comes to implementing them in practice. How would the customers feel about paying such a high non-refundable deposit to make reservations and being charged a different price from other customers? To address these issues there are multiple directions to take. One way is to model the interaction between the firm and customers as a repeated game, where the firm has to consider future customer goodwill. Another broad topic is to incorporate hidden information and reputation concerns. For example, when service valuation is unknown, the reservation policy can be a useful signaling device. Finally, we can incorporate behavioral concerns such as loss aversion and fairness. Even in the simplest one-shot settings, customers may not respond to price discrimination and reservation deposits in the perfectly rational manner as modeled in this paper. Future work can explore how to fine-tune our recommendations.

Chapter 3

Workload Inequality and Fairness Concerns in Service Firms

3.1 Introduction

Service firms often hire multiple workers to supply necessary capacity, but difference in server ability leads to variance in quality. When customers can observe the difference in server quality, they tend to crowd the higher quality sever, and the better workers end up being overloaded with more customers. According to statistics provided by Korean labor organization, workers with higher ability tend to receive more work and burn out more easily than those with lower ability in the same hierarchical rank do (Kim, 2013). This disparity in workload triggers the good workers' fairness concerns and lead them to work less hard. When customers also care for equality and dislike receiving lower quality service than others, crowding and worker demotivation for higher quality server exacerbate.

In education, for example, class size inequality causes envy among teachers (Symour, 1999), and larger class size lowers teacher morale (Glass, 1982). In college, students can easily observe each others' performance and are subject to high level of competition and

fairness concerns. Since instructors' teaching quality is easily shown to the students, when students are given freedom to choose their instructors, better instructors end up with larger classes. Overloaded instructors then may find an incentive to deliberately put less effort to avoid having too much students. "And yes, if I could figure out how to get students to drop my class, I probably would," writes an instructor that teaches math to college students in Texas on her blog, complaining that her class size is much larger compared to other small classes.

In this paper, we consider two ways to deal with worker demotivation due to fairness concerns on workload: eliminating and compensating. A firm can eliminate inequality by distributing customers fairly between workers. This can be done by routing customers evenly or setting the same capacity limit across different workers. On the other hand, the firm can let the inequality in workload persist, but compensate the workers for having more work, which is done by paying workers piece rate. We compare the two remedies in two metrics, expected customer utility and expected quality, to give recommendations under different operational environment.

More specifically, we answer the following questions by modeling the interaction between customers and two workers with different abilities. (1) How does worker fairness concern and workload inequality affect expected customer utility and average quality? (2) What is the best mode of operation that deals with negative effects of worker fairness concerns due to workload disparity and leads to high customer utility and quality? (3) What happens when customers dislike inequality in service quality? What mode of operation should firms choose when customers care for fairness as well?

Our results show that workload disparity leads better workers to work less hard because they will become overcrowded when they provide higher quality service. Eliminating inequality in workload helps fix the situation and provides higher expected customer utility and average quality, but compensating workers per customers they serve works even better.

However, when customers also care for equality in service quality, congestion externality for the higher quality server exacerbates, and a remedy that leads to both high customer expected utility and average quality no longer exists. In this case, quality-utility trade-off exists, and paying workers piece rate leads to high expected quality but low customer expected utility by increasing service quality difference. In this case, distributing workload fairly between workers emerges as an option, which provides reasonable quality without sacrificing customer utility too much.

As far as the authors are aware of, this is the first paper to model fairness concerns of workers in service firms. Usually, papers that consider fairness in congestion prone environment talks about customer fairness concerns about service speed or priority in service. However, our paper considers fairness in workload distribution for workers and fairness in service quality for customers. Variability in service quality due to inherent difference in server ability is a well known problem to operations managers. We believe that this paper can be the first step to model fairness concerns of decision makers in service processes and provide recommendations for service firms based on analytical results.

3.2 Literature Review

This paper relates to literature on fairness concerns of decision makers in economic games. Earlier researches in this stream of literature were experimental works, which showed that people care about fairness so that they may deviate from what standard game theory predicts to achieve more equitable outcomes in various economic games. Kahneman et al. (1986) demonstrate three experiments to show people's are willing to enforce fairness even at some cost to themselves. Güth and Tietz (1990) and Camerer and Thaler (1995) show in bilateral ultimatum experiments that people agree on more equitable outcomes than what maximizes their material payoff. Charness and Rabin (2002) design a range of simple ex-

periment games that show social preferences such as difference aversion models and social welfare models. Many theoretical models were also demonstrated to explain the above empirical results. Just to name a few, Rabin (1993) introduces fairness equilibrium, which reflects people's social preference to help those who help them and hurt those who hurt them. Fehr and Schmidt (1999) introduce inequality aversion to explain how people exploit their market power in competitive markets but not in bilateral bargaining situations. Bolton and Ockenfels (2000) demonstrate an equity, reciprocity, and competition (ERC) model, constructed on the premise that people are motivated by both their pecuniary payoff and their relative standing. These papers study fairness concerns between agents that are playing bilateral games which do not represent the workers' fairness considerations on workload in their workplace.

There is a literature on workers' fairness concerns among peer workers toward their wage, workload, and compensation, to which our work relates more closely. Singh (1995) conducts laboratory experiments to show what people think is a fair allocation of pay between workers based on relative workload, and showed that the fair pay function had a sigmoidal shape characterized by floor and ceiling effects, alluding wage compression. Clark and Oswald (1996) show empirical evidence that job satisfaction is dependent on relative income, and negatively correlated with reference income. Charness and Kuhn (2007) test whether relative wage affects workers' effort levels in a simple gift-exchange labor market and show that workers do not protest underpayment relative to a coworker by withdrawing effort. Frank (1984) examines empirical evidence that when workers are free to choose their coworkers, the competitive wage rates vary substantially less than do individual productivity values. In their paper, relative income status is treated as a good that can either be purchased at the expense of lower income or sold for higher income. Lazear (1989) shows, by a theoretical model, that wage compression is efficient because it reduces uncooperative behavior of workers. In his paper, there is a trade-off between lower in-

centives for hard work and reduction in uncooperative behavior of workers when the firm compresses the wage dispersion. Mirrlees (1971) determines the optimal redistribution of income among workers in the form of taxation. Here, conflict between distribution and incentives arise because the social planner lacks perfect information about the abilities of individuals. Meyer and Mookherjee (1987) design a system of rewards for productive effort that balances distributive considerations with the provision of effort incentives under the uncertainty in production processes and a moral hazard problem. Goel and Thakor (2006) seek for the optimal contract when agents envy one another and show that there are both positive and negative effects of envy to the principal's profit: It decreases the agents' utility, thereby increases the reservation wage, but motivates hard work, and increases productivity. Our work is related to this literature in that it models horizontal fairness concerns among peer workers on workload, and compares between fair distribution of workload and paying incentives to find a way to mitigate the fairness concerns. However, our model differs significantly from any of the above papers since we model a three-party interaction among customers, workers, and a service firm, while customers pose negative externality to one another by making the server crowded and workers cause envy to each other.

Another stream of literature our work has connection to is that on fairness in queues because we are interested in fairness concerns of people involved in service settings which are prone to congestion. Larson (1987), Schmitt et al. (1992), and Rafaeli et al. (2003) study the psychology of customers waiting in line towards fairness in the order they receive service. They talk about how seeing others wait a shorter line is perceived as unfair by customers. Rafaeli et al. (2003) test whether waiting in line should be viewed as a social system with norms or whether behavior in queue can be explained solely by individual's personal interests and cost vs. benefit calculations. Rafaeli et al. (2003) conducts four experiments to see whether technical and physical features of a waiting process can produce perception of unfairness, independent of the actual time-waited experience. Our paper is re-

lated to these works in that we are interested in the participants' behaviors that are affected by their perception on fairness. Also, there are more technical papers that are interested in measuring how fair a queue is based on the order customers are served. Shreedhar and Varghese (1995), Wierman and Harchol-Balter (2003), Avi-Itzhak and Levy (2004), and Raz et al. (2004) propose different measures of order of service fairness in queues. All papers mentioned above, either psychologically or mathematically, study the fairness concerns of customers that rise due to the order they are served or the time they need to wait compared to others. Our work work is different in that we are interested in modeling the fairness concerns of not only the people that are waiting to receive service but also those of the servers that have different workload based on the number of customers they need to serve.

3.3 Model

Service Firm: A monopolistic service firm hires two servers to provide service to customers. In our base model, the firm pays its two servers a fixed equal wage, regardless of their performance. The firm faces a fixed demand, Λ.

Servers: There are two servers of different types, high and low, where the types represent the ability of the workers. We use i to denote the type of a server, $i \in \{H, L\}$. Type-i server has a cost of effort, c_i, where $c_H < c_L$, i.e., the high-type server has lower cost of effort than the low-type server. We use the term effort and quality interchangeably throughout the paper, assuming that higher effort results in higher quality for each server, and denote type-i server's service quality as q_i. We call the traffic of customers served by type-i worker λ_i, and define α as the proportion of customers that are served by the high-type server, i.e., $\lambda_H = \alpha\Lambda$ and $\lambda_H + \lambda_L = \Lambda$. Since servers are paid equally, if the workload between the two is different, the server with higher workload feels unfair and incurs cost δ per addi-

tional customer she serves. We call δ a fairness parameter of the workers, and it captures the degree of workers' aversion against unfairness.

When type-i worker exerts effort level of q_i and serves λ_i customers, then her utility is given as

$$u_i = q_i - \frac{1}{2}c_i q_i^2 + \xi\lambda_i - \delta(\lambda_i - \lambda_{-i})^+,$$

where $-i$ is a subscript used to denote the remaining server other than type-i server. Note that a server cares about her customers' satisfaction, q_i, and has an increasing marginal cost of effort, $\frac{1}{2}c_i q_i^2$. ξ represents a cost or a reward to the workers associated with workload. We assume that $\xi = \kappa - \eta$, where $\kappa \geq 0$ is a reward to the worker per unit of workload. For example, if the firm pays the workers incentive, κ, per customer they serve, then κ represents the piece rate. On the other hand, $\eta \geq 0$ represents the cost workers incur per unit of workload. For simplicity, we assume that $\eta = 0$ throughout the paper, and for our base model, κ also equals zero because there is no additional reward paid to the workers per customer served. Therefore, $\xi = \kappa - \eta = 0$ for our base model.

Customers: Customers who receive service from type-i server enjoy utility of q_i, but incur cost from crowding. The cost of crowding at each server linearly increases in the volume of customers receiving service from the server. The utility function of a customer served by type-i worker is then given as

$$u_i^c = q_i - c\lambda_i,$$

where c is the crowding cost each customer imposes to other customers that are served by the same server.

Equilibrium: We call the base case of our model a "No Intervention model" and label it using a superscript N since the firm does not intervene with the interactions between the workers and the customers. Workers then decide their quality levels, while customers decide which server to receive service from. Workers and customers choose their actions

simultaneously to maximize their utility, and the equilibrium is defined as the following.

Definition 9 *The equilibrium of the simultaneous game between the workers and the customers,* (q_H^N, q_L^N, α^N) *satisfies the following.*
(i) Customers' strategy: Given q_H^N *and* q_L^N*,*

$$\begin{cases} q_H^N - c\alpha^N\Lambda = q_L^N - c(1-\alpha^N)\Lambda, & \text{if } q_H^N - q_L^N < c\Lambda, \\ \alpha^N = 1, & \text{if } q_H^N - q_L^N > c\Lambda. \end{cases} \tag{3.1}$$

(ii) Type-i *server's strategy: Given* α^N *and* q_{-i}^N*,*

$$q_i^N \in \arg\max_{q_i} \; q_i - \frac{1}{2}c_i q_i^2 - \delta(\lambda_i - \lambda_{-i})^+, \tag{3.2}$$

where $\lambda_H = \alpha\Lambda$ *and* $\lambda_L = (1-\alpha)\Lambda$*.*

Measurements of Interest: After solving for the equilibrium, we compute customers' expected utility, U, which is defined as

$$U = \alpha u_H^c + (1-\alpha)u_L^c. \tag{3.3}$$

Also, we are interested in the expected quality, Q, defined as

$$Q = \alpha q_H + (1-\alpha)q_L. \tag{3.4}$$

3.4 Analysis

In this section, we provide analysis for the model to show a negative effect of worker fairness concerns on service quality and consider two ways to alleviate the effect. We first show the expected customer utility and expected quality when workers do not have

any fairness concern to provide benchmark. Then we show the equilibrium results under the base model, where workers are paid equal fixed wage, the elimination remedy, where workload is distributed fairly, and the compensation remedy, where workers are paid a piece rate. Then in section 3.4.5, we compare the different modes of operation based on the two performance measures, expected utility and quality, to find the best one.

3.4.1 No Fairness Benchmark

Before we analyze the effect of fairness concerns of workers, we introduce the case where no worker feels any fairness concern even in the presence of workload inequality as a benchmark. The underlying model for this case is equivalent to the base model introduced in the previous section with the worker fairness concern, δ, being zero, so that the utility function for the i-type worker is becomes

$$u_i = q_i - \frac{1}{2} c_i q_i^2.$$

Note that the utility function no longer depends on the difference of workload, $\lambda_i - \lambda_j$, and the equilibrium outcome can be obtained by solving for (q_H, q_L, α) as in Definition 9, which is summarized as the following.

Lemma 10 *Assume that the firm pays workers fixed equal wage and that customers choose their servers selfishly. When the workers do not have any fairness concern on workload, i.e., $\delta = 0$, then the equilibrium, (q_H^0, q_L^0, α^0), the resulting customer expected utility, U^0,*

and the expected quality, Q^0, satisfy the following.

$$\begin{aligned}
\alpha^0 &= \min\left\{\frac{1}{2} + \frac{1}{2c\Lambda}\left(\frac{1}{c_H} - \frac{1}{c_L}\right), 1\right\}, \\
q_H^0 &= \frac{1}{c_H}, \\
q_L^0 &= \frac{1}{c_L}, \\
U^0 &= \frac{1}{2}\left(\frac{1}{c_H} + \frac{1}{c_L}\right) - \frac{c}{2}\Lambda, \\
Q^0 &= \frac{1}{2}\left(\frac{1}{c_H} + \frac{1}{c_L}\right) + \frac{1}{2c\Lambda}\left(\frac{1}{c_H} - \frac{1}{c_L}\right)^2.
\end{aligned}$$

Note that when workers do not have fairness concern on workload, they choose their level of effort so that the marginal return from effort equals the marginal cost of effort, and their quality decisions do not depend on workload disparity. More customers choose to be served by higher quality server because the higher quality allows the customers to overcome higher cost of crowding.

Note that when the cost of effort for the two workers differ greatly so that the quality difference, $\frac{1}{c_H} - \frac{1}{c_L}$, exceeds a threshold, $c\Lambda$, then no customer chooses the low quality server. This happens when the market size, Λ, is small so that the high quality server alone has enough capacity to serve the entire market. Also, when customers do not care much about crowding with the cost of crowding, c, being negligible, then customers will decide their server based heavily on service quality and will not choose the low quality server. In this paper, however, we only consider the more interesting case where both servers are in service, i.e., $\frac{1}{c_H} - \frac{1}{c_L} < c\Lambda$.

3.4.2 A Model of Fairness Concerns Among Workers in Service Firms

We now show how worker fairness concern combined with workload disparity can lead the high-quality worker to work less hard. In order to do so, we turn to our base model with

non-zero worker fairness concern parameter, $\delta > 0$. Recall that in this case, workers are paid fixed equal wage, while customers choose servers selfishly based on service quality and cost of crowding. The resulting service quality are summarized in the following lemma.

Lemma 11 *When the firm pays workers fixed equal wage and customers choose their servers selfishly, then the equilibrium, (q_H^N, q_L^N, α^N), the resulting expected customer utility, U^N, and the average quality, Q^N, satisfy the following.*

$$\alpha^N = \min\left\{\frac{1}{2} + \frac{q_H^N - q_L^N}{2c\Lambda}, 1\right\}, \tag{3.5}$$

$$q_H^N = \begin{cases} \frac{1}{c_H}\left(1 - \frac{\delta}{c}\right), & \text{if } \delta < c\left(1 - \frac{c_H}{c_L}\right), \\ \frac{1}{c_L}, & \text{if } \delta \geq c\left(1 - \frac{c_H}{c_L}\right), \end{cases} \tag{3.6}$$

$$q_L^N = \frac{1}{c_L}, \tag{3.7}$$

$$U^N = \begin{cases} \frac{1}{2}\left(\frac{1-\delta/c}{c_H} + \frac{1}{c_L}\right) - \frac{c}{2}\Lambda, & \text{if } \delta < c\left(1 - \frac{c_H}{c_L}\right), \\ \frac{1}{c_L} - \frac{c}{2}\Lambda, & \text{if } \delta \geq c\left(1 - \frac{c_H}{c_L}\right), \end{cases}$$

$$Q^N = \begin{cases} \frac{1}{2c\Lambda}\left(\frac{1-\delta/c}{c_H} - \frac{1}{c_L}\right)^2 + \frac{1}{2}\left(\frac{1-\delta/c}{c_H} + \frac{1}{c_L}\right), & \text{if } \delta < c\left(1 - \frac{c_H}{c_L}\right), \\ \frac{1}{c_L}, & \text{if } \delta \geq c\left(1 - \frac{c_H}{c_L}\right). \end{cases}$$

Note that the high-type worker's service quality when she cares about fairness in workload, q_H^N, is lower than that under the case where she does not care about fairness, q_H^0. This is because more customers crowd at the high quality server as her service quality becomes higher compared to the quality at the low-type server. Thus, in order to avoid being overcrowded H-type worker has an incentive to lower her service quality when she is concerned about fairness in workload. This is captured by the term $-\frac{\delta}{c}$ in (3.6). When the cost of crowding, c, is high, less customers will choose the H-type server despite the high

quality so that the H-type worker decreases her service quality by smaller amount. As the fairness concern of the high-quality worker increases, i.e., as δ increases, the high-quality worker has higher incentive to lower her service quality. Moreover, when the high-quality worker's fairness concern becomes very large, then she will choose her service quality as low as the service quality of the low-type worker so that the workload becomes exactly equal between the two servers. Note that the resulting expected customer utility, U, and the average quality, Q, are both lower when workers have fairness concerns than when they do not, i.e., $U^N < U^0$ and $Q^N < Q^0$.

3.4.3 Impact of Workload Inequality on Work Performance

In the previous section, we have seen that when workers are paid the same wage and customers choose their servers selfishly, the high-type worker's service quality drops due to fairness concern on workload. In this section, we consider one remedy to remove the negative effect of fairness concern on service quality: "eliminating" inequality in workload. This can be done by distributing the equal number of customers to the two workers regardless of their service qualities. We call this mode of operation a "fair workload distribution policy" and use F to denote "Fair distribution". An example of this practice in the real world is the equal class enrollment capacities for college core classes taught by multiple instructors in parallel regardless of the instructors' quality.

Under this intervention, customers do not make any decision, but the firm commits to route customers equally between the two workers independent of their service values. The resulting quality levels are not an equilibrium outcome because once the quality level is set by the workers, the firm has an incentive to route more customers to the higher quality server. However, given that the firm can credibly commit to distribute customers equally between workers, $\alpha = \frac{1}{2}$, type-i worker's decision is to choose her service quality that

satisfies

$$q_i^F \in \arg\max_{q_i} \; q_i - \frac{1}{2} c_i q_i^2.$$

The outcome under fair workload distribution scheme is summarized as the following.

Lemma 12 *Let α^F, q_i^F, U^F, and Q^F denote the proportion of customers that are being served by the H-type worker, service quality level of the i-type worker, the expected customer utility, and the average quality, respectively, when the firm commits to set equal capacity limit for the two workers. Then*

$$\begin{aligned} \alpha^F &= \frac{1}{2}, \quad q_H^F = \frac{1}{c_H}, \quad q_L^F = \frac{1}{c_L}, \\ U^F &= \frac{1}{2}\left(\frac{1}{c_H} + \frac{1}{c_L}\right) - \frac{c}{2}\Lambda, \\ Q^F &= \frac{1}{2}\left(\frac{1}{c_H} + \frac{1}{c_L}\right). \end{aligned}$$

When equal workload is given to the workers regardless of their service qualities, the high-quality server no longer has any reason to lower her quality since lowering quality does not reduce workload. Since the firm promised to fix the workload for each worker to half the total workload, the workload term λ_i becomes a constant in type-i worker's utility function, and the worker chooses the quality level that maximizes $q_i - \frac{1}{2} c_i q_i^2$ without worrying about disutility from unequal workload distribution. Thus, there is no incentive for the workers to reduce the quality level below the level where marginal cost of effort equals the marginal utility from delivering quality. Thus, the quality level for the high-quality worker $q_H^F = \frac{1}{c_H}$ is equal to that without worker fairness concerns, $q_H^0 = \frac{1}{c_H}$, which is strictly higher than that under the no intervention model, $q_H^F = \frac{1}{c_H}\left(1 - \frac{\delta}{c}\right)$.

3.4.4 How to Mitigate Worker Fairness Concerns

In the previous section, we saw how a firm can alleviate workers' incentive to lower service quality resulting from fairness concern on workload by eliminating inequality in workload. In this section, we consider an alternative way: compensating for work. In this case, the firm no longer controls the distribution of customers between workers directly, but pays each worker incentive per customer she serves. Customers selfishly decide which server to receive service from based on the service quality and the cost of crowding for each server. We assume that the incentive the firm pays workers per head, $\kappa > 0$, is a fixed constant. Since serving more customers adds positive utility to workers, they no longer feel unfair about serving more customers, i.e, $\delta = 0$.

Given customers' decision, α, the equilibrium (q_H^I, q_L^I, α^I) is defined as in Definition 9, except for the fact that utility for type-i worker is given as

$$q_i - \frac{1}{2}c_i q_i^2 + \kappa\lambda_i.$$

The superscript I represents "incentives".

Solving for the equilibrium gives the following results.

Lemma 13 *When the firm pays piece rate to workers, and the customers choose their workers selfishly to maximize their utilities, then the equilibrium,* (q_H^I, q_L^I, α^I)*, and the*

resulting expected customer utility, U^I, and average quality, Q^I, satisfy the following.

$$\begin{aligned}
\alpha^I &= \min\left\{\frac{1}{2} + \frac{q_H^I - q_L^I}{2c\Lambda}, 1\right\} \\
q_H^I &= \frac{1}{c_H}\left(1 + \frac{\kappa}{2c}\right), \quad q_L^C = \frac{1}{c_L}\left(1 + \frac{\kappa}{2c}\right), \\
U^I &= \frac{1}{2}\left(1 + \frac{\kappa}{2c}\right)\left(\frac{1}{c_H} + \frac{1}{c_L}\right) - \frac{1}{2}c\Lambda, \\
Q^I &= \frac{1}{2}\left(1 + \frac{\kappa}{2c}\right)\left(\frac{1}{c_H} + \frac{1}{c_L}\right) + \frac{1}{2c\Lambda}\left(1 + \frac{\kappa}{2c}\right)^2\left(\frac{1}{c_H} - \frac{1}{c_L}\right)^2.
\end{aligned}$$

Note that when workers are paid piece rate, there is no longer any incentive for the workers to lower their service quality due to fairness concerns. Oppositely, there is now incentive for the workers to put higher effort to attract more customers, and worker-i's service quality increases by $\frac{1}{c_i}\frac{\kappa}{2c}$. As the cost of crowding, c, for customers increases this incentive is reduced because customers will be less likely to crowd for higher quality server, and the effect of quality on customers' decisions will be reduced. Note that the positive effect of compensation on service quality not only affects the high-quality server but also increases the low-quality server. The resulting service qualities for workers, (q_H^I, q_L^I), are higher than those without any intervention, (q_H^N, q_L^N).

3.4.5 Comparing the Modes of Operation

In this section, we compare the three modes of operation, (N) no intervention, (F) fair distribution of workload, and (I) paying workers incentive per customer served, to find a way to deal with worker demotivation due to workload disparity. The measures we are interested in are customer expected utility, U, and expected quality, Q. We examine which mode of operation works well in terms of these measurements under what circumstances and why. The following proposition summarizes the results.

Proposition 14 (Expected Customer Utility) *Let U^N, U^F, and U^I be customer expected*

utility when the firm does not intervene, distributes customers equally between workers, and pays workers piece rate, respectively. Then

$$U^N < U^F < U^I,$$

for all $\delta > 0$.

Proposition 14 shows that when workers dislike having higher workload compared to their colleagues, then 1) just distributing workload equally between workers without changing the payment scheme increases customer satisfaction level, and 2) when the firm pays workers incentive per customer they serve, then customer expected utility becomes even higher.

When the firm eliminates the workload inequality by allocating the same number of customers between the two workers, then the *ex ante* customer utility increases compared to that under the no-intervention case. This is interesting because merely redistributing workload without incurring additional cost increases customer satisfaction. This is due to the fact that by allocating workload equally between workers, the high-quality server's incentive to lower service quality to avoid higher workload goes away. Moreover, equally distributing customers between workers is of additional benefit to overall customer utility because it minimizes the total congestion externality customers experience by avoiding heavy crowd in higher quality server.

Compensating for workload inequality by paying workers piece rate while letting customers and workers choose their own actions selfishly results in the highest customer expected utility when workers are fairness concerned. When workers are paid incentive based on workload, higher workload becomes a positive thing tied to higher payment, and fairness concern on workload disappears. Thus, high-quality worker puts even more effort in quality to attract more customers. In addition, when workers are paid piece rate, the low-quality worker also puts higher effort than she would do without any intervention or under

the fair workload distribution policy.

From the above proposition, we showed that when workers dislike serving more customers than peer workers, then paying workers incentive per customer brings the highest customer expected utility. Now, we turn to expected quality levels and see whether compensating servers based on customers they serve work equally well in terms of quality as well. We compare the expected quality levels under the three modes of operation in the following proposition.

Proposition 15 (Expected Quality) *Let Q^N, Q^F, and Q^I be the expected quality when the firm does not intervene (N), distributes customers equally between workers (F), and pays workers incentive per customer they serve (I), respectively. Then for any given $\kappa > 0$*

$$\begin{aligned} &Q^F < Q^N < Q^I, \quad \textit{if } 0 < \delta < \tilde{\delta}, \\ &Q^N < Q^F < Q^I, \qquad \textit{if } \delta > \tilde{\delta}, \end{aligned}$$

where $\tilde{\delta}$ is given as

$$\frac{1}{\Lambda}\left\{\left(\frac{1}{c_H}-\frac{1}{c_L}\right)^2 - 2\left(\frac{1}{c_H}-\frac{1}{c_L}\right)\frac{\tilde{\delta}}{c_H c} + \frac{\tilde{\delta}^2}{c_H^2 c^2}\right\} = \frac{\tilde{\delta}}{c_H},$$

Before we compare the expected quality levels under different modes of operation, it is worth mentioning two factors that lead to higher expected quality. First, having higher individual service qualities, (q_H, q_L), from the workers results in higher average quality. When the individual quality vector, (q_H, q_L), is large, then it is likely that the expected quality, $Q = \alpha q_H + (1-\alpha)q_L$ is large. Second, average quality level becomes higher when more customers receive service from high quality service, i.e., when the high-quality server has higher exposure to the customers. This can be seen again in the average quality formula, $Q = \alpha q_H + (1-\alpha)q_L$, which increases as α increases.

From Proposition 15, we see that compensating for workload results in the highest expecte quality compared to other two policies. This is because paying workers piece rate increases the individual quality levels, and letting customers choose their servers selfishly increases the exposure of the high-quality server to the customers. Thus, the two reasons for higher average quality mentioned previously satisfy under the piece rate policy, and results in high average quality.

When we compare between the case without any intervention and the case where workload is equally distributed, we need to consider a trade-off between high individual quality levels (high (q_H, q_L)) and high exposure to high-quality server (high α). When the firm does not intervene, individual quality levels are lower, but customer exposure(congestion) to high-quality server is higher. When the firm distributes workload fairly between workers, individual quality levels are higher, but customer congestion at the high-quality server is lower. When worker fairness concern is negligible, i.e., when δ is small, individual quality decrease due to worker fairness concern becomes negligible compared to the quality increase due to high exposure under the no intervention policy. Thus, expected quality is higher when the firm does not intervene and lets the customers choose their own server. However, as worker fairness concern grows, quality decrease overwhelms the benefit of higher exposure of the high-quality service, and fairly distributing workload results in higher expected customer utility.

From the above comparisons in Proposition 14 and Proposition 15, we showed two facts. When workers highly dislike having higher workload than others, then eliminating the worker fairness concern by distributing workload equally between workers increases customer expected utility and expected quality. However, compensating for workload by paying workers incentive per customer they serve results in even higher customer expected utility and quality because having higher workload becomes a positive thing to workers when they are paid per customer they serve.

Note that according to the results we have seen so far, paying workers piece rate and letting customers choose their servers is dominant over all other policies both in terms of utility and quality when workers dislike unequal distribution of workload. However, many service firms, follow other modes of operation besides compensating workers per customer served. In the following section, we answer when these other modes of operation can be a good strategy to consider.

3.5 Interaction with Customer Fairness Concerns

So far, we have looked at how workers' fairness concern on workload distribution affects customer expected utility and average quality. We found ways to alleviate workers' incentive to lower service quality due to fairness concerns. We showed that when worker fairness concern is high, then eliminating workload inequality can increase both average customer utility and average quality. We also showed that customer expected utility and expected quality can increase even further when workers are compensated for workload. Since paying piece rate changes workers' perception on high workload from a negative thing to a positive thing, it results in higher individual service qualities and performs well in both customer utility and average quality. Moreover, since paying piece rate results in the highest performance measures regardless of the parameter values, it seems that paying piece rate is dominant over other two modes of operation.

However, in this section, we introduce customer fairness concerns resulting from unequal service quality and see how customer expected utility and average quality are affected by both worker fairness concerns and customer fairness concerns. We show that when customers care about fairness too, customers may prefer lower but equal quality services over higher but unequal quality services. Therefore, choosing a mode of operation that results in lower but more even individual qualities may be necessary. This leads other modes of

operation besides paying workers piece rate to be preferred mechanism under certain circumstances. Eventually, we make recommendations on what policy a firm should use under what circumstances.

Now we assume that not only workers but also customers dislike being treated unequally. Worker fairness concern results from unequal workload distribution, whereas customer fairness concern stems from difference in service quality, i.e., customers dislike receiving lower quality service than other customers do. Let us denote the degree customers dislike receiving lower quality service using γ, and call it a "customer fairness concern parameter". Customer utility is negatively affected by the difference between service quality the customer receives and that of what other customers receive when she receives an inferior service than some others do. Then we can define the utility of a customer that receives service from type-i worker as

$$u_i = q_i - c\tilde{\lambda}_i - \gamma(\tilde{q}_j - \tilde{q}_i)^+,$$

where we use the expression $\tilde{(\cdot)}$ to represent equilibrium results under the case where customers also care about fairness to distinguish them from the results under the case where only workers have fairness concerns. In the following sections, we first show how customer fairness concern makes the high-quality worker even more crowded by analyzing the equilibrium outcomes under no intervention policy. Then we compare the two remedies, eliminating and compensating, to find the most effective mechanism under different operating conditions.

3.5.1 The Effect of Customer Fairness Concerns in Service Quality

We first consider the no intervention policy where the firm pays workers a fixed equal wage independently of their performance and lets customers choose their servers freely.

The equilibrium under this policy is defined similarly as in Definition 9, but needs some adjustment due to the change in customer utility.

Definition 16 *When both workers and customers have fairness concerns and the firm pays workers equally while letting customers choose their servers selfishly, then the equilibrium of the simultaneous game between the workers and the customers,* $(\tilde{q}_H^N, \tilde{q}_L^N, \tilde{\alpha}^N)$ *is defined as the following.*

(i) Customers' strategy: Given $\tilde{q}_H^N$ *and* $\tilde{q}_L^N$*,*

$$\begin{cases} \tilde{q}_H^N - c\tilde{\alpha}^N\Lambda = \tilde{q}_L^N - c(1-\tilde{\alpha}^N)\Lambda - \gamma(\tilde{q}_H^N - \tilde{q}_L^N)^+, & \text{if } \tilde{q}_H^N - \tilde{q}_L^N < \frac{c\Lambda}{1+\gamma}, \\ \tilde{\alpha}^N = 1, & \text{if } \tilde{q}_H^N - \tilde{q}_L^N > \frac{c\Lambda}{1+\gamma}. \end{cases} \tag{3.8}$$

(ii) Type-i *server's strategy: Given* $\tilde{\alpha}^N$ *and* $\tilde{q}_{-i}^N$*,*

$$\tilde{q}_i^N \in \arg\max_{q_i} \; q_i - \frac{1}{2}c_i q_i^2 - \delta(\tilde{\lambda}_i - \tilde{\lambda}_{-i})^+, \tag{3.9}$$

where $\tilde{\lambda}_H = \tilde{\alpha}\Lambda$ *and* $\tilde{\lambda}_L = (1-\tilde{\alpha})\Lambda$*.*

Solving for the equilibrium gives the following equilibrium.

Lemma 17 *When both workers and customers have fairness concerns and the firm pays workers equally while letting customers choose their servers selfishly, then the equilibrium of the simultaneous game between the workers and the customers,* $(\tilde{q}_H^N, \tilde{q}_L^N, \tilde{\alpha}^N)$ *satisfies*

the following.

$$
\begin{aligned}
\tilde{\alpha}^N &= \min\left\{\frac{1}{2} + \frac{(1+\gamma)(\tilde{q}_H^N - \tilde{q}_L^N)^+}{2c\Lambda}, 1\right\}, \\
\tilde{q}_H^N &= \begin{cases} \frac{1}{c_H}\left\{1 - \frac{\delta(1+\gamma)}{c}\right\}, & \text{if } \delta < \frac{c}{1+\gamma}\left(1 - \frac{c_H}{c_L}\right), \\ \frac{1}{c_L}, & \text{if } \delta \geq \frac{c}{1+\gamma}\left(1 - \frac{c_H}{c_L}\right), \end{cases} \\
\tilde{q}_L^N &= \frac{1}{c_L}, \\
\tilde{U}^N &= \begin{cases} \frac{1}{2}\left\{\frac{1-\delta(1+\gamma)/c}{c_H} + \frac{1}{c_L}\right\} - \frac{\gamma}{2}\left\{\frac{1-\delta(1+\gamma)/c}{c_H} - \frac{1}{c_L}\right\} - \frac{c}{2}\Lambda, & \text{if } \delta < \frac{c}{1+\gamma}\left(1 - \frac{c_H}{c_L}\right), \\ \frac{1}{c_L} - \frac{c}{2}\Lambda, & \text{if } \delta \geq \frac{c}{1+\gamma}\left(1 - \frac{c_H}{c_L}\right), \end{cases} \\
\tilde{Q}^N &= \begin{cases} \frac{1}{2}\left\{\frac{1-\delta(1+\gamma)/c}{c_H} + \frac{1}{c_L}\right\} + \frac{1+\gamma}{2c\Lambda}\left\{\frac{1-\delta(1+\gamma)/c}{c_H} - \frac{1}{c_L}\right\}^2, & \text{if } \delta < \frac{c}{1+\gamma}\left(1 - \frac{c_H}{c_L}\right), \\ \frac{1}{c_L}, & \text{if } \delta \geq \frac{c}{1+\gamma}\left(1 - \frac{c_H}{c_L}\right). \end{cases}
\end{aligned}
\tag{3.10}
$$

The resulting equilibrium is qualitatively similar to the equilibrium outcome in section 3.4.2, where customers do not care about equality. However, worker demotivation due to unequal workload distribution exacerbates due to customer envy when workers are paid equal wage and customers choose their servers selfishly. Note that since customers dislike receiving lower quality service than what other customers receive, more customers tend to crowd in the higher quality server than in the case where customers do not care about fairness, i.e., $\tilde{\alpha}^N = \min\left\{\frac{1}{2} + \frac{(1+\gamma)(q_H^N - q_L^N)^+}{2c\Lambda}, 1\right\} > \min\left\{\frac{1}{2} + \frac{(q_H^N - q_L^N)^+}{2c\Lambda}, 1\right\} = \alpha^N$ given q_H and q_L. Therefore, given a fixed worker fairness parameter, δ, the high type worker has higher incentive to lower her service quality when customers care about equity in service quality than when customers do not mind receiving inferior service to what others receive. Thus, individual service quality from the high-quality server is lower when customers care about fairness than that when customers do not care about equality in service quality.

Note that customer expected utility is lower when customers are concerned about fairness in service quality because in addition to low individual qualities, there is psychological cost for those who receive inferior service. This psychological cost is captured by the term, $-\frac{\gamma}{2}\left\{\frac{1-\delta(1+\gamma)/c}{c_H}-\frac{1}{c_L}\right\}$, in (3.10), and this cost increases as customer fairness concerns, γ, increase. Expected quality, however, may increase due to higher exposure to higher quality service when compared to the case without customer fairness concern.

3.5.2 Impact of Workload Inequality on Work Performance

As in the case without customer fairness concerns, we consider two remedies for exacerbated congestion externalities in the high-quality server. We first consider eliminating the workload inequality by equally distributing customers between workers.

When the workload is set equal for the two workers under the case where customers have fairness concerns on service quality, then the quality levels are determined the same as those under the case without customer fairness concerns. This is because the number of customers that visit the servers no longer depends on the service qualities, and the workers have no incentive to choose other quality levels than those that maximize their utility excluding the psychological cost term due to fairness concern, $\delta(q_i - q_j)^+$. The resulting equilibrium is summarized in the following lemma.

Lemma 18 *Let $\tilde{\alpha}^F$, $\tilde{q}_i^F$, $\tilde{U}^F$, and $\tilde{Q}^F$ denote the proportion of customers that are being served by the H-type worker, service quality level of the i-type worker, customer expected utility, and the expected quality, respectively, when the firm commits to set equal capacity limit for the two workers and customers have fairness concerns on service quality. Then*

$$\tilde{\alpha}^F = \frac{1}{2}, \quad \tilde{q}_H^F = \frac{1}{c_H}, \quad \tilde{q}_L^F = \frac{1}{c_L},$$
$$\tilde{U}^F = \frac{1}{2}\left(\frac{1}{c_H}+\frac{1}{c_L}\right) - \frac{\gamma}{2}\left(\frac{1}{c_H}-\frac{1}{c_L}\right) - \frac{c}{2}\Lambda,$$

$$\tilde{Q}^F = \frac{1}{2}\left(\frac{1}{c_H} + \frac{1}{c_L}\right).$$

Note that when workload is distributed equally between workers, then neither worker fairness concern nor customer fairness concern affect the individual service quality each worker chooses, and thus, does not change the average quality compared to the case without customer fairness concern. Customer expected utility, however, decreases compared to that without customer fairness concerns because there is negative utility from customer envy, $-\frac{\gamma}{2}\left(\frac{1}{c_H} - \frac{1}{c_L}\right)$.

3.5.3 Mitigating Worker Fairness Concerns

As a second remedy to remove worker demotivation from fairness concerns, we consider compensating workers according to their workload. We look at the equilibrium service quality, resulting customer expected utility, and expected service quality when customers have fairness concerns on quality and workers are paid a piece rate, $\kappa > 0$. The equilibrium outcomes are given in the following lemma.

Lemma 19 *When the firm pays piece rate, $\kappa > 0$, to workers, and the customers choose their workers selfishly based on their utility from service and fairness concerns, then the equilibrium, $(\tilde{q}_H^I, \tilde{q}_L^I, \tilde{\alpha}^I)$, and the resulting customer satisfaction level, $\tilde{W}^I$, and average*

quality, $\tilde{Q}^I$, satisfy the following.

$$
\begin{aligned}
\tilde{\alpha}^I &= \min\left\{\frac{1}{2} + \frac{\tilde{q}_H^I - \tilde{q}_L^I}{2c\Lambda}, 1\right\} \\
\tilde{q}_H^I &= \frac{1}{c_H}\left(1 + \frac{\kappa(1+\gamma)}{2c}\right), \quad \tilde{q}_L^I = \frac{1}{c_L}\left(1 + \frac{\kappa(1+\gamma)}{2c}\right), \\
\tilde{U}^I &= \frac{1}{2}\left\{\frac{1}{c_H}\left(1 + \frac{\kappa(1+\gamma)}{2c}\right) + \frac{1}{c_L}\left(1 + \frac{\kappa(1+\gamma)}{2c}\right)\right\} \\
&\quad - \frac{\gamma}{2}\left\{\frac{1}{c_H}\left(1 + \frac{\kappa(1+\gamma)}{2c}\right) - \frac{1}{c_L}\left(1 + \frac{\kappa(1+\gamma)}{2c}\right)\right\} - \frac{c\Lambda}{2}, \\
\tilde{Q}^I &= \frac{1}{2}\left\{\frac{1}{c_H}\left(1 + \frac{\kappa(1+\gamma)}{2c}\right) + \frac{1}{c_L}\left(1 + \frac{\kappa(1+\gamma)}{2c}\right)\right\} \\
&\quad + \frac{1+\gamma}{2c\Lambda}\left\{\frac{1}{c_H}\left(1 + \frac{\kappa(1+\gamma)}{2c}\right) - \frac{1}{c_L}\left(1 + \frac{\kappa(1+\gamma)}{2c}\right)\right\}^2.
\end{aligned}
$$

Note that when workers are paid piece rate, customer fairness concern increases workers' service quality even further. This is due to the fact that customers who dislike receiving inferior service than others are more sensitive to service quality, and workers put more effort in quality to attract these quality-sensitive customers.

3.5.4 Comparing the Modes of Operation

Previously, we showed that paying workers per customer they serve results in both the highest customer expected utility and the highest average quality compared to other modes of operation for all parameter values when workers care about fairness but customers do not. We question why, if that is the case, we observe other modes of operations besides paying servers piece rate in reality. In order to do so, we first show how customer fairness concerns can help or hurt customer satisfaction and average quality under different modes of operations.

Proposition 20 (Customer Fairness Concerns) *Let $\tilde{U}^N$ ($\tilde{Q}^N$), $\tilde{U}^F$ ($\tilde{Q}^F$), and $\tilde{U}^I$ ($\tilde{Q}^I$) be customer expected qualities (expected qualities) when the firm does not intervene (N), dis-*

tribute workload fairly between workers (F), and pays workers incentive based on workload (I), respectively, when both workers and customers care about fairness. Then for any given δ,

there exists $\tilde{\gamma}$ such that for $\gamma > \tilde{\gamma}$,

$$\tilde{U}^N > \tilde{U}^F > \tilde{U}^I.$$

When $\gamma > \tilde{\gamma}$, then $\tilde{U}^N$ increases as δ increases.

Also, there exists $\hat{\gamma}$ such that for $\gamma > \hat{\gamma}$,

$$\tilde{Q}^N < \tilde{Q}^F < \tilde{Q}^I.$$

From this proposition, we can see that when customers care much about fairness, then the mode of operation that gives high average quality results in low customer satisfaction, and vice versa. The order of operational policies from high to low customer expected utility is the reverse of the order from high to low expected quality. Thus, there is a trade-off between quality and customer utility, and no mechanism is unanimously dominant in both performance measures.

For customer expected utility, letting workers and customers choose their actions freely without any intervention is the best mechanism although it results in the lowest individual service qualities. Moreover, when customer fairness concern becomes very high, then the customer expected utility under this best policy, $\tilde{U}^N$, increases as worker envy, δ, increases. This is interesting because as workers' fairness concern, δ, increases, service quality from H-type worker decreases under the no intervention policy, but customers are being more satisfied in expectation by lowering the high-quality server's service quality. In other words, customers fear so much about being the one who receives lower quality service that they would rather remove the possibility to receive higher quality service. That is, customers

prefer eliminating the chance to receive higher quality service so that they do not have to worry about inequality than having a high quality server available but facing positive chance of being a person who receives inferior service. Customers are then willing to sacrifice direct utility from quality in order to avoid any chance of inequality.

Note that paying workers piece rate becomes the worst policy in terms of customer expected utility when customers are highly sensitive about inequality. This is because the policy results in the largest quality gap, $\tilde{q}_H - \tilde{q}_L$, which represents high degree of inequality. Since the high-quality server has more efficiency in effort, if the workers are paid the same piece rate per customer, the high-type worker puts even greater effort than the low quality worker does, and widens inequality in service quality.

Even though paying workers incentive based on workload brings the lowest expected utility to highly envious customers, it remains to deliver the highest average quality to the customers. Recall that there are two ways to achieve high expected quality: 1) to increase individual service qualities, $(\tilde{q}_H, \tilde{q}_L)$, and 2) to expose more customers to higher quality server. Both of these conditions are met when the firm compensates workers for workload. Therefore, $(\tilde{q}_H^I, \tilde{q}_L^I)$ is higher than the individual quality vector under any other policy. Also, because the quality gap, $\tilde{q}_H - \tilde{q}_L$, is the highest under this mechanism, more customers tend to crowd the high-quality server, resulting in high exposure to the high quality service.

Since the order of the three operational policies based on expected customer utility is reverse of the order based on expected quality, we face a situation where we need to decide which performance measure to focus on to choose one operational policy. Depending on the weights we put between utility and quality, the operational policy that is chosen should be different. If the firm solely focuses on quality, paying workers incentive per customer served is good. If it focuses on customer satisfaction level alone, paying workers fixed equal wage while letting customers choose their servers selfishly is good. However, if the firm wants to achieve reasonable level of both quality and customer expected utility, then

distributing workload equally between workers is an option to consider.

Now we can answer why we would see different operational policies are chosen in the real world under various circumstances. Depending on industry types, workers and customers may feel different degree of fairness concerns. Based on the fairness parameters, the operational policy that results in the highest customer satisfaction and the average quality differs. Also, when customers are highly concerned about fairness, the policy one should choose depends on what the industry cares more between expected customer utility and expected quality.

3.6 Discussion

In this section, we talk about why certain service industries adopt certain mode of operation in the real world. Before we talk about examples, we summarize our theoretical findings in figure 3.1. The figure shows that when customer fairness concern is low, then paying workers incentive per customer they serve provides both the highest customer expected utility and the highest expected quality. On the other hand, when customer fairness concern is high, paying workers incentive based on workload results in high expected quality, but suffers from low expected customer utility. In this case, not intervening results in the highest customer satisfaction but delivers poor quality. Equally distributing customers between workers compromises this quality-satisfaction trade-off and delivers reasonable quality without completely ignoring customer satisfaction.

Customer fairness concerns can be high when the server qualities and/or the outcomes of other customers' services are easily observed by customers. Server ranking system that is publicly observable can be of driving factor of high customer fairness concerns as well. Customers seeking for the "best" server, without being satisfied by a good enough service in certain service industries can be a signal of high customer fairness concerns.

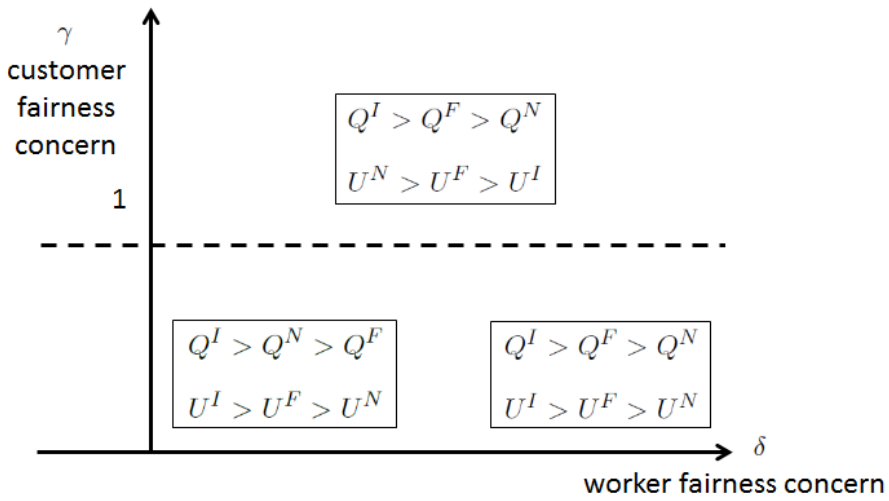

Figure 3.1: *Summary of the findings in Section 3.4 and Section 3.5*

Examples of industry with high customer fairness concerns are health care and education. In the health care industry, patients are often very sensitive about service quality and they seek for the best doctor whose quality can be quantified by the number of successful cases treated or the "scores" the doctor received in a public patient feedback system that shows how smart, kind and responsive he/she is. Also, in college education, students can easily observe teaching quality of instructors and the performance of other students. Students' abilities are judged relatively, and thus, they often care about relative performance even more than their absolute performance. In this competitive environment, students highly prefer being taught by a better instructor.

We consider public hospitals in Korea as an example of service firms that face customers with high fairness concerns and adopt no intervention mechanism. Doctors in these hospitals are paid a fixed equal wage regardless of the number of patients they see, and patients can choose to be treated by a specific doctor if they have preference. Performance measures of doctors are public information, and some patients are happy to wait for months to be seen by the best doctor. It is a common knowledge that many well known doctors tend to spend as little time as possible with patients and often treat them with less care be-

cause they are so busy seeing excessively many patients. Public hospitals in Korea usually target to provide medical coverage to wide range of population, and they care much about patient welfare from fairness. These public hospitals' goal need not be to provide highest quality to the patients. Therefore, adopting no intervention policy can be the best mode of operation for these hospitals since it results in high customer expected utility by providing lower but equal quality of service.

On the other hand, private hospitals can have a different business model. Private hospitals care for providing the patients high quality service for profit rather than giving equal opportunities of treatment to patients. Thus, paying doctors per case they treat and letting patients choose their own doctors can keep the doctors motivated for high quality service. Quality is usually higher at private hospitals than at public hospitals, and high quality doctors have long wait-lists, representing high congestion externality.

In education, the goal is usually to provide quality while keeping the students satisfied. Also, giving students equal opportunity and treating them fairly is an important factor in education. Thus, schools often distribute workload equally among instructors by setting the same class enrollment capacities regardless of the instructors' teaching quality. By doing this schools balance between quality and satisfaction from fairness in the presence of high student fairness concerns. For example, college core classes that are taught by multiple instructors in parallel have equal class enrollment capacities regardless of the instructors' abilities.

Other service firms whose customers do not have such high fairness concerns such as hair salons and law firms usually pay their servers per customer they serve, while letting customers choose their servers freely. This keeps the servers motivated to provide high quality, and customers are satisfied from having freedom to choose their own servers based on service qualities and congestion levels.

3.7 Conclusion

In this paper, we look at ways to deal with fairness concerns of decision makers in service processes. Especially, we are interested in the case where a service firm hires multiple workers to provide enough capacity, and experiences variability in service quality due to inherent difference in server ability. Servers dislike having more work than peer workers do and show incentive to lower service quality when they face higher customer volume due to higher ability.

We consider two ways to deal with worker demotivation due to fairness concerns on workload: eliminating and compensating. Inequality can be eliminated by distributing equal number of customers between workers regardless of their service quality. On the other hand, the firm can pay workers incentive per customer they serve to encourage workers to put high effort to attract more customers. We use two performance measures, expected customer utility and expected quality, to compare the two remedies.

We then answer the following questions using the equilibrium results of our model. (1) How does worker fairness concern affect the expected customer utility and the average quality? (2) Which remedy works better to deal with worker demotivation due to fairness concern? (3) What happens when customers also have fairness concerns on service quality? What mode of operation should firms choose when customers care for fairness as well?

Our results show that when workers care for fairness in workload distribution, the good worker works less hard because otherwise, she will become overloaded. Eliminating inequality by routing equal number of customers between the workers helps increase both expected customer utility and average quality, but paying workers incentive per customer they serve works even better. This is because paying piece rate to workers turns workload inequality from negative thing to a positive thing resulting in higher payment. We then show that when customers also care much for fairness in service quality, then a mode of

operation that results in both high customer expected utility and expected quality does not exist. The firm should decide which operational policy they want to use based on what they focus on between customer utility and expected quality. We then provide real world examples to show how our results explain the current practices of service firms under certain operational parameters.

This paper is one of the earliest efforts made to model different fairness concerns of decision makers in service operations, and there are many possible extensions that can be made. First of all, the performance measures we look at are by no means exhaustive, and there can be other metrics that can be interesting to look at. There can also be other remedies for worker fairness concerns such as awards for service quality. In addition, the concept of fairness may be different from treating the workers equally in the presence of inherent difference in their ability, and our results may change depending on the definition of fairness.

Chapter 4

Signaling Service Quality through Advertising: A Model of Consumer Memory

4.1 Introduction

Quality of service is often difficult for the customers to judge before experiencing it. This is why services are often called "experience goods", which is a terminology introduced by Nelson (1970) to represent products whose values can be verified only after being purchased. When product quality is hard for the customers to judge before purchase, firms use advertisement to signal their quality and persuade more customers to buy their products, especially when they are high-quality. Moreover, if higher quality firms have higher efficiency in signaling due to say, word of mouth or financial advantage, and advertisement messages are informative in that customers who are exposed to more ads tend to believe the firm to be of higher quality, then it seems natural for higher quality firms to advertise vigorously to inform customers about the truth.

However, in reality, advertisement is more intensively used in some service industries than others, and firms in different circumstances adopt different advertisement intensity. For example, ads for fast food restaurants are easy to find, while ads for gourmet restaurants in Old City Philadelphia are not frequently encountered. In this paper, we try to answer how intensively service firms with certain service quality should advertise when their service qualities are not exactly known to the customers. Since advertising is an act to change the customers' belief about the service value, firms' advertisement strategies depend on customers' initial belief about service.

We look at a model where the service value, either high or low, is known only to the firm and customers are only aware of the distribution of the service value. Given its service value, the firm chooses the frequency of advertising messages it releases to maximize profit. Although the firm sends out messages equally frequently to all the customers, each consumer encounters and forgets the messages stochastically, and forms heterogeneity in the number of advertisement messages they remember. Based on the number of messages a customer remembers, she updates her belief about the firm's service quality and strategically decides whether to purchase the service or not by comparing her updated expected value and the expected cost from crowding at the server that will be incurred upon purchase.

We model the interaction between the firm and the customers using a signaling game and solve for the equilibria to find the firm's advertisement strategy. By analyzing the separating equilibrium, we first show that when customers' initial belief about service quality is low, service firms use separating strategy, and they actively use advertisement to attract more customers. Even when the firm chooses a separating equilibrium, because there is randomness in the way customers encounter and process the advertisement messages, advertisements cannot signal service quality perfectly. Therefore, not only a high-quality firm but also a low-quality firm benefits from advertisement for there will be some customers who may think that the firm is of high type after seeing the ad.

We also find that when customers' initial expected service value is high, then not only a low-quality firm but also a high-quality firm will prefer not to signal its quality through advertisement by choosing a pooling strategy. This is due to the fact that no matter how vigorously a high-quality firm advertises, there will be customers who do not encounter the messages and believe the firm to be of low-quality. In this case, the high-quality firm prefers to make advertisement obsolete by choosing the same advertisement intensity as the low-quality firm so that the customers cannot update their initial belief about the service value. In this case, even when a customer receives few or no advertisement messages, she will not lower her expectation about the firm's quality.

There has been very little research on modeling signaling in service operations contexts, even though signaling models have been employed extensively in economics and marketing areas. This paper purports to build a model of (a) signaling decisions of firms in service settings, and (b) customer decisions based on learning and forgetting in congestion-prone environments. We model how congestion externalities interact with signaling efforts in service settings. In the following section, we review the literature and position our work.

4.2 Literature Review

Queueing Models with Strategic Customers: Although we do not model our server as a queue, our paper is related to papers on queueing models with strategic customer decision in that customers who purchase the service exert negative externalities to other customers. Also, customers in our paper compare the cost of crowding at the server to the expected service value in order to make purchasing decisions. A queueing model with customers who strategically make joining/balking decisions was first analyzed by Naor (1969b). In Naor (1969b), each customer, upon arrival to a service center, observes the queue length and decides whether to join or balk from the queue by comparing the value gained from

the service and the expected cost incurred by waiting. The paper finds that levying a toll on the service can induce the system to operate in a socially optimal manner. Mendelson and Whang (1990b) extends Naor (1969b) by having multiple classes of customers where each class differs in the value gained by service and in waiting cost. They analyze an incentive-compatible pricing scheme that leads to a social optimum when customers follow their own utility-maximizing strategy. The queue length is unobservable to customers that arrive in the system, and customers make decisions by comparing their expected waiting time and their value of service.

Queueing Models with Signaling: Since service firms are usually modeled as a queue in the operations management literature, our work also relates to queueing models with signaling. Hassin (1986) analyzes a queueing system, in which service quality is measured by the customers' waiting time. The author considers a revenue-maximizing server who has the option to suppress information about the queue length, which leaves the customers to choose whether to join or balk based on the known distribution of waiting time. The author shows that it is sometimes, but not always, socially optimal to prevent suppression and that it is never optimal to encourage suppression when the revenue maximizer prefers to reveal the queue length. Guo and Zipkin (2007) analyzed how different degrees of precision in a delay-information announcement can have different effects on customers' decisions and so on the overall system. This research explores a queue with balking under three levels of delay information: no information, partial information (the system occupancy) and full information (the exact waiting time). Customers decide whether to stay or balk based on their expected waiting costs, conditional on the information provided. By comparing the three systems, Guo and Zipkin (2007) identifies some cases in which more accurate delay information improves performance and some other cases in which information can actually hurt the provider or the customers. Allon et al. (2007) answers the question of which delay announcements a service provider should use when both firm and consumers

act strategically: the firm choosing its delay announcement and consumers in interpreting this announcement and making joining/balking decisions based on the expected waiting cost. The authors show that, even though the information provided to customers is non-verifiable and non-credible, it improves the profits of the firm and the expected utility of the customers. Further, the precision of information can vary greatly, from as simple as high or low congestion announcement to the true state of the system. They also show that firms may choose to disguise some of the truth by using intentional vagueness to lure customers. Debo and Veeraraghavan (2009) study how a high-quality service firm selects a service rate differently from a low-quality service firm when the service rates are not observable and the service value is unknown to the customers in the market. They show that customers may not follow a threshold type policy in their queue-joining behavior. They find differentiating equilibria in which the high-quality firm selects a slower service rate than the low-quality service firm, even if the cost of speeding up is the same for both firms. Our paper can be distinguished from the papers mentioned above since the papers that model signaling in a queue either deal with signaling about queue length or signaling quality using queue length, whereas we use advertisement as the instrument for signaling.

Signaling and Advertisement: Since we look at advertisement as means to signal a service firm's quality, our work is also related to economics literature that study signaling and advertisement. Signaling games were first brought to attention by Spence (1973) in which in his model there are high and low-ability employees in the job market who are willing to signal their types by their education level. Spence observes that education does not necessarily improve the ability of the employees; however, it may be used as a signal to indicate their ability in the job market. Much advertising literature follows the idea that advertisement signals the quality of products being advertised. Nelson (1974) differentiates between products on a "search good" versus "experience good" basis. With the former, the relevant qualities of the product are evident on inspection, and, because there is little gain

to misrepresentation, advertisements for them can be directly informative. With the latter, crucial aspects of the product's quality are impossible to verify except through use of the product. Thus, for experience goods, the information provided in the advertisement is unverifiable. He shows that, for experience goods, the mere fact that a particular brand of a good is advertised can be a signal to customers that the good has high quality. Kihlstrom and Riordan (1984) extend models of advertising as a signal that are in the spirit of Nelson's argument. Milgrom and Roberts (1986) offer a modeling based on the repeat sales mechanism where both price and advertising are decision variables that may potentially be used as signals of quality. The authors show that, in equilibrium, both price and advertisement may simultaneously be used as signals, with the chosen levels of both differing between high and low-quality firms. Becker and Murphy (1993) treats advertisements and the goods advertised as complements in stable meta-utility functions. Thus, they argue that advertising may not be necessarily informative. Our model is different from the models adopted by the papers mentioned above because we consider a service operations context where customers not only consider the updated valuation but also the cost incurred by crowding when they make their final decisions. Also, customers in our model cannot distinguish the firm's type exactly even after knowing the firm's strategy since the way they encounter and forget the signals is stochastic in nature.

In this paper: (i) We explore the signaling efforts a firm in an environment where customer decision are influenced by congestion. (ii) We study the impact of stochastic nature of customer learning and forgetting on firm's signaling strategies. Thus, we extend signaling studies to stochastic environments with congestion externalities.

4.3 Model

Firm: We examine a monopolistic service firm selling an experience good to customers. Based on the quality of service it provides, the firm can be of either type, high or low, which is known to the firm but not to the customers. We represent the type using $\theta \in \{h, \ell\}$. When the firm is of "high-type", it offers service value $V = v_h$, and when it is of "low-type", it offers service value $V = v_\ell$ to its customers, where $v_h > v_\ell$. Nature chooses the type, $\theta \in \{h, \ell\}$, of the firm, and the type is high with probability p_0 and low with probability $1-p_0$, $\Pr\{V = v_h\} = 1-\Pr\{V = v_\ell\} = p_0$. This prior probability is common knowledge in the market.

Firm's Quality Signaling: Since the service value is unobservable to the customers, the firm signals its quality by allocating capacity (e.g. labor, workforce) on informative advertising effort. We represent this advertising effort using $a_\theta \in (0, \infty)$, which denotes the rate at which the firm sends its informative advertisement messages when it is of type-θ. When the firm makes an advertising effort, it incurs cost of signaling. We assume that signaling cost is linear in signaling intensity. Specifically, when the firm is of type θ and chooses intensity a_θ, it spends $c_\theta a_\theta$ per unit time on signaling effort. We assume that the firm has higher efficiency in signaling its quality if its service quality is higher, i.e., $c_h < c_\ell$.[1].

Firm Objective: We denote the total number of customers that purchase the service from the firm when it is of type-θ as λ_θ, which is endogenously determined as a result of customers' equilibrium decisions. As the traffic of customers that arrive to the service firm increases, congestion incurs from crowding, which has negative effect on customers' utility. Given its type, θ, the firm maximizes its profit, R_θ, (revenue net of signaling cost) by choosing the right advertisement intensity, a_θ. The revenue accrued from each customer

[1]This is in accordance with extant signalling literature. See Spence (1973). It is also likely that high effort may be required to signal higher quality. This needs to be modeled and is part of ongoing effort

per unit time is normalized to one (wlog), and the profit can be written as

$$R_\theta = \lambda_\theta - c_\theta a_\theta.$$

The Market: There are Λ customers in the potential market, which is weakly greater than the effective arrival rate, λ_θ. We assume that both service quality and level of congestion at the server remain unobservable to the customers. We also assume that consumers are rational utility maximizers. All customers are homogeneous in their aversion to congestion, i.e. they all incur the same time-linear cost of crowding, c. Each customer in the potential market makes a decision whether to purchase the service or not. This decision is based on the value of service she expects and the expected cost of congestion. Reneging is not permitted in the model. Absent any signaling effort by the firm, all consumers would follow the same purchasing strategy since they are *ex ante* homogeneous. However, the signaling effort of the firm creates heterogeneity among customers in terms of the information they possess.

Consumers' Information Set: Learning and Forgetting: We adopt the "forgetting model" introduced in Nelson (1974) to describe the information set of the customers in this paper. Recall that customers are exposed to the signaling efforts of the firm. Specifically, customers stochastically confront advertising messages from the firm that emanate at rate a, and the number of messages a customer confronts over time is a Poisson process with rate a. In addition, customers stochastically 'forget' the messages they have received before. For example, a customer who has received a advertising messages and has forgotten f of them 'remembers' $a - f$ messages. Nelson (1974) assumes that the rate at which customers forget messages is proportional to the number of messages they remember at present. Given that a customer remembers i advertising messages, forgetting follows a Poisson process with rate $i \cdot f$ where f is the forgetting rate per message. Then each customer's learning

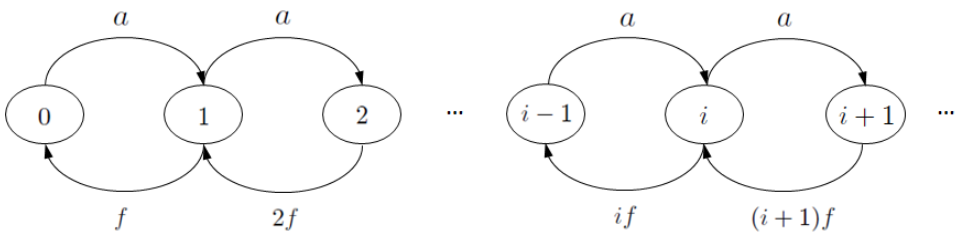

Figure 4.1: *Customers' Learning and Forgetting Process*

and forgetting process can be illustrated by the Markov chain, shown in Figure 4.1, whose state space, $\{0, 1, 2, \cdots\}$, representing the number of advertising messages the customer remembers. Note that this process at the individual customer level could be thought of as an $M/M/\infty$ queue with an arrival rate of a, and the individual servers serving at a rate f. Then, the stationary distribution of the number of messages that remains in any consumer's information set follows a Poisson distribution with rate a/f:

$$\pi_i = \frac{(a/f)^i e^{-a/f}}{i!}, \quad \forall i \tag{4.1}$$

where π_i represents the probability of a customer remembering i advertisement messages in the steady state. Essentially, the signaling process introduces heterogeneity in the consumer population that leads to different choice behavior for customer types based on the differences in their information sets.

Note that the Markov Chain characterized by the forgetting process of the advertising messages at the individual customer level is aperiodic since all the states can be visited in any number of transitions. It is also positive recurrent since $\sum_{n=1}^{\infty} \frac{1}{n!} \left(\frac{a}{f}\right)^n = e^{a/f} < \infty$ when a and f are finite.(See Ross (1996), pg.254.) Then from the ergodicity of the chain, the time averages of samples from this population have the same distribution as the stationary distribution given by (4.1).

Customers' Belief Updating Process: Recall that customers are unaware of the service value of the firm. However, each customer updates her prior belief about the quality of the

firm's service using the number of advertisement messages she remembers in the steady state to arrive at a decision whether to purchase the service. From now on, we call customers who remember i messages type-i customers. Note that when the firm decides its strategy, (a_h, a_ℓ), customers observe the strategy. However, since customers encounter the advertisement messages stochastically, we assume that they do not know the frequency of advertisement messages they are receiving. Type-i customers then use Bayes rule to compute the probability that the frequency of advertisement messages they receive is a_h to know the probability of the firm being a high-type as follows:

$$\Pr(v_h|i) = 1 - \Pr(v_\ell|i) = \frac{\Pr(i|v_h)p_0}{\Pr(i|v_h)p_0 + \Pr(i|v_\ell)(1-p_0)}, \tag{4.2}$$

where $\Pr(i|v_\theta)$ is the conditional probability that a customer remembers i advertisement messages given that the firm is of type-θ and that it is advertising with intensity a_θ. $\Pr(v_\theta|i)$ is the updated probability of the firm being θ-type given that the customer remembers i messages.

Then with this updated belief, customers can compute the expected service value given the number of advertisement messages they remember. Let us denote v_i as the value of service expected by a customer who remembers i advertisement messages. Then,

$$v_i = v_h \Pr(v_h|i) + v_\ell \Pr(v_\ell|i).$$

Customers' Purchasing Decisions: Each type-i customer makes her joining-balking decision by comparing the expected service value v_i and the expected cost of crowding in the system. We represent the purchasing decision of a type-i customer using $q_i \in [0, 1]$, where q_i is the probability of a type-i customer purchasing the service. If the updated service value for type-i customers dominates the expected cost of crowding, type-i customers purchase the service (i.e. $q_i = 1$). On the other hand, type-i customers balk when the congestion cost

is higher than the service value (i.e. $q_i = 0$). When the value gained and the cost of crowding are equal, customers are indifferent between purchasing the service and not, and joins with some probability $q_i \in (0, 1)$. We restrict ourselves to purchasing strategies that are identical within customers of the same type. In other words, all customers of type-i adopt the same purchasing strategy q_i. Now, in order to characterize the customers' strategy, q_i, let us consider what λ_θ is. Given that nature has chosen θ for the firm's type, the proportion of type-i customers in the entire market is

$$\Pr(i|v_\theta) = \frac{(a_\theta/f)^i e^{-a_\theta/f}}{i!}, \tag{4.3}$$

for all $\theta \in \{h, \ell\}$. Since nature chooses v_h with probability p_0 (and v_ℓ with probability $(1 - p_0)$), the unconditional distribution of the customers in the market is given by

$$\Pr(i) = \Pr(i|v_h)p_0 + \Pr(i|v_\ell)(1 - p_0).$$

Note that since customers do not know the exact type of the firm and also the advertisement intensity, a_θ, customers can at the very best guess that there are $\Pr(i)$ proportion of customers in the market that remember i advertisement messages. Since customers do not know the type of the firm, they cannot know the exact effective arrival rate either. We use $\bar{\lambda}$ to denote the effective arrival rate perceived by the customers as opposed to the real effective arrival rate, λ_θ. Given that every type-i customer adopts q_i, the effective arrival rate $\bar{\lambda}$ that customers believe is given as $\Lambda \sum_{j=0}^{\infty} q_j \Pr(j)$ and the net value a customer of type-i believes she will earn from joining the service is given as

$$v_i - c\Lambda \sum_{j=0}^{\infty} q_j \Pr(j).$$

Then the joining decision of customers can be summarized as

$$\begin{cases} q_i = 1 & \text{if} \quad v_i - c\Lambda \sum_{j=0}^{\infty} q_j \Pr(j) > 0, \\ q_i = 0 & \text{if} \quad v_i - c\Lambda \sum_{j=0}^{\infty} q_j \Pr(j) < 0, \\ q_i \in (0,1) & \text{if} \quad v_i - c\Lambda \sum_{j=0}^{\infty} q_j \Pr(j) = 0, \end{cases} \tag{4.4}$$

where q_i for $\forall i$ is the best response to $q_{-i} = \{q_0, q_1, \cdots, q_{i-1}, q_{i+1}, \cdots\}$.

The Equilibrium: The equilibrium we seek is a Markov Perfect Bayesian Nash Equilibrium (Fudenberg and Tirole (1991)) where customers update their strategies in Bayesian fashion and makes purchasing decisions to maximize their payoffs on all Markov paths reached with positive probability.

Definition 21 *The equilibrium of the game between the firm and the customers is of the form* $[(a_h^*, a_\ell^*), (q_i^*, i \in \{0, 1, \cdots\})]$*, so that*

(i) (Profit Maximization) given q_i *for all* i*, the firm chooses* (a_h^*, a_ℓ^*) *that maximizes its profit.*

$$a_\theta^* = \arg\max_{a_\theta} \left\{ \Lambda \sum_{j=0}^{\infty} q_j \Pr(j|v_\theta) - c_\theta a_\theta \right\}, \ \forall \theta.$$

(ii) (Customer Rationality) given a_θ *for all* θ *and* q_{-i}*,* q_i *maximizes type-*i *customers' payoffs.*

$$q_i^* = \arg\max_{q_i} \left\{ v_i - c\Lambda \sum_{j=0}^{\infty} q_j \Pr(j) \right\}, \ \forall q_{-i}, \ \forall i.$$

Recall that the profit function of the firm is given as $R_\theta(a_\theta) = \lambda_\theta - c_\theta a_\theta$ when the type is θ. The actual effective arrival rate λ_θ is given as

$$\lambda_\theta = \sum_{j=0}^{\infty} q_j \Pr(j|v_\theta).$$

Note that since the firm knows its type, the actual effective purchase rate uses the conditional probability, $\Pr(i|v_\theta)$, to denote the proportion of customers that remember i number of messages instead of the unconditional probability, $\Pr(i)$. Since the revenue accrued from each customer per unit time is normalized to one in our model, the type-θ firm's profit per unit time (R_θ) is calculated by the number of customers purchasing the service per unit time subtracted by the signalling cost per unit time:

$$R_\theta = \Lambda \sum_{j=0}^{\infty} q_j \Pr(j|v_\theta) - c_\theta a_\theta \tag{4.5}$$

where $\Pr(i|v_\theta)$ is given as in (4.3).

4.4 Results

In this section, we analyze the equilibrium of the game to describe the firms advertising decisions. We intend to show when signaling is valuable to the firm by identifying the condition under which the firm chooses a separating equilibrium over a pooling equilibrium. Also, we show in this section that there is a situation where signaling harms the firms profit so that the firm deliberately chooses to remain ambiguous about its service quality even for the high quality firm. In order to do so, we show the following structural results. First, we show that if the high-quality firm advertises more than the low-quality firm, a customer who remembers more advertisement messages will have stronger beliefs about the firm having high quality. The opposite will also be shown if the high-quality firm advertises less. (i.e. the customer will tend to think a service has high quality if she remembers fewer messages). Based on this result, we show that customers follow a threshold type policy so that they purchase the service only if they remember more than a certain number of advertisement messages when the firm chooses to separate with advertising more intensively

when it is of high-quality (and the opposite is true when the firm advertises more intensively when it is of low-quality.) Then we show that a low quality firm does not advertise more than a high quality firm in equilibrium. Then we identify a condition under which separating equilibrium is played by the firm as opposed to a pooling equilibrium.

In the following Lemma we show how customer belief changes based on the number of advertisement messages she remembers when the firm's strategy is given.

Lemma 22

1. *If* $a_h > a_\ell$*, then* $\Pr(V = v_h|i)$ *is increasing in* i*.*

2. *If* $a_h < a_\ell$*, then* $\Pr(V = v_h|i)$ *is decreasing in* i*.*

From Lemma 22, we can see that if a firm chooses higher advertisement intensity when it is of high-type than when not, then customers who remember higher number of advertisement messages has a stronger belief about the firm being a high quality firm.

Next, we show that if a high-quality firm advertises more, then customers who remember more advertising messages will expect higher value from the service than those who remember fewer messages. This fact follows directly from Lemma 22 which suggests that, when a high-quality firm releases advertising messages more frequently, customers with more messages in mind believe more strongly that the service is of high quality. Since their perceived probability of high quality is higher, the expected value follows to be higher. Similarly, the opposite is again shown: if a high-quality firm advertises less, then the opposite is true (i.e. fewer remembered messages implies higher expected service value).

Lemma 23

1. *If* $a_h > a_\ell$*, then* v_i *is increasing in* i*.*

2. *If* $a_h < a_\ell$*, then* v_i *is decreasing in* i*.*

Now, from the customers' belief and expected value of the service, we can characterize the strategy of each customer type.

Lemma 24 *Suppose there exists an equilibrium such that $q_j \geq 0$ for all j. Then*

1. *If $a_h > a_\ell$, then $q_i > 0$ implies $q_{i+1} > 0$.*

2. *If $a_h < a_\ell$, then $q_i > 0$ implies $q_{i-1} > 0$.*

First, let us examine Lemma 24(1), where the high-quality firm advertises more. If customers who remember i signaling messages choose to join the service, then customers who remember $(i+1)$ messages must also. Note that this follows from Lemma 23, in which we have shown that $(i+1)$-type customers are expecting higher value from the service than i-type customers do. Since both types of customers are facing the same expected waiting cost, it is natural that $(i+1)$-type customers join if i-type customers with even lower expected value join the queue. The reverse argument is applicable to the case where high-quality firm advertises less than low-quality firm. In this case, if i-type customers choose to join, then $(i-1)$-type customers should also join. The fewer messages the customers remember, the more likely they are to join the service.

Lemma 25 *Suppose there exists an equilibrium such that $q_j \geq 0$ for all j. Then*

1. *If $a_h > a_\ell$, then $q_i = 0$ implies $q_{i-1} = 0$.*

2. *If $a_h < a_\ell$, then $q_i = 0$ implies $q_{i+1} = 0$.*

This result is similar to Lemma 24 except for that the latter describes customers' balking behavior. It states that if high-quality service firm advertises more than the low-quality firm and if customers who remember i messages choose not to purchase the service, then those who remember less than i messages should also. It is again followed by the fact that all types of customers expect the same cost of crowding, but the ones with less number

of messages in mind have lower expected service value than the ones that have more. If customers expecting higher value do not purchase the service, then those expecting less will not either. The opposite is true for the case in which the low-quality firm advertises more than the high-quality firm. If customers with i messages do not purchase the service, those with $(i+1)$ will not as well.

Based on what we have analyzed about customers' reaction to the advertising levels of the firms, the customers' best response strategy can be summarized as the following.

Lemma 26 *Let a_θ and q_i be the decision of the firm of type-θ and the customers of type-i, respectively. Then*

1. If $a_h > a_\ell$, $q_0 = 0$ and there exists $i > 0$ with $q_i > 0$, then there exists $K > 0$ such that

$$\begin{cases} v_i - c\sum_{j=0}^{\infty}\Lambda[q_j \Pr(j|v_h)p_0 + q_j \Pr(j|v_\ell)(1-p_0)] < 0, & q_i = 0, & \forall i = 0, \cdots K-1 \\ v_i - c\sum_{j=0}^{\infty}\Lambda[q_j \Pr(j|v_h)p_0 + q_j \Pr(j|v_\ell)(1-p_0)] \geq 0, & q_i \in [0,1], & \forall i = K \\ v_i - c\sum_{j=0}^{\infty}\Lambda[q_j \Pr(j|v_h)p_0 + q_j \Pr(j|v_\ell)(1-p_0)] > 0, & q_i = 1, & \forall i = K+1, \cdots \end{cases}$$

2. If $a_\ell > a_h$, $q_0 = 1$ and there exists $i > 0$ with $q_i = 0$, then there exists $K > 0$ such that

$$\begin{cases} v_i - c\sum_{j=0}^{\infty}\Lambda[q_j \Pr(j|v_h)p_0 + q_j \Pr(j|v_\ell)(1-p_0)] > 0, & q_i = 1, & \forall i = 0, \cdots K-1 \\ v_i - c\sum_{j=0}^{\infty}\Lambda[q_j \Pr(j|v_h)p_0 + q_j \Pr(j|v_\ell)(1-p_0)] \geq 0, & q_i \in [0,1], & \forall i = K \\ v_i - c\sum_{j=0}^{\infty}\Lambda[q_j \Pr(j|v_h)p_0 + q_j \Pr(j|v_\ell)(1-p_0)] < 0, & q_i = 0, & \forall i = K+1, \cdots \end{cases}$$

Lemma 26 characterizes the equilibrium purchasing behavior in the market, for any given advertising choice made by the firms. When the firm's strategy is to advertise more if it has high quality and less if low quality, customers follow a threshold-type policy in which customers who remember more than a critical-K number of messages all purchase the service and those who remember less than K all balk. Those with exactly K messages may purchase the service with some weakly positive probability. For the case in which a high-quality firm advertises less and low-quality firm advertises more, customers again respond with a threshold policy but in an opposite direction: Customers who remember less than a critical-K number of messages all join the service, those who remember more

do not, and those with exactly K joins with some weakly positive probability.

What remains to be studied are the firms' strategies. In the next steps, we show that there cannot be an equilibrium in which a low-quality firm advertises more than a high-quality firm; the only possible case in equilibrium for the high-type firm to signal more than or equal to the low-type firm's advertisement intensity.

Lemma 27 $a_h^* \geq a_\ell^*$.

If we assume that the low-type firm advertises more than the high-type firm, then we have shown in Lemma 26 that customers' best response to this strategy is to purchase the service if they remember less than some threshold number of advertising messages. Then, it is always better off for the firm to advertise less because more customers will purchase the service if they do not remember many advertisement messages. Moreover, since advertising is costly, their profit decreases even more when they advertise. Thus, the firms' best response to such strategy of customers' is not to advertise, which contradicts to the fact that the low-quality firm is advertising more than the high-quality firm.

Now we analyze the equilibrium and show the conditions under which the firm decides to follow a separating strategy over pooling strategy. Before we do so, we note that when the firm pools, advertisement loses its signaling purpose, and the number of advertisement messages customers remember no longer has any meaning. Customers, thus, no longer can update their beliefs about service quality, and use solely the initial *ex ante* service value to make their purchasing decisions. The following Lemma summarizes the set of equilibria and the condition under which each equilibrium is sustainable.

Proposition 28 *For a given c_h, c_ℓ, c, p_0, v_h, and v_ℓ, the following holds.*
1. A separating equilibrium is of the form $[(a_h^*, a_\ell^*) \in F_s, q^* \in C_s]$, where

$$F_s := \{(a_h, a_\ell) : a_h > a_\ell \geq 0\},$$
$$C_s := \{(q_0, q_1, \cdots) : q_i = 0 \textit{ for } i < K,\ q_i = 1 \textit{ for } \forall i > K,\ q_K \in [0, 1]\},$$

and $K \in \{0, 1, 2, \cdots\}$.

2. A pooling equilibrium is of the form $[(\tilde{a}_h^*, \tilde{a}_\ell^*) \in F_p, q^* \in C_p]$*, where*

$$F_p := \{(a_h, a_\ell) : a_h = a_\ell = \tilde{a} \geq 0\},$$
$$C_p := \{(q_0, q_1, \cdots) : q_i = \min\left\{1, \tfrac{p_0 v_h + (1-p_0)v_\ell}{c\Lambda}\right\}, \textit{ for } \forall i = 0, 1, 2, \cdots\},$$

3. There exist thresholds $\tilde{v}_1 > 0$ *and* $\tilde{v}_2 > 0$ *such that only separating strategy exists for* $p_0 v_h + (1 - p_0)v_\ell < \tilde{v}_1$*, and only pooling equilibrium exists for* $p_0 v_h + (1 - p_0)v_\ell > \tilde{v}_2$.

The above equilibrium states that when separating equilibrium exists, it is of the form where the firm advertises more intensively when it is high-quality than when it is low-quality. On the other hand, under a pooling equilibrium, the advertisement intensity can be any number $\tilde{a} > 0$. However, since the advertisement messages do not have any signaling function, advertisement cost is a true waste without any added value, and the most efficient pooling equilibrium is not to advertise at all. However, in the real world, firms may fail to coordinate well to reach the most efficient equilibrium and end up advertising with positive intensity although the ads do not add any value.

When customers' initial *ex ante* belief on service value is low, not many customers will purchase the service if there is no way to update their belief about the service quality. Thus, the firm needs advertisement to induce more purchase, and it will choose a separating equilibrium (because if the firm pools, then advertising loses signaling function and customers will make decisions solely based on their low initial expected value). If the firm is of high-quality, it chooses higher advertisement intensity, a_h^*, than the intensity, a_ℓ^* it would have chosen in case it were of low-quality, so that more customers can remember larger number of advertisement messages and purchase the service. What is interesting is that even when the firm is of low-quality, it benefits by choosing a separating strategy. This is because the customers cannot perfectly distinguish the firm's type by directly observing the adver-

tisement intensity, but can only imperfectly distinguish it by the number of advertisement messages they remember, which is an outcome of their stochastic advertisement encountering and forgetting process. When the prior belief about the quality is low, a low-quality firm also chooses to separate so that advertisement does not lose its function. Then due to the stochastic nature of advertisement encountering process, there will be some customers that remember large number of the low-quality firm's advertisement messages and believe that the firm is of higher quality than what it actually is. Thus, the low-quality firm will separate well enough to preserve the function of advertisement, but at the same time will try to advertise as much as possible so that large enough number of customers can remember many advertisement messages so as to purchase the service.

On the other hand, when customers' initial *ex ante* belief on service value is high, large number of customers will purchase the service based on the initial belief even without advertisement. In this case, the firm is better off not creating heterogeneous belief about service quality among customers. Again, because of the stochastic nature of advertisement message encountering and forgetting process, even when the firm is of high-type and chooses a high advertisement intensity, there will be some customers who end up remembering very small number of advertisement messages and believe the service to be of low-quality. Had the firm chosen pooling equilibrium, larger number of customers could have purchased the service based on their high prior belief. However, when the firm separates to signal its quality, it loses sales from customers who end up remembering only a small number of messages. This effect is even more salient when the firm is of low-type because the firm will choose lower advertisement intensity, and more customers tend to remember few advertisement messages. Thus, when the initial expected service value is high, the firm has an incentive to deliberately be ambiguous about its quality, and this is interestingly the case even for a high-quality firm.

4.5 Conclusion

In this paper, we have looked at a model in which a firm's service can have either high or low quality, and the quality of service is kept as the firm's individual information so that the customers are left uncertain about the firm's quality. The firm sends signals to customers by advertising to update customers' belief about its service quality so that it can convince the customers to purchase the service. We modeled the interaction between the firm and the customers as a signaling game and found Bayesian Nash equilibria of the game.

By analyzing the equilibria, we have shown that when customers' initial expectation about service quality is low, then the firm benefits by choosing separating equilibrium no matter what the firm's type is. When the firm separates, it advertises more intensively when it is of high-quality than when it is of low-quality. Customers purchase the service only when they remember more than a certain threshold number of advertisement messages. In this case, high-quality firm chooses to advertise vigorously so that more customers can remember more advertisement messages than the threshold value. A low-quality firm, on the other hand, benefits from the fact that advertisement messages are encountered and forgotten randomly by customers, i.e., that they are imperfect signals of the true quality. Even though the low-quality firm advertises with lower intensity than it would have done if it were a high-quality firm, there will be some customers who happen to be exposed to the advertisement messages frequently and end up believing that the firm is of high-quality and purchase the service.

On the other hand, when customers' initial expected service value is high, not only a low-quality firm but also a high-quality firm will prefer to choose pooling equilibrium to make advertisement meaningless. When the firm chooses a pooling equilibrium, then advertisement messages no longer incorporate any information about service quality, and customers cannot update their belief about service quality. Note that even when the firm

separates, advertisement messages are imperfect means to signal quality. Thus, even when the high-quality firm advertises with very high intensity, there will be customers who happen not to encounter the messages and end up misjudging the quality of service. Thus, when initial expected service quality is high, not only a low-quality firm but also a high-quality firm prefers to pool to make advertisement obsolete and have the customers make their purchasing decisions based on their high initial expected valuation.

Future step that will be interesting to work on is to relax the assumption that high-quality firm has higher efficiency in signaling. As in Kihlstrom and Riordan (1984) it will also be interesting to look at the equilibria in which the high-quality firms require investment in specialized assets that increases fixed costs, but decreases the marginal cost of advertising. Introducing price of the service as a decision parameter can be another way to extend the model as in many advertising papers from economics do. Since not much research has been conducted in models that signal quality by advertising in a service operations context, we suppose many variants can evolve from our model.

Chapter 5

Conclusion

This dissertation focuses on modeling the decision makers' behavior in service operations management problems. Three projects that study different operational problems service firms face will comprise the thesis.

The first project looks at a service firm's pricing decisions for reservations when customers are given an option to make reservations to avoid waiting. We consider three policies, fully-prepaid reservations, partially-prepaid reservations, and no reservation, which are regimes found in the advance selling literature as advance selling, option selling, and spot selling. We show that when the customers have homogeneous *ex ante* belief about their service valuations, then the firm can maximize profit by selling fully-prepaid reservations. The firm is better off by charging lower price to reservation customers than the price of service charged to walk-ins. However, when the potential market becomes large, then it is better not to offer reservations, and only serve walk-in customers. We also find that when customers have heterogenous expectations about service valuation or when the firm overbooks, then selling partially-prepaid reservations becomes an optimal policy.

The second project considers how a service firm can alleviate workers' demotivation due to fairness concerns when the firm hires multiple workers with different service quali-

ties. We propose two options, eliminating and compensating for inequalities in workload, and compare the two remedies using expected customer utility and expected quality. We show that eliminating inequality in workload works, but compensating workers per customers they serve works even better in terms of both expected customer utility and average quality. However, when customers also care for equality in service quality, a remedy that leads to both high customer expected utility and average quality no longer exists. In this case, distributing workload fairly between workers, which provides reasonable quality without sacrificing customer utility too much, emerges as an option.

The third project answers how service firms should choose their advertisement intensities depending on their service qualities when the qualities are not exactly known to the customers. We look at a signaling game between the firm and the potential customers and show that the firm is better off by playing a separating equilibrium regardless of the service quality when customers' initial belief about quality is low. Because advertisement messages are encountered by customers stochastically and thus, cannot signal the quality perfectly, low quality firm also benefits from signalling due to some customers that happen to believe the firm to have high quality by chance. On the other hand, when customers initially hold a high belief about quality, then the firm is better not to introduce noise by signalling, and chooses to play a pooling equilibrium.

Bibliography

Alexandrov, A., M.A. Lariviere. 2012. Are reservations recommended? *Manufacturing and Service Operations Management* **14**(2) 218–230.

Allon, Gad, Achal Bassamboo, Itay Gurvich. 2007. We will be right with you: Managing customers with vague promises. Kellogg Working Paper.

Avi-Itzhak, Benjamin, Hanoch Levy. 2004. On measuring fairness in queues. *Advances in Applied Probability* 919–936.

Becker, Gary S., Kevin M. Murphy. 1993. A simple theory of advertising as a good or bad. *The Quarterly Journal of Economics* **108**(4) 941–964.

Bolton, Gary E, Axel Ockenfels. 2000. Erc: A theory of equity, reciprocity, and competition. *American economic review* 166–193.

Boyacı, T., Ö. Özer. 2010. Information acquisition for capacity planning via pricing and advance selling: When to stop and act? *Operations Research* **58**(5) 1328–1349.

Camerer, Colin, Richard H Thaler. 1995. Anomalies: Ultimatums, dictators and manners. *The Journal of Economic Perspectives* **9**(2) 209–219.

Charness, Gary, Peter Kuhn. 2007. Does pay inequality affect worker effort? experimental evidence. *Journal of Labor Economics* **25**(4) 693–723.

Charness, Gary, Matthew Rabin. 2002. Understanding social preferences with simple tests. *The Quarterly Journal of Economics* **117**(3) 817–869.

Chu, L., H. Zhang. 2011. Optimal preorder strategy with endogenous information control. *Management Science* **57**(6) 1055–1077.

Cil, Eren B, Martin A Lariviere. 2013. Saving seats for strategic customers. *Operations Research* **61**(6) 1321–1332.

Clark, Andrew E, Andrew J Oswald. 1996. Satisfaction and comparison income. *Journal of public economics* **61**(3) 359–381.

Collins, G. 2010. Table for 2? get ready to wait in line. URL `http://www.nytimes.com/2010/06/09/dining/09reservations.html?pagewanted=all&_r=0`.

Debo, Laurens G., Senthil K. Veeraraghavan. 2009. Firm service rate selection when service rates are not observable and service value is unknown to the market. Chicago Booth Research Paper No. 09-32.

DeGraba, P. 1995. Buying frenzies and seller-induced excess demand. *RAND Journal of Economics* **26**(2) 331–342.

Fay, S., J. Xie. 2008. Probabilistic goods: A creative way of selling products and services. *Marketing Science* **27**(4) 674–690.

Fehr, Ernst, Klaus M Schmidt. 1999. A theory of fairness, competition, and cooperation. *The quarterly journal of economics* **114**(3) 817–868.

Fisher, M., A. Raman. 1996. Reducing the cost of demand uncertainty through accurate response to early sales. *Operations Research* **44**(1) 87–99.

Frank, Robert H. 1984. Are workers paid their marginal products? *The American economic review* **74**(4) 549–571.

Fudenberg, Drew, Jean Tirole. 1991. Perfect bayesian equilibrium and sequential equilibrium. *Journal of Economic Theory* **53** 236–260.

Gallego, G., Ö Şahin. 2006. Inter-temporal valuations, product design and revenue managemet Working Paper, Columbia University, New York.

Georgiadis, George, Christopher S Tang. 2012. Optimal reservation policies and market segmentation. *Available at SSRN 1860398* .

Glass, Gene V. 1982. School class size: Research and policy. .

Goel, Anand M, Anjan V Thakor. 2006. Optimal contracts when agents envy each other. *Draft December* **18** 2006.

Guo, P., P. Zipkin. 2007. Analysis and comparison of queues with different levels of delay information. *Management Science* **53**(6) 962–970.

Güth, Werner, Reinhard Tietz. 1990. Ultimatum bargaining behavior: A survey and comparison of experimental results. *Journal of Economic Psychology* **11**(3) 417–449.

Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue*. Kluwer Academic Publishers, Norwell.

Hassin, Refael. 1986. Consumer information in markets with random product quality: The case of queues and balking. *Econometrica* **54**(5) 1185–1195.

Jerath, K., S. Netessine, S. Veeraraghavan. 2010. Revenue management with strategic customers: Last-minute selling and opaque selling. *Marketing Science* **56**(3) 430–448.

Kahneman, Daniel, Jack L Knetsch, Richard H Thaler. 1986. Fairness and the assumptions of economics. *Journal of business* S285–S300.

Kihlstrom, Richard E., Michael H. Riordan. 1984. Advertising as a signal. *The Journal of Political Economy* **92**(3) 427–450.

Kim, Yongsung. 2013. Why do good workers leave the company? burn out syndrom. URL `http://biz.chosun.com/site/data/html_dir/2013/10/27/2013102701959.html`.

Larson, Richard C. 1987. Or forumperspectives on queues: Social justice and the psychology of queueing. *Operations Research* **35**(6) 895–905.

Lazear, Edward P. 1989. Pay equality and industrial politics. *Journal of political economy* 561–580.

Li, C., F. Zhang. 2013. Advance demand information, price discrimination, and preorder strategies. *Manufacturing and Service Operations Management* **15**(1) 57–71.

MacInnis, Kate. 2012. Class-size envy. URL `http://mathdancing.wordpress.com/2012/09/14/class-size-envy/`.

Mendelson, H., S. Whang. 1990a. Optimal incentive-compatible priority pricing for the m/m/1 queue. *Operations Research* **38**(5) 870–883.

Mendelson, Haim, Seungjin Whang. 1990b. Optimal incentive-compatible priority pricing for the m/m/1 queue. *Operations Research* **38**(5) 870–883.

Meyer, Margaret A, Dilip Mookherjee. 1987. Incentives, compensation, and social welfare. *The Review of Economic Studies* **54**(2) 209–226.

Milgrom, Paul, John Roberts. 1986. Price and advertising signals of product quality. *The Journal of Political Economy* **94**(4) 796–821.

Mirrlees, James A. 1971. An exploration in the theory of optimum income taxation. *The review of economic studies* **38**(2) 175–208.

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* **29**(3) 315–335.

Naor, P. 1969a. The regulation of queue size by levying tolls. *Econometrica* **37**(1) 15–24.

Naor, P. 1969b. The regulation of queue size by levying tolls. *Econometrica* **37**(1) 15–24.

Nasiry, J., I. Popescu. 2012. Advance selling when consumers regret. *Management Science* **58**(6) 1160–1177.

Nelson, Phillip. 1970. Information and consumer behavior. *The Journal of Political Economy* **78**(2) 311–329.

Nelson, Phillip. 1974. Advertising as information. *The Journal of Political Economy* **82**(4) 729–754.

Png, I. P. L. 1989. Reservations: Customer insurance in the marketing of capacity. *Marketing Science* **8**(3) 248–264.

Prasad, A., K. Stecke, X. Zhao. 2011. Advance selling by a newsvendor retailer. *Production and Operations Management* **20**(1) 129–142.

Rabin, Matthew. 1993. Incorporating fairness into game theory and economics. *The American economic review* 1281–1302.

Rafaeli, Anat, Efrat Kedmi, Dana Vashdi, Greg Barron. 2003. Queues and fairness: A multiple study experimental investigation.

Raz, David, Hanoch Levy, Benjamin Avi-Itzhak. 2004. A resource-allocation queueing fairness measure. *ACM SIGMETRICS Performance Evaluation Review*, vol. 32. ACM, 130–141.

Ross, Sheldon M. 1996. *Stochastic Processes*. John Wiley and Sons, Inc.

Schmitt, Bernd H, Laurette Dube, France Leclerc. 1992. Intrusions into waiting lines: does the queue constitute a social system? *Journal of Personality and Social Psychology* **63**(5) 806.

Seymour, Liz. 1999. Reduced student load creates class envy among some teachers. URL `http://articles.latimes.com/1999/jan/11/news/mn-62518`. Los Angeles Times.

Shreedhar, Madhavapeddi, George Varghese. 1995. Efficient fair queueing using deficit round robin. *ACM SIGCOMM Computer Communication Review* **25**(4) 231–242.

Shugan, S. M., J. Xie. 2000. Advance pricing of serivces and other implications of separating purchase and consumption. *Journal of Service Research* **46**(3) 37–54.

Shugan, S.M., J. Xie. 2005. Advance-selling as a competitive marketing tool. *International Journal of Research in Marketing* **22**(3) 351–373.

Singh, Ramadhar. 1995. " fair" allocations of pay and workload: Tests of a subtractive model with nonlinear judgment function. *Organizational Behavior and Human Decision Processes* **62**(1) 70–78.

Spence, Michael. 1973. Job market signaling. *The Quarterly Journal of Economics* **87**(3) 355–374.

Stokey, N.L. 1981. Rational expectations and durable goods pricing. *The Bell Journal of Economics* **12**(1) 112–128.

Tang, C., K. Rajaram, A. Alptekinoğlu. 2004. The benefit of advance booking discount programs: Model and analysis. *Management Science* **50**(4) 465–478.

Wierman, Adam, Mor Harchol-Balter. 2003. Classifying scheduling policies with respect to unfairness in an m/gi/1. *ACM SIGMETRICS Performance Evaluation Review*, vol. 31. ACM, 238–249.

Yu, M., R. Kapuscinski, H. Ahn. 2008. Advance selling: The effect of capacity and customer behavior Working Paper, University of Michigan, Ann Arbor.