Pricing and Operational Performance in Discretionary Services

Chunyang Tong

School of International Business Administration, Shanghai University of Finance and Economics,

tong.chunyang@mail.shufe.edu.cn

Sampath Rajagopalan

Marshall School of Business, University of Southern California, raj@marshall.usc.edu

Abstract

In many services, for example website or landscape design, the value or quality derived by a customer depends upon the service time and this valuation differs across customers. Customers procure the service based on the expected value to be delivered, prices charged and the timeliness of service. We investigate the performance of the optimal pricing scheme as well as two commonly used pricing schemes (fixed fee and time-based pricing) for such services on important dimensions such as revenue, demand served, and utilization. We propose a novel model that captures the above features and wherein both service rate and demand are endogenous and functions of the pricing scheme. In particular, service time is an outcome of the pricing scheme adopted and the heterogeneous valuations of customers, unlike in the queueing-based pricing literature. We find that the service system may benefit from a greater variance in consumer valuations, and the performance of pricing schemes is impacted by the shape of the distribution of customers' valuation of service time and the responsiveness desired by customers. Both the fixed fee and time-based schemes do well relative to the optimal pricing scheme in terms of revenue in many plausible scenarios, but there are substantial differences between the pricing schemes in some important operational metrics. For instance, the fixed fee scheme serves more customers and has higher utilization than the time-based scheme. We also explore variants of the fixed and time-based schemes that have better revenue performance and show that the two-part tariff which is a combination of fixed and time-based pricing can do as well as the optimal scheme in terms of revenue.

Keywords: Discretionary Service, Pricing Scheme, Service Performance, Queuing.

History of the paper: Received: January 2012; accepted: May 2013 by Michael Pinedo after three revisions

1 Introduction

In many service systems, the time that a service provider devotes to a customer is a key determinant of service value. Usually, the value (or quality) obtained from this type of service increases with service time and completion of the service process may be subjective. This is in contrast to traditional manufacturing systems and routine service work where value may depend solely on whether or not the work has been completed, as indicated in Hopp et al. (2007) who provide several examples of such service tasks referred to as "discretionary" tasks. Examples of such services, which we refer to henceforth as "discretionary services" include website design, landscape design, and several white-collar, information-based services (Karmarkar and Apte (2007), Wang et al. (2010)). Such services may be termed as "discretionary" because the service time is not exogenous but is at the discretion of the service provider and/or customer.

Management of such discretionary services poses interesting challenges because while customers value additional service time, longer service times can degrade service responsiveness and result in fewer customers being served and lower revenues. An important lever available to the firm in managing such services is the pricing scheme adopted and prices charged. So, a revenue-maximizing firm has to balance three important factors: the service value provided, the pricing scheme used and corresponding fees charged, and responsiveness. The traditional queuing theory-based pricing literature provides methodologies to study the trade-off between the latter two dimensions: the customers' payment and responsiveness (see Hassin and Haviv (2003) for a comprehensive review). However, this literature assumes that service time and its variation are exogenous and do not impact the value derived from the service. We explicitly consider the nature of the *relationship between service time and value* in making pricing decisions to address this important gap in the literature. So, unlike in traditional queue-based pricing literature, service times are not exogenous but an *outcome* of heterogenous consumer valuations of service time and the pricing schemes adopted. The focus here is on identifying the set of prices and service times that optimize the performance of such discretionary service systems and evaluating two commonly used pricing schemes.

We consider a service system with a monopolistic service provider and a stream of rational customers with varying service expectations who arrive stochastically. Customers share their heterogeneous needs and expectations with the firm and the service outcome (i.e. service time and value) and payment made depend on the pricing scheme. Two key features of discretionary services are captured in our model of the relationship between service time and value: 1) Service value increases with service time, but the incremental value gained from additional service time diminishes. 2) The value derived may vary across customers due to differences in service expectations. We first study the optimal or first-best pricing scheme under which the firm charges different prices and provides varying service time (value) to different customers. It serves as a useful benchmark to compare and evaluate some commonly used pricing schemes. In

particular, we investigate two ubiquitous pricing schemes used in such services – fixed fee and time-based billing. Lowenhahl (2005), based on an extensive survey, suggests that these are the two most commonly used pricing schemes in professional services. In the fixed fee scheme all customers pay the same fee and get a consistent outcome. This is implicitly the pricing scheme widely studied in the queuing theory-based pricing literature but this literature does not consider the relationship between service value and service time. Some website design services charge flat fees and so do some legal consulting firms preparing wills and trusts. Another commonly used pricing scheme is the time-based scheme wherein the firm posts a certain fee or rate per unit time and consumers "consume" some service time based on their own needs. In fact, an important difference between products and services is that services can be sold based on the service time consumed. Several website design firms, legal firms preparing will and trusts and home-cleaning services charge based on time. Managers of website and graphic design services usually struggle with the choice of whether to charge a fixed or time-based fee.[1]

We also investigate how pricing schemes impact system performance in such discretionary services. To this end, we build a model wherein both the service rate and demand served are endogenous and in particular, they are functions of the pricing scheme. This novel feature of our model allows us to explore how the choice of pricing scheme impacts service system performance along various dimensions, including revenue, demand served, utilization, value delivered, congestion, etc. We consider a service setting wherein the service outcome is observable and verifiable, and the firm and customer can have an influence on the service time in each service encounter, as in the examples discussed earlier.

Several key insights emerge from our analysis, some of which are summarized here. Using the optimal pricing scheme, counter to conventional wisdom in the queuing literature that greater variance in service times degrades system performance, the firm can benefit from a more heterogenous valuation among customers. Both fixed fee and time-based schemes do well relative to the optimal pricing scheme in terms of revenue and they can exhibit better performance than the optimal pricing scheme on some important system metrics. Relative to the optimal scheme, the time-based scheme over-serves customers by providing a higher service value and charging customers more while the fixed scheme under-serves customers by providing a lower service value and charging customers less. Among the three pricing schemes, the optimal scheme serves the largest number of customers while the time-based scheme serves the fewest. Higher utilization is not always accompanied by greater congestion; the optimal and fixed-fee schemes have higher utilization than the time-based scheme but typically have less congestion. The choice of fixed or time-based fee depends critically on the shape of the distribution of customers' valuation of

[1] See http://rosmarin-search-marketing.com/blog/2010/07/07/the-hourly-rate-vs-flat-fee-debate/ and http://www.claytowne.com/beats-digging-ditches/flat-fee-versus-hourly-rates-how-to-charge-for-your-web-or-graphic-design-services/ for funny and colorful comparisons of fixed versus hourly pricing for website design and some related services.

service time and the responsiveness desired by consumers. Overall, this work provides some valuable guidance about the relative merits of fixed and time-based pricing schemes.

The remainder of this paper is organized as follows: Section 2 provides a literature review and in Section 3, we propose a model for discretionary services. In Section 4, we derive the optimal prices and service times for the three pricing schemes mentioned earlier, and compare them in terms of revenue performance. We analytically compare the three pricing schemes on some important metrics of operational performance in Section 5. We provide extensions of fixed and time-based pricing schemes in Section 6. Key insights derived from our analysis are discussed in Section 7.

2 Literature Review

There are two broad streams of literature related to our paper: (i) work on the interaction between the value from a service and service time, which has a recent and limited history and (ii) work on pricing in services which has a long history. Starting with (i), Chase (1981) referred to services where service time may be discretionary as having high customer contact. Karmarkar and Pitbladdo (1995) provide a framework for analyzing service processes and describe key characteristics of such services that are intangible. This type of service is termed as a "discretionary task" in Hopp et al. (2007) and as a "customer intensive service" in Anand et al. (2011). Perhaps the first formal model-based analysis of the first stream is the work by Hopp et al. (2007). Their focus is on the firm's dynamic control over the workload to maximize the value derived from a service but not the firm's pricing scheme which is our focus. They identify quality as an additional buffer to conventionally viewed buffers of time and capacity in service systems. Debo et al. (2008) study the "demand or service inducement" effect in "expert" or "credence" services where there is information asymmetry between the firm and the customers about the quality of the service, under a given time-based fee structure and provide insights on how to limit "service inducement". Both of the above papers focus on dynamic control of the workload by the service provider who adjusts the service time as a function of the system load. Also, both papers treat demand as exogenous while we treat customer demand as endogenously determined by the firm's choice of payment scheme and the prices therein. Further, the focus of Debo et al. (2008) is not on the service time-value relationship or pricing issues unlike in our work and their focus is on "expert" services where the service provider has an informational advantage over the customer and effectively determines the service time; examples are surgery, computer or appliance repair. Pinker and Shumsky (2000) explore the efficiency-quality trade-off in a stochastic service system with exogenous demand.

Anand et al. (2011) explore the firm's trade-off between quality and speed under an endogenous demand model. While they consider a fixed fee structure and assume that customers

have a homogenous service time-value relationship, we allow for customers to be heterogenous in their service time-value relationship and explore the impact of different pricing schemes on service performance. Also, variation in service time is exogenous and has no impact on service value in their model. The dynamic aspect of the quality-speed trade-off is analyzed in Kostami and Rajagopalan (2013). Akan et al. (2011) consider the speed-quality trade-off in a healthcare setting and relate service quality (or value) to the time spent by the physician (an "expert") in an endogenous demand model and consider asymmetry in information about physician skills. While we do not model information asymmetry issues, they assume that customers have a homogenous service time-value relationship and the firm commits to an expected service time. Also they do not consider pricing schemes such as time-based pricing. In earlier work, Lovejoy and Sethuraman (2000) pointed out that time and quality are substitutes and that there are speed-quality trade-offs in a manufacturing setting.

The literature on pricing in services is long and belongs to two broad categories: those that consider congestion and responsiveness as we do and those that do not. There is a long history of studies on congestion control in service queues in the presence of rational customers. To keep the literature review concise, readers are referred to Hassin and Haviv (2003) for a recent comprehensive review; a few recent examples are Chen and Frank (2004), Ata and Shneorson (2006), Randhawa and Kumar (2008), Bitran et al. (2008) and Cachon and Feldman (2011), Li et al. (2012). Ha (1998) and Ha (2001) are the first works to analyze incentive-compatible pricing schemes and admission control in settings wherein customers can choose their service time. In his models, customers trade off between their service time in the system (and the corresponding cost paid) and the effort put into the service, a tradeoff that is absent in our model. See Hsu et al. (2009) and the references therein for more studies on incentive-compatible pricing and optimal scheduling in service systems. Cachon and Feldman (2011) point out the value of comparing two commonly used pricing schemes, in their case per-use pricing and subscription pricing. The performance comparison among different pricing schemes (static or dynamic pricing) in a queueing framework with different customer classes is investigated in Hall et al. (2009). Ata and Shneorson (2006) is similar to our work in allowing both demand or price and service rates to be endogenous. In all the above works, service value is not a function of service time, a key element of discretionary services.

There is a literature on pricing of services that does allow for customers to obtain heterogenous utility from usage of a service. Representative examples of this usage-based pricing literature in both monopoly and competitive settings are Essengaier et al. (2002), Sundararajan (2004), and Bala and Carr (2010), but these works do not consider congestion effects, a key aspect of our model. Roels et al. (2010) compare the performance of contracting schemes such as fixed-fee, time and materials and performance-based contracts when managing collaborative services wherein the output is dependent on both the vendor's and clients' efforts which are

not verifiable. Unlike them, we do not consider moral hazard issues; on the other hand, Roels et al. (2010) do not consider the service time-value relationship, congestion effects, etc. Ren and Zhou (2008) study the choice of contractual forms in call center outsourcing under a queueing framework but they do not incorporate a service time-value relationship and assume exogenous demand.

The trade-offs between fixed versus time-based payment schemes have been explored in some industry domains. In the legal world, fixed fee was very common for legal services but gradually gave way to hourly billing. Polinsky and Rubinfeld (2003) and Shepherd and Cloud (1999) focus on the moral hazard or "agency" issue in legal services that arises because the law firm has an incentive to spend too much (too little) time in an hourly billing (fixed fee) scheme. But, unlike in our work, these works do not capture the service time-value relationship or congestion effects and demand is assumed to be exogenous. Welton and Dismuke (2008) provide persuasive evidence that nursing care time (also called nursing "intensity") varies significantly across patients and point to the need for variable billing based on service time rather than charging a fixed fee for nursing services but do not present an analytical model.

Overall, a key contribution of our work is investigating the impact of a widely used set of pricing schemes on system performance in discretionary services, which to the best of our knowledge is the first effort in this topic. Also, we contribute to the queue-based pricing literature by endogenizing both service rate and demand as functions of the pricing scheme.

3 Model

We start with an example to motivate the model – consider a firm that specializes in designing websites with similar content and scope. Customers arrive randomly to avail of this service. Each service encounter typically comprises of two phases. The first phase comprises of consultations during which the firm understands the customer's service needs and both the firm and customer learn about the time required to achieve a desired set of features and quality level, and/or some other preparatory tasks performed by the firm to create a barebones website. These tasks are necessary to the second phase of the service process and are common to all customers. In the second phase, the designer creates the website incorporating features desired by a customer such as a more colorful website, more animation or a more flexible site that can handle a variety of browsers, etc. So, the first phase represents the non-discretionary part of the service process and in the second phase, the firm performs discretionary tasks based on the customer's expectations discussed in the first phase. Additional features and enhancements to the website require more time but they add more value from the customer's perspective. The amount of incremental value derived is likely to be concave in the service time and the value will vary across customers. Let $\tau_0 > 0$ denote the time required for the initial phase which represents the minimum and non-

discretionary part of total service time (similar to the minimum service duration $\tau_{\min}$ in Hopp et al. (2007)), which is common to all customers and is not customer-dependent. Since τ_0 is a non-discretionary part of the service encounter and provides a common value to all customers, we normalize the value derived from the first phase to be zero and refer to the initial phase $(0, \tau_0)$ as a diagnosis phase henceforth for ease of exposition.[2] Let $\tau \geq \tau_0$ be the total service time for each service encounter and so the service time for the second phase is $(\tau - \tau_0)$.

A website design firm could use a variety of pricing schemes but many typically charge fixed fees (see www.123triad.com) or hourly fees (http://buildinternet.com/2009/12/a-discussion-on-hourly-rates-in-web-design/). A website designer may be busy at times and customers may have to wait for service, incurring a waiting cost. So, the full price paid by a customer consists of both the nominal price charged by the firm plus the congestion cost, as in the congestion pricing literature (Hassin and Haviv (2003)). The waiting time here does not necessarily refer to a customer's physical waiting experience. Rather, it is meant to capture the responsiveness of the service even in contexts where services may be scheduled based on appointments. For example, a customer who has to wait for several weeks for a website design because the firm is busy does not represent "physical waiting" but captures the level of responsiveness. Robinson and Chen (2011) refer to this type of waiting as "the second type of waiting costs (compared to the direct waiting cost)".

We propose a queueing model for this problem setting that captures the key elements described above and in which both demand rate and service rate are endogenous – we describe this model next[3]. A monopolistic firm serves customers who arrive according to a Poisson process. The market potential is assumed to be ample and the arrival rate (or demand served) is denoted by λ which is determined endogenously as a function of customers' net utilities as described next. Customers derive a net utility from the service which depends on the value they derive from the service, the amount they have to pay for the service and the time they have to wait to receive service. Each customer arriving for service is characterized by his type α (to be specified shortly) which is a random variable with a known density function $f(.)$, with mean $\hat{\alpha} = E(\alpha)$ and $\alpha \in [\underline{\alpha}, \bar{\alpha}](\bar{\alpha} > \underline{\alpha} > 0)$. α is unknown to both the customer and the firm before the customer goes through the diagnostic phase. During the diagnosis phase, the firm understands the customer's expectations and the customer understands the time it will take to achieve a certain service value and so α becomes known to both. At the risk of some abuse of notation but to simplify the exposition, we do not distinguish between the random variable α and its realization. Differences in α values represent variation in expectations and willingness to pay for quality and features. The gross value derived by a customer of type α from a service

[2]We can easily let the duration of the first phase τ_0 be a random variable without changing the following analysis and insights.

[3]This description of the model is partially based upon a summary of our model prepared (and graciously shared) by Professor Refael Hassin for a forthcoming survey on "Rational Queueing".

of total length τ is given by:

$$\begin{aligned} v(\alpha,\tau) &= \sqrt{\frac{\tau-\tau_0}{\alpha}},\ \text{if } \tau \geq \tau_0, \\ v(\alpha,\tau) &= 0,\ \text{if } \tau \in [0,\tau_0) \end{aligned} \tag{1}$$

The value function (1) captures the two key properties described earlier: (i) service value is increasing and concave with service time[4]; (ii) the value placed on service time varies across customers due to differences in service expectations, captured by the parameter α. While service value increases with service time in a concave fashion in the above model as in Hopp et al. (2007) and Anand et al. (2011), it is different along some important dimensions. First, we allow the value derived to differ across customers. Second, unlike in Anand et al. (2011), the maximum value derived is not assumed to be the same across all customers. We could allow for a more general service value versus service time function that is increasing and concave. For example, our main results and insights hold if the value function is generalized to $(\sqrt{\frac{\tau-\tau_0}{\alpha}+b^2}-b)$ where b is an additional parameter that can capture a finite slope at τ_0 for the value-time curve (1). But to keep the model parsimonious, we use (1) henceforth without loss of generality.

The firm charges a price $p(\alpha)$ for a customer of type α. We primarily consider three pricing schemes in our analysis but consider some extensions of these schemes in Section 6. First, we characterize the optimal pricing scheme wherein the firm charges a price $p(\alpha)$ and provides a service time $\tau(\alpha)$ with a corresponding value $v(\alpha,\tau)$ to customer type α. Then, we explore two commonly used pricing schemes in services: fixed fee and time-based fee (Lowenhahl (2005)). In the fixed fee structure, the firm charges every customer the same fee f and provides the same committed value v_f. This is the de facto pricing scheme studied in the queuing-based pricing literature: all customers get the same service value and pay the same price but with varying service times. But, unlike in traditional queue-based pricing literature where service time is exogenous, service time here is an outcome of heterogenous consumer valuations of service time and the pricing schemes adopted. If the firm uses a time-based fee structure, it charges the same rate r_t per unit of service time to all customers but each customer chooses the service value they desire and the corresponding service time. For instance, a website design firm may charge a certain rate per hour and customers choose the features they desire and equivalently the service time, based on the consultations in the initial diagnosis phase. We assume that the characteristics of the pricing scheme used are posted and known to the customer before they arrive for service, as is typical in the literature. The focus of this paper is on the impact of static pricing schemes on system performance and so we do not consider dynamic pricing policies. Also, the payment and service time for each customer does not depend on the system state. Each

[4]In fact, we can model the value-time relationship as a concave function (that need not be nondecreasing). Since both the firm and customers will not choose a service time that has negative marginal value of additional time, we can restrict ourselves to the part of the value-time curve that has a non-negative slope.

customer incurs a waiting cost upon arriving for service and it is linear in the wait time. The waiting cost is homogenous and is equal to $\beta' W(\lambda, g(\tau))$ where $\beta' \geq 0$ is the fixed waiting cost per unit waiting time and $W(\lambda, g(\tau))$ is the expected waiting time which is a function of the demand rate λ and the distribution of service times $g(\tau)$ that arises from the heterogeneity in consumer valuations. The monopolistic firm works as a M/G/1 queueing system operated on a First-Come-First-Serve (FCFS) basis.

A customer joins the unobservable queue if his net utility (service value minus price minus expected waiting cost) is nonnegative and we assume a mixed joining strategy on the part of the customers (as in Chapter 3, Hassin and Haviv (2003)). We note that customers are symmetric a priori when deciding to procure the service because their valuation of service time, i.e. α is realized only after the initial diagnosis. Also, consistent with many queuing models, customers do not leave the system once they join the queue.

The firm wishes to maximize its long run average rate of revenues under the constraint that the expected net utility of any joining customer is nonnegative. In our context, working faster or slower does not have an impact on the cost associated with service provision. So revenue-maximization is an appropriate objective for the firm. Its decision variables are the posted functions of price $p(\alpha)$ and service duration, $\tau(\alpha)$. All customers and the firm are assumed to be risk-neutral. The net value or utility obtained by a customer of type α who pays price, $p(\alpha)$, and receives service for a total duration $\tau(\alpha)$ is:

$$v(\alpha) - p(\alpha) - \beta' W(\lambda, g(\tau))$$

Let $E(\tau)$ denote the expected service time, $\mu = 1/E(\tau)$ the average service rate and $CV(\tau)$ the coefficient of variation in service times. So, the net utility for a customer of type α is given by[5]:

$$v(\alpha) - p(\alpha) - \beta' W(\lambda, g(\tau)) = \sqrt{\frac{\tau(\alpha) - \tau_0}{\alpha}} - p(\alpha) - \beta' \frac{\lambda}{(\mu - \lambda)\mu} \frac{(1 + CV^2(\tau))}{2}.$$

We note that both the average service rate μ and coefficient of variation in service times $CV(\tau)$ are functions of the distribution of α values. In the ensuing analysis, we replace $\beta' \frac{(1+CV(\tau)^2)}{2}$ with β where β incorporates the effect of $(1 + CV(\tau)^2)$ but does not vary with the pricing decision. This approximation is formally justified in Appendix B but an intuitive and simplified justification for the approximation is as follows. Even if there is a small variation in service times induced by different pricing schemes (and Appendix B shows that it is indeed very small), its impact on the coefficient of variation of the total system, including the Poisson arrival process and the initial service time τ_0, tends to be negligible. It is worth noting that we are not ignoring the variation in α and its impact on system performance and in fact, it plays a significant role in the choice of the pricing scheme.

[5]Note that we have taken customers' waiting time using Pollaczek-Khinchin (PK) formula, not sojourn time, as our measure of congestion because customers get negative utility primarily from waiting, not from the service time.

4 Comparison of pricing schemes

The key decision studied in this work is the fee structure chosen by the firm, which endogenously determines both the demand rate and service rate. Next, we explore the optimal prices and service times that will maximize revenue in each of the three pricing schemes mentioned above. We first analyze the optimal pricing scheme and then study two commonly used pricing schemes: the fixed and time-based schemes. For each pricing scheme, we first provide a formulation of the revenue maximization problem together with the market clearing or equilibrium conditions. We then provide a transformation of the revenue maximization problem that facilitates analysis and derivation of the optimal service times and prices. After characterizing each pricing scheme in Sections 4.1 to 4.3, we investigate the sub-optimality of the fixed and time-based schemes in terms of revenue relative to the optimal scheme. In the rest of the paper, we use subscripts "f" and "t", respectively, to represent parameters and variables associated with fixed and time-based payment schemes.

4.1 The Optimal Pricing Scheme

Based on our discussion in Section 3, under the optimal pricing, the monopolistic firm charges different prices $p(\alpha)$ and provides varying service times $\tau(\alpha)$ (and values) to different customer types such that the expected net utility is zero, i.e.

$$E[v(\alpha) - p(\alpha) - \beta\frac{\lambda}{(\mu-\lambda)\mu}] = 0. \tag{2}$$

Under the optimal pricing scheme, the firm can extract all the consumer surplus (Naor (1969)) because the queue is unobservable. The firm's objective is to determine the optimal prices and service times while satisfying the constraint (2) so as to maximize expected revenue, that is,

$$\begin{aligned} \max_{\{p(\cdot),\tau(\cdot)\geq\tau_0\}} R &= \int \lambda p(\alpha) f(\alpha) d\alpha = \lambda E(p), \\ s.t. \quad & (2) \end{aligned} \tag{3}$$

where $E(p)$ is the expected price.

Denote $r := \frac{E(p)}{E(\tau)}$, where $E(\tau) = 1/\mu$ denotes the average total service time. r can be viewed as a proxy rate that measures the average amount charged per unit service time across all customers. Then using (2) and noting that $E(v)$ is the expected value, we have:

$$E(p) = E(v) - \beta\frac{\lambda}{(\mu-\lambda)\mu} = rE(\tau) = \frac{r}{\mu}.$$

After some algebraic manipulation using the above equation, the market clearing or equilibrium demand rate is given by:

$$\lambda = \mu \frac{1}{1+\frac{\beta}{\mu E(v)-r}}.$$

So, the firm's revenue maximization problem can be written as:

$$\max_{\{r,\tau(\cdot)\geq\tau_0\}} R = \lambda E(p) = \lambda \frac{r}{\mu} = r \frac{1}{1+\frac{\beta}{\mu E(v)-r}}. \tag{4}$$

Note that $E(v)$ and μ are functions of $\tau(\alpha)$ but we suppress this for ease of exposition. Therefore, the firm's problem is to choose a value for r and the service times $\tau(\cdot)$. Define $\gamma := \frac{1}{\hat{\alpha}E(\frac{1}{\alpha})}$. Since $\frac{1}{x}$ is a convex function, we have $\gamma \in (0,1)$ due to Jensen's theorem. Note that γ is a function of the distributional shape of α. In particular, it becomes smaller when the variation of α is larger. More interestingly, it becomes smaller (bigger) when the distribution is right-(left-) skewed, i.e. with a longer right (left) tail.[6] Using the superscript * for all the optimal quantities hereafter, we then have the following characterization of the optimal policy.

Proposition 1. *In the optimal scheme:*

i) The optimal service time for customer type α is $\tau^*(\alpha) = \tau_0 + \tau_0 \frac{\gamma\hat{\alpha}}{\alpha}$ and $\mu^* = \frac{1}{2\tau_0}$; the service value for customer type α is $v^*(\alpha) = \frac{\sqrt{\tau_0\gamma\hat{\alpha}}}{\alpha}$ and $E(v^*) = \sqrt{\frac{1}{\gamma}}\sqrt{\frac{\tau_0}{\hat{\alpha}}}$;

ii) Define $B := \mu^* E(v)^* = \frac{1}{2\sqrt{\tau_0\hat{\alpha}\gamma}}$, the optimal effective rate per unit of time is $r^* = B + \beta - \sqrt{(B+\beta)\beta}$, and the optimal expected revenue is $R^* = \beta(\sqrt{\frac{B+\beta}{\beta}} - 1)^2$. Further, $E(p)^*$ decreases in β and increases in τ_0;

iii) The optimal revenue R^* decreases in γ, and convexly decreases in β and τ_0.

Note that the optimal revenue is derived based on the effective rate per unit of time r^*. The corresponding price for each customer type can be implemented in a variety of ways, given the optimal expected price $E(p)^*$. In particular, a plausible condition is one under which each customer gets zero utility ex-post. The following Corollary states without proof the optimal price for each customer type under this condition.

Corollary 1. *Suppose $v(\alpha) - p(\alpha) - \beta\frac{\lambda}{(\mu-\lambda)\mu} = 0$ for any α, then the optimal price charged to customer type α is $p^*(\alpha) = \frac{\sqrt{\tau_0\gamma\hat{\alpha}}}{\alpha} - 2\tau_0[\sqrt{(B+\beta)\beta} - \beta]$.*

In the optimal scheme, the firm charges a lower price and provides less service time to customers with higher α, i.e. to customers who derive lower incremental value from a given amount of service time. Note that the marginal value of service time decreases for all customers as additional service time is provided but this value is lower for customers with higher α. Hence, customers with higher α receive less service time and correspondingly less value; so, they are charged a lower price. Using the website design example, customers that have higher α will get less service time and will get fewer enhancements or a less flexible website. Interestingly, the

[6] Suppose α is a Beta distribution $Beta(a,b)$ with domain (0,1), where $a > 2, b > 2$ and both are integers, it can be verified that $\hat{\alpha} = \frac{a}{a+b}$, the square of the coefficient of variation is $\frac{b}{a(a+b+1)}$, and $E(\frac{1}{\alpha}) = \frac{a+b-1}{a-1}$. Thus $\gamma = \frac{(a+b)(a-1)}{(a+b-1)a}$. Fixing a, we can see that γ becomes smaller when b increases, i.e distribution becomes more right-skewed and has higher variation.

service time and value in the optimal scheme are closely linked to τ_0. Since the initial service time τ_0 is similar to a minimum service time, higher values of τ_0 prevent the firm from serving more customers to avoid higher congestion costs. Furthermore, a larger value of τ_0 results in a longer service time for each service encounter so as to use capacity more effectively. This is similar in spirit to having larger batch sizes when the setup time increases in production lot sizing models. In Proposition 1, the optimal average service time is $2\tau_0$ due to our specific choice of value-time curve (1). Adding a parameter to the value-time curve (for instance, letting $v(\tau) = \kappa\sqrt{\frac{\tau-\tau_0}{\alpha}}$, where $\kappa > 0$) can be used to change the ratio of service time to τ_0 so as to adapt the model to a variety of business settings without changing the underlying rationale. In the website design example, an implication of the Proposition is that a designer that spends more time with customers in the initial consultation phase will tend to spend more time and provide greater value on average during the second phase, i.e. the discretionary part of the service that provides additional features and quality enhancements. This will in turn mean that she will have to charge a higher price and serve fewer customers.

It is interesting that the optimal revenue increases as γ decreases; as noted earlier, γ becomes smaller when the service time valuation distribution, captured by α, is more heterogenous or right-skewed. So, given a mean value of α, a more heterogenous set of value-time curves creates more opportunity for the firm to differentiate among customers in terms of service values and prices and achieve higher revenues. Thus, unlike in a traditional queuing model where greater variation in service times deteriorates system performance, greater variance in customer valuations (which translates into higher variance in service times) generates higher revenues here. Furthermore, skewness of the distribution matters too; if there are many customers with low α values but a few with very high α values (i.e. a right-skewed distribution), this will result in higher revenues too.

4.2 Fixed-fee scheme

In the fixed-fee scheme, the firm charges the same price f and provides the same value v_f, i.e. a consistent level of satisfaction to all customers. Given any (f, v_f), the service time is $\tau_f(\alpha) = \tau_0 + \alpha v_f^2$ for customers of type α. Expected service time $E(\tau_f)$ and expected service rate μ_f are given by:

$$E(\tau_f) = \tau_0 + \hat{\alpha} v_f^2,\ \mu_f = \tfrac{1}{\tau_0 + \hat{\alpha} v_f^2}.$$

As in § 4.1, the market-clearing condition is obtained by setting the expected net utility equal to 0:

$$v_f - f - \beta \frac{\lambda_{ef}}{(\mu_f - \lambda_{ef})\mu_f} = 0 \tag{5}$$

which can be solved to obtain the equilibrium demand rate λ_{ef} as:

$$\lambda_{ef} = \frac{(v_f - f)\mu_f^2}{\mu_f(v_f - f) + \beta}.$$

Similar to the approach used in Section 4.1, we make the following transformation: Let $r_f := \frac{f}{E(\tau_f)}$. Then we have the following problem formulation that maximizes revenue, after some algebra and rearrangement of terms as in section 4.1:

$$\max_{\{r_f, \tau_f(.) \geq \tau_0\}} R_f = \lambda_{ef} r_f E(\tau_f) = \lambda_{ef} r_f / \mu_f = r_f \frac{1}{1 + \frac{\beta}{v_f \mu_f - r_f}} \tag{6}$$

where v_f and μ_f are both functions of $\tau_f(.)$.

Proposition 2. *In the fixed-fee scheme:*

i) the service time for a customer of type α is $\tau_0(1 + \frac{\alpha}{\hat{\alpha}})$, $\mu_f^* = \frac{1}{2\tau_0}$, $v_f^* = \sqrt{\frac{\tau_0}{\hat{\alpha}}}$;

ii) the optimal fee charged is $f^* = 2\tau_0(b + \beta - \sqrt{(b+\beta)\beta})$, with the effective rate per unit time $r_f^* = b + \beta - \sqrt{(b+\beta)\beta}$, where $b = \frac{1}{2\sqrt{\tau_0 \hat{\alpha}}}$. f^* decreases in β and increases in τ_0;

iii) the optimal revenue is $R_f^* = \beta(\sqrt{\frac{b+\beta}{\beta}} - 1)^2$ which is independent of γ and convexly decreases in β and τ_0.

Some interesting observations emerge when we compare the fixed fee scheme with the optimal pricing scheme. While the *average* service time is the same in both schemes, the service time is higher for low valuation customers (with higher α) in the fixed fee pricing scheme, while it is lower in the optimal scheme. For instance, in the fixed fee scheme, the website design firm will spend more time with higher α customers to achieve the same level of satisfaction with their website. This is in sharp contrast to the optimal pricing scheme which spends *less* time with the high α customers. Thus, the fixed fee scheme effectively charges a lower rate per unit time to low valuation customers and subsidizes these customers at the expense of the high valuation customers. Of course, from the customer's point of view, since they all receive the same value and are charged the same fee, the fixed fee scheme may appear to be more *equitable*. An increase in the congestion penalty β results in a lower price charged to all consumers but does not impact average service time. Thus, the higher congestion penalty is absorbed completely by lowering the price and does not impact service times. However, an increase in β (or τ_0) does impact the revenue performance of the fixed fee scheme relative to the optimal scheme, as will be discussed later. Interestingly, unlike the optimal scheme, the optimal revenue in the fixed fee scheme does not depend on γ and this suggests the limitation of the fixed fee scheme in not effectively exploiting the heterogeneity and skewness in customer valuations of service time. From Propositions 1 and 2, the average service time in the optimal and fixed fee schemes are identical while $\frac{E(v^*)}{v_f^*} = \sqrt{\frac{1}{\gamma}}$. So, γ represents the extent to which the optimal scheme can extract additional service value while having the same average service time as the fixed scheme.

4.3 Time-based scheme

In the time-based scheme, the firm charges a rate r_t per unit time. Customers, *after* going through the initial service phase τ_0, make the optimal choice of the service time τ_t depending on the rate r_t charged by the firm. Note that unlike in the optimal pricing scheme and the fixed fee scheme, the customer decides the service time. In fact, if the firm were to decide the rate and the service time for the customers, this is equivalent to the optimal pricing scheme based on our formulation of the maximization problem (4) in Section 4.1 (in this case, the proxy rate r^* can be treated as a real rate). In other words, when a website design firm posts a rate per hour and determines the time required to design a customer's website, this is equivalent to the optimal pricing scheme. So, here we model commonly observed real-world scenarios where the firm posts a rate but the customers have a say about the consumption of service time depending on their needs. For instance, a website design firm may let the customer decide the design features and enhancements, wherein the customer is fully aware of the time required for such enhancements and so the customer effectively makes the service time decision. In some service contexts, the control over service time consumption may be exercised by the firm or customers depending on the pricing scheme adopted. A recent example of such an approach, in a different setting than ours, can be found in Cachon and Feldman (2011), wherein a subscription scheme loses direct control over customers' visit frequency to a service facility while a per-use pricing scheme keeps this control. The optimal τ_t is the solution to:

$$\tau_t = \arg\max_{\{\tau_t \geq \tau_0\}} (v(\tau_t) - r_t(\tau_t - \tau_0)) = \tau_0 + \tfrac{1}{4r_t^2\alpha}.$$

Thus, the service time is higher for customers with a smaller α, as in the optimal pricing scheme and unlike in the fixed fee scheme. The net value (value less price) for a customer of type α is then $\frac{1}{4r_t\alpha} - r_t\tau_0$. The expected service time $E(\tau_t)$, expected service rate μ_t and the expected service value $E(v_t)$ are given by:

$$E(\tau_t) = \tau_0 + \tfrac{1}{4r_t^2\hat{\alpha}\gamma},\ \mu_t = \tfrac{1}{E(\tau_t)},\ E(v_t) = \tfrac{1}{2r_t\gamma\hat{\alpha}}.$$

As before, we use the market clearing condition to get the equilibrium demand rate λ_{et}. Then, after some algebra, the firm's revenue maximization problem is given by:

$$\max_{\{r_t\}} R_t = r_t\frac{\lambda_{et}}{\mu_t} = r_t\frac{1}{1 + \frac{\beta}{\mu_t E(v_t) - r_t}}. \tag{7}$$

Proposition 3. *The revenue function $R_t(r_t)$ is a concave function, and the optimal $r_t^* \in (r_1, r_2)$, where $r_1^2 = \frac{\sqrt{2}-1}{k}$, $r_2^2 = \frac{1}{k}, k := 4\tau_0\gamma\hat{\alpha}$. Also, r_t^* decreases in γ, β as well as in τ_0.*

Thus, an increase in the initial diagnosis time τ_0 results in a lower rate charged and correspondingly, the average service time increases. In turn, this results in higher value provided and a higher price charged to customers. In response to an increase in the congestion penalty

β, the firm lowers the rate and correspondingly the service time increases. Thus, unlike in the fixed fee and the optimal pricing schemes, an increase in the congestion penalty does impact service times and interestingly results in higher rather than lower service times. The decrease in the optimal rate with γ suggests another interesting phenomenon – recall that a decrease in γ implies a more heterogenous or right-skewed distribution of the customer valuation of service time. Thus, the optimal rate in the time-based scheme is higher when customer valuations are more heterogenous or more right-skewed. This is because the firm has the flexibility to discourage the low valuation (high α) customers from consuming more service time without deterring too many high valuation customers by raising the rate per unit time.

While a closed-form solution for the optimal rate cannot be obtained unlike in the optimal and fixed fee schemes, we can use the range (r_1, r_2) within which r_t^* lies to assess how the optimal rate in the time-based scheme compares with that in the fixed fee and optimal pricing schemes. A comparison among these three rates, while interesting in itself, has a significant impact on system performance, as discussed in the next Section.

Proposition 4. $r_t^* \geq r^* \geq r_f^*$.

Thus, the rate charged in the time-based scheme is always greater than the effective average rate charged in the optimal pricing scheme which in turn is greater than that in the fixed fee scheme. From Propositions 1 and 2, the structure of r^* and r_f^* are identical except that B in r^* is replaced by b in r_f^*, where $B = \frac{1}{\sqrt{\gamma}} b$. So the difference in these two rates is determined solely by γ, which in turn is determined by the distribution of α and the concavity of the service value-time curve. As variation in α increases, smaller is the value of γ and greater is the gap between the rates. Why is the rate in the time-based scheme never smaller than the effective rate in the optimal scheme? If the average rate were lower in the time-based scheme, then customers would consume too much service time since they optimize their own net value, ignoring the effect on waiting time and the potential for greater value and revenue that can be elicited from other customers by the firm. Hence, the firm charges a higher effective rate.

4.4 Comparison of revenues among the pricing schemes

Now we compare the revenue performance of the pricing schemes. As discussed in section 4.2, the optimal scheme is more effective in terms of value and capacity allocation by better exploiting the concavity of the value-time curve and the heterogeneity in customer valuations of service time. In addition, the effective rate per unit time r^* in the optimal scheme is greater than that in the fixed scheme r_f^* as indicated in Proposition 4. These differences are accentuated depending on the shape of and variability in the distribution of α. Recall that R^* decreases in γ while R_f^* is independent of γ, so γ or equivalently the distribution of α is a key factor that drives the ability of the optimal scheme to achieve higher value and translate this into higher revenues.

The parameters β and τ_0 also impact the relative revenues as shown in the next result.

Proposition 5. *For any given τ_0, the relative revenue difference $\frac{R^*}{R_f^*}$ increases in β. For any given β, $\frac{R^*}{R_f^*}$ increases in τ_0.*

As seen from Propositions 1 and 2, both R^* and R_f^* decrease in β and τ_0. However, compared to the fixed fee scheme, the revenue of the optimal scheme diminishes at a relatively slower rate as β becomes greater. This is because the optimal scheme provides a higher service value so that customers are more tolerant of congestion, as compared to the fixed scheme. So the relative disadvantage of the fixed scheme is exacerbated in the presence of stronger sensitivity to congestion. The same phenomenon occurs at higher values of τ_0 for similar reasons.

While the fixed scheme is inferior to the optimal scheme, both of these schemes possess two levers – controlling service time *and* price. In contrast, the time-based scheme possesses just one lever: the rate charged per unit time. However, compared with the fixed scheme, the time-based scheme does exploit the heterogeneity in α and the concavity of the value-time curve. This is because low valuation customers (with higher α) will refrain from consuming more service time. Notice that service times in the time-based scheme decrease with α as in the optimal scheme, unlike the fixed fee scheme wherein higher service time is allocated to customers with higher α. The following result compares the revenue between the fixed and time-based schemes.

Proposition 6. *There exist a set of thresholds for γ, β and τ_0 which determine whether the time-based or fixed fee scheme has higher revenue. That is, there exist $\bar{\gamma}(\beta, \tau_0)$, $\bar{\beta}(\gamma, \tau_0)$ and $\bar{\tau_0}(\gamma, \beta)$ such that when $\gamma < \bar{\gamma}(\beta, \tau_0)$, or $\beta < \bar{\beta}(\gamma, \tau_0)$, or $\tau_0 < \bar{\tau_0}(\gamma, \beta)$, we have $R_t^* > R_f^*$. Otherwise, we have $R_f^* \geq R_t^*$.*

Whether the time-based or fixed fee scheme dominates in terms of revenue depends on which of two forces prevail – the direct control over two levers (rather than one) in the fixed fee scheme versus the ability of the time-based scheme to exploit customers' heterogenous value-time relationships. For instance, greater variability in α implies a lower value of γ and in this case, the advantage of the time-based scheme in terms of allocating service time more effectively among customers is harnessed to a greater degree; the time-based scheme consequently can provide a higher average service value for the same average service time. When this value-gaining effect in the time-based scheme is strong, represented by a small γ, the disadvantage due to the inability to control service times (unlike in the fixed-fee scheme) tends to be dominated. On the other hand, when this value-gaining effect is weak, the absence of control over service time hurts the time-based scheme. Taking the extreme case of deterministic α, where $\gamma = 1$, the fixed scheme achieves the optimal solution and thus dominates the time based scheme. Thus, when the distribution of α is more heterogenous or right-skewed, implying a smaller γ, the time-based scheme dominates. For example, if there are a few website design customers who are willing to pay very little for enhancements (i.e. with high α), then a time-based scheme may be a better

choice.

The values of β and τ_0 also influence the relative dominance of the time-based and fixed fee schemes. As shown before, the average service time in the time-based scheme actually increases in β because the rate r_t^* decreases in β. So the congestion in the time-based scheme is aggravated by a longer service time when the congestion penalty β is higher. In contrast, the average service time in the fixed scheme is independent of β. Hence, when β becomes sufficiently large, although the demand served decreases in both the time-based and fixed schemes, the service rate decreases in the time-based scheme while it stays constant in the fixed-scheme. As a result, the revenue in the time-based scheme begins to diminish at a faster rate than in the fixed fee scheme when β becomes greater. So the time-based scheme becomes less attractive in the presence of a large congestion penalty. Similarly, a smaller τ_0 tends to speed up the service process which can counter the tendency of the time-based scheme to have longer service times and mitigate the higher congestion.

Table 1: Sub-optimality of the Fixed and the Time-based Pricing Schemes

Minimum Service-time	Congestion Penalty	α Uniform Distributed (0,1)		α Triangular Distributed (0,1)	
		Fixed	Time-based	Fixed	Time-based
τ_0	β'	$\frac{R-R_f}{R}$	$\frac{R-R_t}{R}$	$\frac{R-R_f}{R}$	$\frac{R-R_t}{R}$
0.1	1	2.5%	0.4%	1.6%	0.3%
	3	3.0%	1.1%	1.7%	1.1%
	7	2.8%	2.4%	1.8%	2.4%
0.3	1	2.6%	0.6%	1.7%	0.6%
	3	2.8%	1.8%	1.8%	1.8%
	7	3.0%	4.1%	1.9%	4.2%
0.7	1	2.6%	0.9%	1.7%	0.9%
	3	2.9%	2.8%	1.8%	2.8%
	7	3.2%	5.9%	2.0%	4.5%

Note: R_f, R_t and R represent the optimal revenue obtained from the pure fixed and time-based scheme and the optimal scheme, respectively.

We performed numerical experiments to compare the revenues of the fixed and time-based pricing schemes with each other and with the optimal scheme as a function of the two parameters β and τ_0. We considered the following set of values $\{1, 3, 7\}$ for β and $\{0.1, 0.3, 0.7\}$ for τ_0. Also, we considered a discrete distribution for α with 20 different discrete values ranging from 1 to 2 and considered both the uniform and triangular distributions for α. The triangular distribution was used as it is widely used in project management to represent task duration. The uniform distribution was used because of its larger spread and variance which is valuable in checking the robustness of the approach. The results (see Table 1) suggest that the fixed fee and time-based pricing schemes perform quite well relative to the optimal pricing scheme. The maximum optimality gap in revenue for the time-based pricing scheme is 5.9% while the maximum gap is 3.2% for the fixed fee scheme. Note that both of the maximum gaps occur when α has

a uniform distribution, which is less likely in reality. The optimality gap for the time-based scheme is greater at higher β and τ_0 values while it does not appear to show any systematic relationship with β and τ_0 for the fixed fee scheme.

5 Operational Performance

While revenue is clearly an important measure for comparing the performance of the pricing schemes, managers and policy makers routinely use other important performance measures such as value provided to customers, utilization, demand served, congestion, etc. to evaluate and manage service systems. In governmental and non-profit organizations, number of customers served and congestion are important measures. Furthermore, although we have shown that the gap in revenue between the fixed and time-based schemes and the optimal scheme is not large, it is not clear if this is true for other measures of operational performance. For example, can the demand served and gross service value, which are important dimensions for a system, be substantially different across the three pricing schemes even if their revenues are similar? Next, we compare the three pricing schemes in terms of commonly used operational performance metrics.

In discretionary services, there is a speed-value trade-off because speeding up the process allows the firm to serve more customers and/or reduce congestion but it reduces the gross value $v(\tau)$ derived by the consumers. So, we explore the impact of pricing schemes on the gross value.

Proposition 7. *We have: (i)* $E(v_t^*) \geq E(v^*) \geq v_f^*$*; and (ii)* $\mu_t^* \leq \mu_f^* = \mu^*$.

The optimal scheme does not provide the highest gross value but is clearly the best at balancing the trade-off between speed (service rate) and value. The time-based scheme provides the longest service time and correspondingly the highest average service value. As shown earlier, the time-based scheme can sometimes have higher revenues than a fixed fee scheme and so both the firm and customers are better off in some scenarios. The time-based scheme "errs" on the side of providing too much value. The fixed fee scheme provides less average value than the optimal scheme (even though average service times are the same in both schemes) because the fixed fee scheme allocates too much service time to low-valuation (high α) customers and so the low α customers end up "subsidizing" the high α ones, dragging down the average value provided. As shown in the proof of Proposition 7 (see Appendix A), the relative disadvantage of the fixed scheme in terms of a lower service value increases when γ is smaller, i.e. when the distribution of α is more variable and/or right-skewed.

The next result sheds light on the relative dominance of the payment schemes in terms of several key metrics. Let RC denote revenue per customer (RC) – this metric is used by the sales function and is especially useful in discretionary services as all customers are not alike and some customers value the service more and may consume more and thus provide greater revenue.

Proposition 8. *We have: (i)* $RC_t^* \geq RC^* \geq RC_f^*$*; (ii)* $\rho^* \geq \rho_f^* \geq \rho_t^*$*; (iii)* $\lambda^* \geq \lambda_f^* \geq \lambda_t^*$*; and (iv) Both utilization rate* (ρ) *and demand served* (λ) *in all three pricing schemes decrease in* β *and* τ_0*.*

Recall that the time-based scheme has the highest rate charged per unit time. The higher rate charged together with the longer service time results in a higher RC for the time-based scheme. While the fixed fee scheme has the same average service time as the optimal scheme, the effective average rate charged is lower in the fixed fee scheme and so it ends up having lower RC. Despite the higher RC, the time-based scheme does not always have higher total revenue than the fixed fee scheme because it has lower demand served.

In the optimal pricing scheme, the optimal choice of price and service time allows the firm to achieve the highest utilization (ρ) while achieving the maximum revenue. Interestingly, the fixed fee scheme always achieves higher utilization than the time-based scheme despite having lower average service time. The impact of β on utilization is as expected – a higher β will result in lower congestion at the optimum and this can be achieved only by reducing the utilization. An increase in τ_0 also decreases utilization levels and the rationale is interesting. Recall that the firm has no control over τ_0, the initial service time, and from our results, optimal total service time τ to each customer is proportional to τ_0. Hence, the firm has to compensate for the longer average service time (which will increase congestion) by lowering the utilization.

The number of customers served by the system (λ) is an important measure in scenarios where the demand potential is high but the service system has limited capacity and cannot serve everyone. Recall that the effective rate charged in the time-based scheme is highest followed by the optimal scheme and the fixed fee scheme (Proposition 4). One would expect this to have an inverse effect on the demand served. Interestingly, demand served is not lower in the optimal scheme relative to the fixed fee scheme despite the higher effective rate charged per customer. Demand served is lowest in the time-based scheme primarily because this scheme has high average service times. The firm compensates for this by having lower utilization levels but the lower utilization combined with a lower service rate implies that the demand served is lowest for the time-based scheme. Thus, the time-based scheme "pays" for the longer service time by having lower number of customers served which in turn negatively impacts its total revenue. This negative impact is even more significant at higher β and τ_0 values and the time-based scheme does worse than the fixed fee scheme.

In scenarios where a decision-maker's focus is on serving as many customers as possible (without losing the focus on revenue), the optimal scheme comes out best. But the fixed fee scheme does quite well in terms of revenue and demand served and may be a good compromise if the optimal pricing scheme is difficult to implement in such environments. From the results of the numerical experiments discussed in Section 4, we find that both demand served and service rate can be over 10% lower in the time-based scheme as compared to the fixed scheme,

even though revenues in both pricing schemes are close to each other (e.g., within 1.5% for $\alpha \sim U(1,2)$, $\tau_0 = 0.1$, $\beta = 1$). This finding is valuable because it shows that a firm can focus on performance metrics other than revenue, without sacrificing much revenue. The relatively slight change in revenue, despite substantial differences in demand served or service rate, comes from the discretionary nature of the service. For instance, in the time-based scheme, although the demand served is smaller than in the fixed scheme, customers "consume" more service time, get higher gross value or quality and pay more if the distribution of α is more variable and/or right-skewed; so the combined effect is such that the revenues are not significantly different.

Next, we compare the congestion levels across the different pricing schemes.

Proposition 9. *There exist thresholds for the waiting cost β and the diagnosis time τ_0 which determine the order of waiting times for the three pricing schemes. Specifically,*

i) $W_t^* \geq W^* \geq W_f^*$ when $\beta \leq \bar{\beta}_2$ for any given τ_0, or $\tau_0 \leq \bar{\tau}_2$ for any given β;

ii) $W^* \geq W_f^* \geq W_t^*$ when $\gamma \geq 0.828427$, $\beta \geq \bar{\beta}_1$, or $\tau_0 \leq \bar{\tau}_1$. Further, $\bar{\beta}_1 \geq \bar{\beta}_2$ for any given τ_0, $\bar{\tau}_1 \geq \bar{\tau}_2$ for any given β;

iii)$W^* \geq W_t^* \geq W_f^*$ otherwise.

In most realistic scenarios under which β and/or τ_0 are not too large, we have $W_t^* \geq W^* \geq W_f^*$. One might expect higher utilization levels to result in greater congestion in a single-server system setting but interestingly we observe that this is not often the case. Recall that the time-based scheme has the lowest utilization level. The fixed fee scheme generally has lower congestion than the time-based scheme despite having higher utilization. The optimal pricing scheme has the highest utilization but may have lower congestion than the time-based scheme, although it always has higher congestion than the fixed fee scheme. We also used the numerical tests described in section 4 to explore differences in congestion levels in the three pricing schemes since the results in Proposition 9 are parameter dependent. For τ_0 values between 0.1 and 0.7 and β values between 1 and 7, we find that congestion in the time-based scheme was highest followed by the optimal scheme and fixed fee scheme in all instances. Thus, the time-based scheme seems to have the highest congestion in most realistic scenarios.

Table 2 summarizes the operational performance of the three pricing schemes characterized above.

Table 2: Comparison of Operational Performance

Operational Metric	Optimal Scheme	Fixed Scheme	Time-based Scheme
Service Value	medium	lowest	highest
Service Rate	medium	medium	lowest
Revenue Per Customer	medium	lowest	highest
Utilization	highest	medium	lowest
Demand	highest	medium	lowest
Waiting Time(in most cases)	medium	lowest	highest

6 Extensions

There exist several real-world pricing schemes that are variants of the fixed and time-based schemes studied above and next, we discuss two variants. First, we consider a variant of the fixed pricing scheme wherein the firm commits to a service value but restricts the maximum service time provided. The fixed pricing scheme studied in Section 4.2 is a special case of this scheme in the sense that customers' service time is not restricted. Second, we consider adding a fixed fee to the time-based pricing scheme. Given that the revenue of the fixed and time-based pricing schemes studied earlier are found to be close to the optimal scheme in many plausible settings, the improvement in revenue of these variants is likely to be modest. We use subscript 2 to denote any variables associated with the variant of fixed and time-based pricing schemes in this Section.

6.1 Fixed fee with a maximum service time

When the firm restricts the maximum service time $\bar{\tau}$ in the fixed pricing scheme, the service value is committed to all joining customers with a restriction that the service time will not go beyond $\bar{\tau}$. That is, the firm posts a set of $(v_{f2}, f2, \bar{\tau})$, where v_{f2} represents the service value committed to the customers only if their service time does not violate the maximum service time $\bar{\tau}$, and $f2$ denotes the price paid. Note that the fixed scheme studied in Section 4 is a special case in the sense that $\bar{\tau} > \tau_0 + \bar{\alpha} v_{f2}^2$ based on our model of value-time curve.

For the sake of tractability, we assume there are two types of customers, i,e, $\alpha = \alpha_L$ with probability q, and $\alpha = \alpha_H$ with probability $1-q$, with the expectation $\hat{\alpha} = q\alpha_L + (1-q)\alpha_H$.We restrict ourselves to the case such that customers with type α_L will get the committed service value of v_{f2} and slow customers (of type α_H) will get service value of $\sqrt{\frac{\bar{\tau}-\tau_0}{\alpha_H}}$ reflecting the imposed maximum service time. Denoting $\tau_1 = \bar{\tau} - \tau_0$, the expected service time $E(\tau_{f2})$, service rate μ_{f2} and service value $E(v_{f2})$ are given by,

$$E(\tau_{f2}) = q(\tau_0 + \alpha_L v_{f2}^2) + (1-q)\bar{\tau} = \tau_0 + q\alpha_L v_{f2}^2 + (1-q)\tau_1,\ \mu_{f2} = \tfrac{1}{E(\tau_{f2})},$$
$$E(v_{f2}) = q v_{f2} + (1-q)\sqrt{\tfrac{\tau_1}{\alpha_H}}.$$

We use an approach similar to the one for fixed fee scheme in section 4.2 to derive the equilibrium condition and the revenue maximization problem – the details can be found in Appendix A. We then have the following result which compares this variant of the fixed fee scheme to the original fixed fee and the optimal schemes on key measures of operational performance.

Proposition 10. *Assuming a binary distribution for α, with a maximum service time imposed, the committed service value $v_{f2}^* \geq v_f^*$. Further, the average service value $E(v_{f2}) = E(v^*) \geq v_f^*$, and the utilization rate $\rho^* \geq \rho_{f2}^* \geq \rho_f^*$, and the demand served $\lambda^* \geq \lambda_{f2}^* \geq \lambda_f^*$.*

Recall that the fixed scheme (without a maximum service time imposed) studied in previous sections deviates from the optimal scheme by 1): providing lower gross service value, 2) possessing a lower utilization rate, and 3) serving less demand. Proposition 10 shows that with a maximum service time imposed, this underperformance will be mitigated and in particular, the fixed scheme with a maximum service time imposed can achieve the same average service value as the optimal scheme. This mitigating effect is driven by the fact that the fixed pricing scheme by its nature lacks direct control over service times among customers but this weakness is partially overcome by directly imposing a maximum service time.

6.2 Time-based fee plus a fixed fee

While many service providers charge a pure time-based fee, some of these firms often tack on a fixed fee. For example, landscape design and website design firms may charge a fixed (or a minimum) fee and then an hourly rate. Similarly, some law practices charge a flat fee plus an hourly rate [7](Robertson and Calloway (2008)). Such pricing is also often referred to as a two-part tariff in other contexts. We now analyze such a pricing scheme that comprises of a fixed fee F_2 for the initial phase $(0, \tau_0)$ and a time-based rate r_2 for the main phase. For conciseness, we do not describe the model again since it is identical to the time-based model described earlier (section 4.3) except for the addition of the fixed fee F_2. As before, customers having decided to procure the service and *after* going through the initial service phase τ_0, make the optimal choice of the additional service time depending on the rate r_2 and it is equal to $\frac{1}{4r_t^2\alpha}$. We then have the following result on the optimal two-part tariff.

Proposition 11. *The optimal two-part tariff* (F_2^*, r_2^*) *is given by* $F_2^* = 2\tau_0[B+\beta-\sqrt{(B+\beta)\beta}]-\frac{\sqrt{\tau_0}}{2\sqrt{\gamma\hat{\alpha}}} \geq 0$ *and* $r_2^* = \frac{1}{2\sqrt{\tau_0\gamma\hat{\alpha}}}$. *Further, this two-part tariff has the same optimal service time, expected service rate and optimal revenue as the optimal scheme.*

It is interesting to see that the two-part tariff achieves the same outcome as the optimal scheme. This result is consistent with the literature in economics and operations management on congestion-based pricing (Bitran et al. (2008) and the references therein) which has shown that the two-part tariff is optimal in many scenarios. However, this literature does not incorporate the relationship between service time and service value that is a key element of discretionary services and so our result extends previous results in the literature to this important setting. Thus, the two-part tariff which is essentially a combination of the fixed and time-based pricing can achieve the same outcome as the optimal scheme and is indeed an appealing alternative.

[7] http://members.mobar.org/billablehours/Appendix_D.htm

7 Discussion and Conclusion

7.1 Discussion

The optimal scheme is clearly ideal in the sense that it maximizes revenue and utilization and also serves the largest number of customers. Further, it generates only a moderate level of congestion and provides a moderate level of value to consumers by appropriately trading off service speed and value. However, even though they may yield lower revenues, the fixed fee and time-based schemes are frequently used in practice. This may be because they are: (a) simpler to implement or (b) perceived more positively by the consumer as the fixed fee scheme commits to a certain value and the time-based scheme cedes control over service time to the customer. Our analysis has identified the limitation of these simpler schemes: the fixed scheme does not allocate the service time among heterogenous customers effectively; the time-based scheme is more effective at allocating time among different customers, but the firm only has a single lever, the rate per hour. Our study indicates that the relative merit/drawback of these pricing schemes is strongly influenced by the shape of the distribution of customers' valuation of service time and the responsiveness desired by customers. We have also shown that a two-part tariff pricing structure can achieve the same outcome as the optimal scheme. However, Cachon and Feldman (2011) point out in their comparison of subscription and per use pricing for shared facilities, which shares some similarity to fixed fee and time based pricing in our context, that a two-part tariff may not always be desirable and this is perhaps why simple schemes such as fixed fees and hourly billing are used in practice. Furthermore, our numerical study suggests that the loss in revenue from using a simple fixed or time-based pricing schemes is likely to be small in many scenarios due to the inherent speed-value trade-off in discretionary services. Further, the fixed fee and time-based schemes exhibit superior performance along some other important operational metrics. For instance, the time-based scheme can provide higher service value than the optimal scheme for a slight sacrifice in revenue. Thus, our study can be seen as providing some justification for the widespread use of fixed fee and time-based schemes in practice despite their revenue-suboptimality.

Our study also provides some insights to help a firm choose between the fixed fee and time-based schemes. When the initial consultation/diagnosis time is likely to be short relative to the total service time (for example, a home cleaning service), then a time-based scheme is better. On the other hand, when initial consultation phase is long relative to total service time as may be the case for website designs with signficant user interaction, then a fixed fee scheme may be better. When congestion penalties are low, the time-based scheme is better. For instance, customers may not mind waiting for a while for some less urgent services (implying lower congestion penalty) and in such cases, a time-based scheme is better.

While the above conclusions are based solely on a revenue comparison, the firm may consider

other service performance measures too. A social planner may want a service system to serve more customers even at the expense of service value and revenues. In such scenarios, a fixed fee may be better even if it means lower revenues relative to a time-based scheme since it always serves more customers and also has lower congestion. This may be true for instance for services provided by government agencies or non-profits. Also, a website design firm may charge fixed fees even if there may be significant variation in the time required for different projects. Even if a time-based scheme may result in higher revenues, the firm may charge fixed fees because this results in more customers being served and less congestion which may be important criteria in the short run as it establishes its reputation in the marketplace. Also, a fixed fee is easier to implement since service time does not have to be tracked for each customer. Overall, both the fixed fee and time-based schemes do quite well in terms of revenue despite their simplicity and they are superior on some metrics of service performance relative to the optimal scheme.

7.2 Conclusion

The management of discretionary services poses interesting challenges because customers value additional time provided by the firm but longer service times can result in greater delays and lower productivity. Our analysis indicates that the fee structure chosen by a firm is an important lever in the management of such service systems. While the optimal pricing scheme dominates along many dimensions such as revenue, demand served and utilization, the fixed fee and time-based schemes do quite well in terms of revenue, are easier to implement and may be perceived as fairer. Also, they have some distinct advantages along certain dimensions as discussed in section 5. Our study also sheds light on when and why the fixed-fee scheme outperforms the time-based scheme in terms of revenue as well as when both schemes do well relative to the optimal pricing scheme.

There are several interesting avenues for future work and we mention a couple of them here. First, we could consider the agency issues associated with either payment scheme identified in the law and economics literature (Shepherd and Cloud (1999)): fixed fee may induce shorter service times while time-based payment may induce longer service times. Note that this is relevant only when customers have no influence on the service time. Second, we could consider a competitive scenario where two or more firms decide on the pricing scheme. In a competitive setting, the time-based scheme might be more attractive to high valuation customers and the fixed fee scheme might be more attractive to customers with low valuation customers. It is not clear which scheme will dominate or whether both schemes may coexist.

Acknowledgement

We are grateful to an anonymous AE and two referees for their comments and suggestions that have improved this paper substantially. We thank Professors Refael Hassin, Albert Ha and,

in particular, Ramandeep Randhawa for their valuable comments on our work and we are also grateful to seminar participants at Fudan University, UCLA, MIT, UT Austin, UNC Chapel Hill and the Winter Operations Conference at University of Utah for their feedback.

References

Akan, M., T. Dai, S. Tayur. 2011. Imaging room and beyond: The underlying economics behind physicians' test-ordering behavior in outpatient services. *Carnegie-Mellon University. Pittsburgh* **working paper**.

Anand, K.S., M. F. Pac, S.K. Veeraraghavan. 2011. Quality-speed conundrum: Tradeoffs in customer-intensive services. *Management Science* **57** 40–56.

Ata, B., S. Shneorson. 2006. Dynamic control of an m/m/1 service system with adjustable arrival and service rates. *Management Science* **11** 1778–1791.

Bala, R., S. Carr. 2010. Usage-based pricing of software services under competition. *Journal of Revenue and Pricing Management* **9**(3) 204–216.

Bitran, G., P. Rocha e Oliveira, A. Schilkrut. 2008. Managing customer relationships through pricing and service quality. *MIT* **working paper**.

Cachon, G.P, P. Feldman. 2011. Pricing services subject to congestion: charge per-use fees or sell subscriptions? *Manufacturing & Service Operations Management* **13**(2) 244–260.

Chase, R.B. 1981. The customer contact approach to services: Theoretical bases and practical extensions. *Operations Research* **29**(4) 698–706.

Chen, H., M. Frank. 2004. Monopoly pricing when customers queue. *IIE Transactions* **36** 569–581.

Debo, L.G., B. Toktay, L.K. Wassenhove. 2008. Queueing for expert services. *Management Science* **54**(8) 1497–1512.

Essengaier, S., S. Gupta, Z.J. Zhang. 2002. Pricing access services. *Marketing Science* **21**(2) 139–159.

Ha, Albert Y. 1998. Incentive-compatible pricing for a service facility with joint production and congestion externalities. *Management Science* **44**(12) 1623–1636.

Ha, Albert Y. 2001. Optimal pricing that coordinates queues with customer-chosen service requirements. *Management Science* **47**(7) 915–930.

Hall, J. M., P. K. Kopalle, D. F. Pyke. 2009. Static and dynamic pricing of excess capacity in a make-to-order environment. *Production and Operations Management* **18** 411–425.

Hassin, R., M. Haviv. 2003. *To queue or not to queue: Equilibrium behavior in queueing systems*. Kluwer Academic Publishers, Norwell, MA.

Hopp, W.J., S. M.R. Iravani, G. Y. Yuen. 2007. Operations system with discretionary task completion. *Management Science* **53**(1) 61–77.

Hsu, V.N., S.H. Xu, B.Jukic. 2009. Optimal scheduling and incentive compatible pricing for a service system with quality of service guarantees. *Manufacturing & Service Operations Management* **11**(3) 375–396.

Karmarkar, U.S., U.M. Apte. 2007. Operations management in the information economy: Information products, processes, and chains. *Journal of Operations Management* **25** 438–453.

Karmarkar, U.S., R. Pitbladdo. 1995. Service markets and competition. *Journal of Operations Management* **12** 397–411.

Kostami, V., S. Rajagopalan. 2013. Speed quality tradeoffs in a dynamic model. *forthcoming at Manufacturing & Service Operations Management* .

Li, L., L Jiang, L. Liu. 2012. Service and price competition when customers are naive. *Production and Operations Management* **21** 747–760.

Lovejoy, W. S., K. Sethuraman. 2000. Congestion and complexity costs in a plant with fixed resources that strives to make schedule. *Manufacturing & Service Operations Management* **2**(3) 221–239.

Lowenhahl, B. 2005. *Strategic Management of Professional Service Firms*. Copenhagaen Business School Press, Denmark.

Naor, P. 1969. On the regulation of queue size by levying tolls. *Econometrica* **38** 13–24.

Pinker, E.L., R.A. Shumsky. 2000. The efficiency-quality tradeoff of crosstrained workers. *Manufacturing & Service Operations Management* **2**(1) 32–48.

Polinsky, A. M., D. L. Rubinfeld. 2003. Aligning the interests of lawyers and clients. *American Law and Economics Review* **5**(1) 165–188.

Randhawa, R., S. Kumar. 2008. Usage restriction and subscription services: operational benefits with rational customers. *Manufacturing & Service Operations Management* **10** 429–447.

Ren, Z.J., Y.P. Zhou. 2008. Call center outsourcing: Coordinating staffing level and service quality. *Management Science* **54**(2) 369–383.

Robertson, M.A., J.A. Calloway. 2008. *Winning alternatives to the billable hour: strategies that work*. American Bar Association; 3rd Edition.

Robinson, L., R. Chen. 2011. Estimating the implied value of the customer's waiting time. *Manufacturing & Service Operations Management* **13**(Winter) 53–57.

Roels, G., U. S. Karmarkar, S. Carr. 2010. Contracting for collaborative services. *Management Science* **56**(5) 849–863.

Shepherd, G. B., M. Cloud. 1999. Time and money: Discovery leads to hourly billing. *University Of Illinois Law Reveiw* **92** 91–179.

Sundararajan, A. 2004. Nonlinear pricing of information goods. *Management Science* **50**(12) 1660–1673.

Wang, X., L. Debo, A. Scheller-Wolf, S. Smith. 2010. Design and analysis of diagnostic service centers. *Management Science* **56**(11) 1873–1890.

Welton, J. M., C. E. Dismuke. 2008. Testing an inpatient nursing intensity billing model. *Policy, Politics, & Nursing Practice* **9**(2) 103–111.

Appendix A: Proofs

Proof of Proposition 1. In the optimal pricing scheme, the firm chooses a set of service time $\tau(\alpha)$ and price $p(\alpha)$ for each customer type α to maximize the revenue, i.e.

$$\max_{\{(\tau(\cdot)\geq\tau_0, p(\cdot)\}} \lambda E(p) \Leftrightarrow \max_{\{r,\tau(\cdot)\geq\tau_0\}} r\frac{1}{1+\frac{\beta}{\mu E(v)-r}},$$

based on the transformation illustrated in Section 4.1. Note that $\mu E(v)$ is only a function of $\tau(\alpha)$ for $\alpha \in (\underline{\alpha}, \bar{\alpha})$ and the decisions on the service times and the effective rate are decoupled. Hence, we first maximize $\mu E(v)$ for a given r and then find the optimal r.

(i) Denoting $\tau_1(\alpha) = \tau(\alpha) - \tau_0$ as the service time for customers of type α after the first phase which lasts τ_0 time units, it follows that $\mu = \frac{1}{\tau_0 + E(\tau_1)}$. We have:

$$\mu E(v) = \frac{\int_{\underline{\alpha}}^{\bar{\alpha}} \sqrt{\frac{\tau_1(\alpha)}{\alpha}} f(\alpha)d\alpha}{\tau_0 + E(\tau_1)} = \frac{\int_{\underline{\alpha}}^{\bar{\alpha}} \sqrt{\frac{\tau_1(\alpha)}{\alpha}} f(\alpha)d\alpha}{\tau_0 + \int_{\underline{\alpha}}^{\bar{\alpha}} \tau_1(\alpha) f(\alpha)d\alpha}, \tag{8}$$

where $f(.)$ is the probability density function of α.

Let us fix $t := E(\tau_1) = \int_{\underline{\alpha}}^{\bar{\alpha}} \tau_1(\alpha) f(\alpha)d\alpha$, the optimal scheme has to be the one such that $\int_{\underline{\alpha}}^{\bar{\alpha}} \sqrt{\frac{\tau_1(\alpha)}{\alpha}} f(\alpha)d\alpha$ is maximized. That is,

$$\begin{aligned} \max_{\{\tau_1(\cdot)\geq 0\}} \quad & \int_{\underline{\alpha}}^{\bar{\alpha}} \sqrt{\frac{\tau_1(\alpha)}{\alpha}} f(\alpha)d\alpha \\ s.t. \quad & \int_{\underline{\alpha}}^{\bar{\alpha}} \tau_1(\alpha) f(\alpha)d\alpha = t \end{aligned}$$

The Lagrangian function associated with this optimization problem is $L := \int_{\underline{\alpha}}^{\bar{\alpha}} \sqrt{\frac{\tau_1(\alpha)}{\alpha}} f(\alpha)d\alpha + \omega(t - \int_{\underline{\alpha}}^{\bar{\alpha}} \tau_1(\alpha) f(\alpha)d\alpha) = \int_{\underline{\alpha}}^{\bar{\alpha}} l(\alpha)d\alpha$, where ω is the Lagrangian multiplier and $l(\alpha) = \sqrt{\frac{\tau_1(\alpha)}{\alpha}} f(\alpha) - \omega t f(\alpha) - \omega\tau_1(\alpha) f(\alpha)$.

We get $\tau_1^*(\alpha) = \frac{1}{4\omega^2\alpha}$ based on the first order condition and $t = \frac{1}{4\omega^2} E(\frac{1}{\alpha})$, since $\frac{\partial l(\alpha)}{\partial \tau_1(\alpha)} = f(\alpha)\frac{1}{\sqrt{\alpha}}\frac{1}{2\sqrt{\tau_1(\alpha)}} - \omega f(\alpha)$, and $\frac{\partial^2 l(\alpha)}{\partial \tau_1(\alpha)^2} = -\frac{1}{4} f(\alpha)\frac{1}{\sqrt{\alpha}}\tau_1(\alpha)^{-\frac{3}{2}} < 0$. It follows that $\tau_1^*(\alpha) = \frac{t}{\alpha E(\frac{1}{\alpha})}$.

Plugging $\tau_1(\alpha)^*$ back into equation (8), we get $\mu E(v) = \frac{\sqrt{E(\frac{1}{\alpha})}\sqrt{t}}{\tau_0 + t} = \frac{\sqrt{E(\frac{1}{\alpha})}}{\frac{\tau_0}{\sqrt{t}} + \sqrt{t}}$. So the optimal t^* that maximizes $\mu E(v)$ is the one such that $\frac{\tau_0}{\sqrt{t}} + \sqrt{t}$ is minimized. It is easy to verify that the unique solution is $t^* = \tau_0$ and $\tau_1^*(\alpha) = \frac{\tau_0}{\alpha E(\frac{1}{\alpha})}$. Using $\gamma := \frac{1}{\hat{\alpha} E(\frac{1}{\alpha})}$, we get $\tau_1^*(\alpha) = \frac{\tau_0\gamma\hat{\alpha}}{\alpha}$, $\tau^*(\alpha) = \tau_0 + \frac{\tau_0\gamma\hat{\alpha}}{\alpha}$ and $v^*(\alpha) = \frac{\sqrt{\tau_0\gamma\hat{\alpha}}}{\alpha}$. Further, $E(\tau^*) = \tau_0 + E(\tau_1) = 2\tau_0$, $\mu^* = \frac{1}{2\tau_0}$ and $E(v^*) = \sqrt{\tau_0\gamma\hat{\alpha}} E(\frac{1}{\alpha}) = \sqrt{\frac{1}{\gamma}}\sqrt{\frac{\tau_0}{\hat{\alpha}}}$.

(ii) Now we find the optimal r. Using $B = \mu^* E(v^*) = \frac{1}{\sqrt{4\tau_0\gamma\hat{\alpha}}}$, the optimal revenue is derived by solving $\max_{\{r\}} R(r) = r\frac{1}{1+\frac{\beta}{B-r}}$. Note we require the condition $r \in [0, B)$ to maintain a stable queueing system(otherwise, the traffic intensity would be more than 1). Since $\frac{\partial R(r)}{\partial r} = \frac{r^2 - 2(B+\beta)r + B(B+\beta)}{(B+\beta-r)^2}$, and $\frac{\partial^2 R(r)}{\partial^2 r} = \frac{-2\beta(B+\beta)}{(B+\beta-r)^2} < 0$, the optimal r^* satisfies the first order

condition: $r^2-2(B+\beta)r+B(B+\beta)=0$, with solutions for this quadratic equation being $r^*=B+\beta\pm\sqrt{(B+\beta)\beta}$. Since $r^*=B+\beta+\sqrt{(B+\beta)\beta}$ violates the conditions required for stability of queueing, the unique solution is then $r^*=B+\beta-\sqrt{(B+\beta)\beta}$, and the optimal revenue $R^*=[B+\beta-\sqrt{(B+\beta)\beta}]\frac{\sqrt{(B+\beta)\beta}-\beta}{\sqrt{(B+\beta)\beta}}=\beta(\sqrt{\frac{B+\beta}{\beta}}-1)^2$ after some algebra. Based on the equilibrium condition (see Section 4.1) $\lambda=\mu\frac{1}{1+\frac{\beta}{\mu E(v)-r}}$, the utilization rate in the optimal pricing scheme is $\rho^*=\frac{\lambda}{\mu}=\frac{\sqrt{(B+\beta)\beta}-\beta}{\sqrt{(B+\beta)\beta}}$. Therefore the waiting cost is $WC^*=\beta\frac{1}{(\frac{\mu^*}{\lambda^*}-1)\mu^*}=2\tau_0(\sqrt{(B+\beta)\beta}-\beta)$. As a result, $E[p^*(\alpha)]=\sqrt{\tau_0\gamma\hat{\alpha}}E(\frac{1}{\alpha})-2\tau_0(\sqrt{(B+\beta)\beta}-\beta)=\sqrt{\frac{\tau_0}{\hat{\alpha}\gamma}}-2\tau_0(\sqrt{(B+\beta)\beta}-\beta)$ by the definition of γ. Also $\tau_0=\frac{1}{4B^2\hat{\alpha}\gamma}$ by the definition of B, we can rewrite $E(p^*(\alpha))$ as $E[p^*(\alpha)]=2B\tau_0-2\tau_0(\sqrt{(B+\beta)\beta}-\beta)=2\tau_0[B+\beta-\sqrt{(B+\beta)\beta}]=\frac{1}{2\gamma\hat{\alpha}}\frac{B+\beta-\sqrt{(B+\beta)\beta}}{B^2}$. Now, $\frac{B+\beta-\sqrt{(B+\beta)\beta}}{B^2}=\frac{1}{B(1+\sqrt{\frac{\beta}{B+\beta}})}$ decreases in B, which decreases in τ_0. So $E[p^*(\alpha)]$ increases in τ_0.

(iii) Since $R^*=\beta(\sqrt{\frac{B+\beta}{\beta}}-1)^2$, R^* decreases in γ because R^* increases in B and B decreases in γ. Further, we have $\frac{\partial R^*}{\partial\beta}=2-\frac{B+2\beta}{\sqrt{B\beta+\beta^2}}<0$, and $\frac{\partial^2 R^*}{\partial\beta^2}=\frac{B^2}{2(B\beta+\beta^2)^{\frac{3}{2}}}>0$ after some algebra. Thus, R^* convexly decreases in β. □

Proof of Proposition 2. In the fixed pricing scheme, the firm chooses the committed service value v_f and the fixed fee f to maximize the revenue, i.e.

$$\max_{\{v_f,f\}}\lambda_f f\Leftrightarrow\max_{\{r_f,v_f\}}r_f\frac{1}{1+\frac{\beta}{\mu_f v_f-r_f}},$$

based on the transformation discussed in Section 4.2. As in proof of Proposition 1, we optimize v_f and r_f separately since $\mu_f v_f$ and r_f are decoupled.

(i)For any given r_f, the optimal committed value v_f^* is chosen such that $v_f\mu_f=v_f\frac{1}{\tau_0+\hat{\alpha}v_f^2}=\frac{1}{\frac{\tau_0}{v_f}+\hat{\alpha}v_f}$ is maximized, which directly leads to $v_f^*=\sqrt{\frac{\tau_0}{\hat{\alpha}}}$. Thus the service time for customer type α is $\tau_0+\tau_0\frac{\alpha}{\hat{\alpha}}$, the average service time is $E(\tau_f^*)=\tau_0+\hat{\alpha}(v_f^*)^2=2\tau_0$, and the service rate $\mu_f^*=\frac{1}{2\tau_0}$.

(ii) Using $b=v_f^*\mu_f^*=\frac{1}{2\sqrt{\hat{\alpha}\tau_0}}$, the revenue function $R_f(r_f)$ in (6) can be expressed as $R_f(r_f)=r_f\frac{b-r_f}{b-r_f+\beta}$. Note that it is required that $b>r_f$ to maintain a stable queueing system. The first order condition is:

$$\frac{\partial R_f}{\partial r_f}=\frac{(b-r_f)^2+\beta(b-2r_f)}{(b-r_f+\beta)^2},\tag{9}$$

and the second order condition is:

$$\frac{\partial^2 R_f}{\partial^2 r_f}=-\frac{2\beta(b+\beta)}{(b-r_f+\beta)^3}<0.$$

Hence, the revenue function given the optimal committed value is a concave function of r_f. The first order condition (9) gives rise to $(b-r_f)^2+\beta(b-2r_f)=0$. Solving this quadratic

equation in r_f and deleting the solution that violates the queueing stability condition, we get the optimal $r_f^* = b + \beta - \sqrt{(b+\beta)\beta}$ and the optimal fee $f^* = r_f^* \frac{1}{\mu_f} = 2\tau_0(b + \beta - \sqrt{(b+\beta)\beta})$. It is straightforward to show that f^* increases in τ_0 and decreases in β.

(iii) Plugging r_f^* into $R_f(r_f)$, the optimal revenue can be written as $R_f^* = \beta(\sqrt{\frac{b+\beta}{\beta}} - 1)^2$ after some algebra. Note that b is independent of γ so R_f^* is independent of γ. Further, R_f^* convexly decreases in β because $\frac{\partial R_f^*}{\partial \beta} = 2 - \frac{b+2\beta}{\sqrt{b\beta+\beta^2}} < 0$, and $\frac{\partial^2 R_f^*}{\partial \beta^2} = \frac{b^2}{2(b\beta+\beta^2)^{\frac{3}{2}}} > 0$ □

Proof of Proposition 3. Based on the illustration in Section 4.3, the firm chooses the rate r_t to maximize the revenue in the time-based pricing scheme, i.e.

$$\max_{\{r_t\}} \quad r_t \frac{\lambda_t}{\mu_t} = r_t \rho_t = r_t \frac{1}{1 + \frac{\beta}{\mu_t E(v_t) - r_t}},$$

In the following, we first show that the revenue function is concave in r_t, and then characterize some properties regarding the optimal r_t^*.

Denoting $k := 4\tau_0\gamma\hat{\alpha}$, we have $\mu_t = \frac{1}{\tau_0 + \frac{\tau_0}{r_t^2 k}} = \frac{k r_t^2}{\tau_0(1+kr_t^2)}$, and $E(v_t) = \frac{1}{2r_t \frac{k}{4\tau_0}} = \frac{2\tau_0}{k r_t}$ (see Section 4.3). Defining $\phi(r_t) := \mu_t E(v_t) - r_t = \frac{r_t(1-kr_t^2)}{1+kr_t^2}$, the revenue of the time-based scheme can be expressed as $R_t = r_t \frac{1}{1+\frac{\beta}{\phi(r_t)}}$. To maintain a stable queueing system, it is required that $\phi(r_t) > 0$, which is equivalent to $1 > kr_t^2$. So we have that $r_t^* < r_2$ where $r_2 = \sqrt{\frac{1}{k}}$. We next show that $\phi(r_t)$ is concave for $r_t \in (0, r_2)$: The first derivative of $\phi(r_t)$ is $\phi'(r_t) = \frac{1-4kr_t^2-k^2r_t^4}{(1+kr_t^2)^2}$, and the second derivative is $\phi''(r_t) = \frac{4kr_t(kr_t^2-3)}{(1+kr_t^2)^3}$. Thus, we have $\phi''(r_t) < 0$ for $r_t \in (0, r_2)$ because $kr_t^2 < 1 < 3$.

Recall that the revenue function is $R_t = r_t \frac{1}{1+\frac{\beta}{\phi(r_t)}}$, so the optimal rate r_t^* can only be located in the range within which $\phi(r_t)$ decreases in r_t, which implies that $\phi'(r_t^*) < 0$. That is, $1 - 4k(r_t^*)^2 - k^2(r_t^*)^4 < 0$. Solving this quadratic inequality we get $r_t^* > r_0$, where $r_0^2 = \frac{\sqrt{5}-2}{k}$.

Summarizing the above, we have that $\phi(r_t)$ concavely decreases for $r_t \in (r_0, r_2)$. As a result, the revenue function $R_t(r_t) = r_t \frac{\phi(r_t)}{\phi(r_t)+\beta}$ is concave because

$$\frac{\partial^2 R_t}{\partial r_t^2} = \frac{\beta\phi'(r_t)\phi(r_t) + r\beta\phi''(r_t)\phi(r_t) + \beta^2\phi'(r_t) + r\beta^2\phi''(r_t) - 2r\beta[\phi'(r_t)]^2}{(\phi(r_t)+\beta)^3} < 0.$$

Therefore, the unique optimal rate r_t^* satisfies the first order condition of $R_t'(r_t) = 0$, i.e.,

$$\phi(r_t^*)^2 + \beta[\phi(r_t^*) + r_t^*\phi'(r_t^*)] = 0 \Leftrightarrow r_t^*(1 - k(r_t^*)^2)^2 + 2\beta[2 - (1 + k(r_t^*)^2)^2)] = 0. \tag{10}$$

Thus, the optimal rate r_t^* satisfies $2 - (1 + k(r_t^*)^2)^2 < 0$. Solving this inequality we get that $r_t^* > r_1$, where $r_1^2 = \frac{\sqrt{2}-1}{k}$. Putting these together, we get that $r_t^* \in (r_1, r_2)$. In the rest of our proof, we restrict ourselves to the support of (r_1, r_2).

Now we show r_t^* decreases in k, which also implies that r_t^* decreases in γ as well as τ_0 because

k was defined as $k = 4\tau_0\gamma\hat{\alpha}$. Recall that both $\phi(r_t)$ and $\phi'(r_t)$ decrease in r_t, and $\phi'(r_t) < 0$ for $r_t \in (r_1, r_2)$. Also note that (10) can be re-expressed as:

$$r_t^* = \frac{\phi(r_t^*)^2 + \beta\phi(r_t^*)}{-\beta\phi'(r_t^*)}.$$

Now we prove by contradiction. Suppose r_t^* increases as k increases. Then the right hand side of above equation $\frac{\phi(r_t^*)^2+\beta\phi(r_t^*)}{-\beta\phi'(r_t^*)}$ decreases in k because the numerator decreases and the denominator increases in k (note $\phi(r_t)$ concavely decreases in $r_t \in (r_1, r_2)$). Contradiction occurs because the left hand side of above equation was assumed to be increasing in k.

The equation (10) can also be rewritten as $\beta = \frac{\phi(r_t)^2}{-(\phi(r_t)+r_t\phi'(r_t))}$. Based on the fact that $\phi(\cdot)$ concavely decreases, we know that $\phi(r_t) + r_t\phi'(r_t)$ decreases in r_t. So when β increases, r_t^* has to be decreasing.

□

Proof of Proposition 4. $r^* \geq r_f^*$ is obvious by viewing the expressions for r^* and r_f^* in Propositions 1 and 2 (note that $B \geq b$).

Now we show that $r_t^* \geq r^*$. Define $\delta = \frac{\beta}{B}$, and $c = 1 + \delta - \sqrt{(1+\delta)\delta}$, where $\delta \in (0, \infty)$. Alternatively, $\delta = \frac{(1-c)^2}{2c-1}$. So, $r^* = B + \beta - \sqrt{(B+\beta)\beta} = Bc$ (note that $c \in (\frac{1}{2}, 1)$).

To show $r_t^* \geq r^*$, it suffices to show that $\frac{\partial R_t}{\partial r_t}|_{r_t=r^*} \geq 0$, which is equivalent to having $[\phi(r_t)^2 + \beta(\phi(r_t) + r\phi'(r_t))]|_{r_t=r^*=Bc} = r_t(1 - kr_t^2)^2 + 2\beta[2 - (1 + kr_t^2)^2] \geq 0$ (see the proof of Proposition 3).

Based on the fact that 1) $\beta = \delta B$; 2) $\delta = \frac{(1-c)^2}{2c-1}$; 3)$kB^2 = 1$ (by the definitions of k and B), and plugging $r_t = r^* = Bc$ into the inequality above, we get $r_t^* \geq r^*$ if and only if:

$$\begin{aligned} Bc(1 - kB^2c^2)^2 + 2\frac{(1-c)^2}{2c-1}B[2 - (1 + kB^2c^2)^2] &= Bc(1 - c^2)^2 + 2\frac{(1-c)^2}{2c-1}B[2 - (1 + c^2)^2] \\ &= \frac{B(1-c)^2}{2c-1}(3c^3 - 4c^2 - c + 2) \geq 0 \end{aligned}$$

Denoting $L(c) := 3c^3 - 4c^2 - c + 2 \geq 0$, we know that $r_t^* \geq r^*$ if and only if $L(c) \geq 0$ for $c \in (\frac{1}{2}, 1)$. Since $L(c = 1) = 0$, and $L'(c) = 9(c + \frac{1}{9})(c - 1) \leq 0$ for $c \in (\frac{1}{2}, 1)$, we have $L(c) \geq 0$ for $c \in (\frac{1}{2}, 1)$. This concludes the proof that $r_t^* \geq r^*$. □

Proof of Proposition 5. Based on Propositions 1 and 2, the relative difference is $\frac{R^*}{R_f^*} = (\frac{\sqrt{1+\frac{B}{\beta}}-1}{\sqrt{1+\frac{b}{\beta}}-1})^2$. Denote $g(\beta) := \frac{\sqrt{1+\frac{B}{\beta}}-1}{\sqrt{1+\frac{b}{\beta}}-1}$. Since $g'(\beta) = \frac{1}{(\sqrt{\frac{b}{\beta}+1}-1)^2}\frac{1}{2\beta^2}(\frac{\sqrt{B+\beta}}{\sqrt{b+\beta}} - \frac{\sqrt{b+\beta}}{\sqrt{B+\beta}} + \frac{B\sqrt{\beta}}{\sqrt{B+\beta}} - \frac{b\sqrt{\beta}}{\sqrt{b+\beta}})$ and $B \geq b$, we have $g'(\beta) \geq 0$. So $\frac{R^*}{R_f^*}$ increases in β.

The relative revenue difference $\frac{R^*}{R_f^*}$ increasing in τ_0 for a given β can be proved in a similar way as above, details are skipped.

□

Proof of Proposition 6. The proof proceeds as follows: we first show the structure of the revenue function in the fixed as well as the time-based scheme. We then investigate how the revenue dominance changes with γ by fixing β and τ_0 , with τ_0 by fixing β and γ, with β by fixing τ_0 and γ, respectively.

Let $M(r) := \mu E(v)|_{r_t=r} = \frac{2r}{1+kr^2}$. Note that we require $M(r) > r$ to maintain a stable queueing system by viewing equation (6) and (7).

Based on Propositions 2 and 3, we have

$$\begin{aligned} R_t^* &= r_t^* \frac{1}{1+\frac{\beta}{M(r_t^*)-r_t^*}}, \\ R_f^* &= r_f^* \frac{1}{1+\frac{\beta}{b-r_f^*}}. \end{aligned}$$

For any given β and τ_0, we know from Proposition 3 that R_t^* monotonically decreases in γ while R_f^* is independent of γ. In addition, when γ, and thus k approaches zero, R_t^* obviously is greater than R_f^*. So, $R_t^* \geq (<) R_f^*$ when γ is smaller (greater) than some threshold $\bar{\gamma}(\beta, \tau_0)$, respectively.

For any given β and γ, we next show that the revenue of the time-based scheme dominates that of the fixed scheme if and only if τ_0 is less than some threshold. By viewing the expressions for the optimal revenue of time-based and fixed schemes listed above, we know that $R_t^* \geq R_f^*$ if and only if $M(r_t^*) \geq \hat{M}$, where $\hat{M}$ solves $r_t^* \frac{1}{1+\frac{\beta}{\hat{M}-r_t^*}} = r_f^* \frac{1}{1+\frac{\beta}{b-r_f^*}}$. It can be easily verified that the first derivative of $M(r_t^*)$ with respect to k is $M'(k) = \frac{(1-k(r_t^*)^2)\frac{\partial r_t^*}{\partial k} - (r_t^*)^3}{(1+k(r_t^*)^2)^2} < 0$, based on the results of Proposition 3. Since $M(r_t^*)$ decreases in k and k linearly increases in τ_0, we conclude that the revenue of the time-based scheme dominates that of the fixed scheme if τ_0 is less than some threshold $\bar{\tau_0}(\beta, \gamma)$. Otherwise, the fixed scheme dominates the time-based scheme in terms of revenue.

For any given τ_0 and γ, we show that the revenue of the time-based scheme dominates that of the fixed scheme if and only if β is less than some threshold. As $M(r_t^*)$ increases in r_t^* for any given k, and r_t^* is shown to be decreasing in β as proved in Proposition 3, analogously we have that the revenue of the time-based scheme dominates that of the fixed scheme if β is less than some threshold $\bar{\beta}(\tau_0, \gamma)$. Otherwise, the fixed scheme dominates the time-based scheme in terms of revenue. □

Proof of Proposition 7. (i) Based on Propositions 1, 2 and 3, we know that $E(v^*) = \sqrt{\frac{1}{\gamma}}\sqrt{\frac{\tau_0}{\hat{\alpha}}}$, $v_f^* = \sqrt{\frac{\tau_0}{\hat{\alpha}}}$, and $E(v_t^*) = E(\frac{1}{2r_t^*\alpha}) = \frac{1}{2r_t^*\gamma\hat{\alpha}}$, respectively. Since $r_t^* \leq \frac{1}{\sqrt{k}}$, and $k = 4\tau_0\gamma\hat{\alpha}$, we have that

$$E(v_t^*) \geq \frac{1}{2\frac{1}{2\sqrt{\tau_0\gamma\hat{\alpha}}}\gamma\hat{\alpha}} = \sqrt{\frac{\tau_0}{\gamma\hat{\alpha}}} = E(v^*) \geq v_f^*.$$

(ii) Also, $\mu_t^* = \frac{1}{\tau_0 + \frac{1}{4(r_t^*)^2\hat{\alpha}\gamma}} \leq \frac{1}{\tau_0 + \frac{k}{4\hat{\alpha}\gamma}} = \frac{1}{2\tau_0}$, we conclude that $\mu_t^* \leq \mu_f^* = \mu^*$ using Propositions 1 and 2. □

Proof of Proposition 8. We prove parts (i), (ii) and (iii) of the Proposition sequentially and the proof of part (iv) relating to utilization is included in the proof of part (ii) and the proof of part (iv) relating to demand rate is included in the proof of part (iii).

(i) From Propositions 4 and 7, we know that $r_t^* \geq r^* \geq r_f^*$, $\mu_t^* \leq \mu_f^* = \mu^*$, and $RC = r/\mu$ for the three schemes. So, we have $RC_t^* \geq RC^* \geq RC_f^*$.

(ii) We first show that $\rho^* \geq \rho_f^*$ and characterize the conditions under which $\rho_f^* \geq \rho_t^*$. We then show that these conditions are satisfied in our setting.

From Proposition 1, we have $r^* = B + \beta - \sqrt{(B+\beta)\beta}$. From (4), the utilization rate is $\rho^* = \frac{1}{1+\frac{\beta}{B-r^*}} = 1 - \sqrt{\frac{\beta}{B+\beta}}$. Analogously, $\rho_f^* = 1 - \sqrt{\frac{\beta}{b+\beta}}$. Since $B \geq b$, we have $\rho^* \geq \rho_f^*$. Also, it is obvious that both ρ^* and ρ_f^* decrease in β and τ_0 given the definitions of B and b, respectively.

Following the proof of Proposition 3, the utilization rate in the time-based scheme is $\rho_t = \frac{\phi(r_t^*)}{\phi(r_t^*)+\beta}$, where $\phi(.)$ is defined in the proof of Proposition 3 and r_t^* satisfies the first order condition

$$\phi(r_t^*)^2 + \beta[\phi(r_t^*) + r\phi'(r_t^*)] = 0 \Leftrightarrow r_t(1 - kr_t^2)^2 + 2\beta[2 - (1 + kr_t^2)^2)] = 0.$$

It follows that $\frac{\phi(r_t^*)}{\beta} = -\frac{\phi(r_t^*)+r_t^*\phi'(r_t^*)}{\phi(r_t^*)} = \frac{2[[1+k(r_t^*)^2]^2-2]}{1-k^2(r_t^*)^4}$. As shown in Proposition 3, r_t^* decreases in β, so $\frac{\phi(r_t^*)}{\beta}$ decreases in β. Therefore, we know that ρ_t decreases in β.

We now show $k(r_t^*)^2$ decreases in k. We prove by contradiction as above: suppose $k(r_t^*)^2$ increases in k, then $r_t^*[1 - (kr_t^*)^2]$ decreases in k as r_t^* has been shown to be decreasing in k. But this contradicts the relation (10): $r_t^*[1 - (kr_t^*)^2] = 2\beta[(1 + (kr_t^*)^2 - 2]$. So, $k(r_t^*)^2$ decreases in k. As a result, we have that $\frac{\phi(r_t^*)}{\beta}$ decreases in k. So ρ_t^* also decreases in τ_0 since $k = 4\gamma\tau_0\hat{\alpha}$.

Now we show $\rho_f^* \geq \rho_t^*$. From (6) and (7), we have the expressions for ρ_f^* and ρ_t^*, respectively. After some algebra, we have that:

$$\rho_f^* \geq \rho_t^* \text{ if-and-only-if } \frac{r_t^*(1 - k(r_t^*)^2)}{(1 + k(r_t^*)^2)} \leq \frac{1}{2}(\sqrt{\frac{1}{\hat{\alpha}\tau_0}} - 2r_f^*), i.e. \frac{r_t^*(1 - k(r_t^*)^2)}{(1 + k(r_t^*)^2)} \leq \sqrt{(b+\beta)\beta} - \beta. \tag{11}$$

For expositional ease, we denote $k(r_t^*)^2 = \theta$, where $\theta \in (\sqrt{2} - 1, 1)$ based on Proposition 3. So the relation (10) can be re-expressed as $\beta = \frac{r(1-\theta)^2}{2[(1+\theta)^2-2]}$. Plugging this into equation (11), we get the following after some algebra:

$$\rho_f^* \geq \rho_t^* \text{ if and only if } h(\theta) := \frac{r\sqrt{\gamma}}{2}\frac{(1-\theta)^2}{(1+\theta)^2[(1+\theta)^2-2]}[(1+\theta)^2 - \frac{4}{\sqrt{\gamma}}\theta\sqrt{\theta}] \geq 0. \tag{12}$$

Now we verify that $a(\theta) := (1+\theta)^2 - \frac{4}{\sqrt{\gamma}}\theta\sqrt{\theta}$ decreases on $(\sqrt{2} - 1, 1)$. For this purpose,

we check that its first derivative $a'(\theta) = 2(1+\theta) - \frac{6}{\sqrt{\gamma}}\sqrt{\theta} < 0$ for $\theta \in (\sqrt{2}-1, 1)$. Solving the quadratic inequality of $a'(\theta) = 2(1+\theta) - \frac{6}{\sqrt{\gamma}}\sqrt{\theta} < 0$, we get $\sqrt{\theta} \in (\frac{2}{\frac{3}{\sqrt{\gamma}}+\sqrt{\frac{9}{\gamma}-4}}, \frac{\frac{3}{\sqrt{\gamma}}+\sqrt{\frac{9}{\gamma}-4}}{2})$. Since $\frac{3}{\sqrt{\gamma}} + \sqrt{\frac{9}{\gamma}-4}$ decreases in γ, the tightest range of $\sqrt{\theta}$ would be the ones such that $\gamma = 1$, which is $(\frac{2}{3+\sqrt{5}}, \frac{3+\sqrt{5}}{2})$. It is easy to verify that $(\sqrt{\sqrt{2}-1}, 1) \subset (\frac{2}{3+\sqrt{5}}, \frac{3+\sqrt{5}}{2})$. So we have that $(1+\theta)^2 - \frac{4}{\sqrt{\gamma}}\theta\sqrt{\theta}$ decreases in $\theta \in (\sqrt{2}-1, 1)$. As a result, we get that $h(\theta)$ decreases in $\theta \in (\sqrt{2}-1, 1)$.

Since $h(\theta = 1) = 0$, we get $h(\theta) \geq 0$ for $\theta \in (\sqrt{2}-1, 1)$. Thus $\rho_f^* \geq \rho_t^*$. This concludes the proof that $\rho^* \geq \rho_f^* \geq \rho_t^*$.

(iii) Finally, we show $\lambda^* \geq \lambda_f^* \geq \lambda_t^*$.

As shown in Proposition 7, $\mu^* = \mu_f^* > \mu_t^*$. Since $\rho^* \geq \rho_f^* \geq \rho_t^*$ as proved above and $\lambda^* = \rho^*\mu^*$, we have $\lambda^* \geq \lambda_f^* \geq \lambda_t^*$. Since both ρ^* and μ^* decrease in τ_0 and β, so does λ^*. Similarly, it is easy to show that λ_f^* decreases in both β and τ_0.

As shown in Proposition 3, $E(\tau_t) = \tau_0 + \frac{1}{4\gamma\hat{\alpha}(r_t^*)^2}$ and r_t^* decreases in β, we have that the average service time $E(\tau_t)$ increases in β. Also as shown earlier, the utilization rate ρ_t^* decreases in β and τ_0. So the demand $\lambda_t^* = \rho_t^*\mu_t^*$ decreases in β as well as τ_0. □

Proof of Proposition 9. In the following, we will first show $W^* \geq W_f^*$, and then characterize the conditions under which $W^* \geq W_t^*$ and $W_f^* \geq W_t^*$.

Recall that the waiting time $W(\lambda, \mu) = \frac{\lambda}{(\mu-\lambda)\mu} = \frac{1}{(\frac{\mu}{\lambda}-1)\mu}$. Because $\mu^* = \mu_f^*$, and $\lambda^* \geq \lambda_f^*$ as shown in Propositions 7 and 8, we have that $W^* \geq W_f^*$.

From Propositions 1, 2 and 3 and using the equilibrium demands and the service rates, we get $W^* = 2\tau_0(\sqrt{(B+\beta)\beta} - \beta)$; $W_t^* = \frac{1}{4r_t^*\gamma\hat{\alpha}} - r_t^*\tau_0 = \frac{\tau_0}{k} - r_t^*\tau_0$, where $k = 4\tau_0\gamma\hat{\alpha}$, and r_t^* satisfies (10); and $W_f^* = \sqrt{\frac{\tau_0}{\hat{\alpha}}} - 2r_f^*\tau_0$.

After some algebra, we have that:

$$W_t^* \geq W_f^* \quad \text{if-and-only-if} \quad \sqrt{\frac{1}{\hat{\alpha}\tau_0}} - 2r_f^* \leq \frac{1-k(r_t*)^2}{kr_t^*}, i.e. \sqrt{(b+\beta)\beta} - \beta \leq \frac{1-k(r_t^*)^2}{2kr_t^*}; \quad (13)$$

$$W_t^* \geq W^* \quad \text{if-and-only-if} \quad \sqrt{\frac{1}{\hat{\alpha}\tau_0}} - 2r^* \leq \frac{1-k(r_t^*)^2}{kr_t^*}, i.e. \sqrt{(B+\beta)\beta} - \beta \leq \frac{1-k(r_t^*)^2}{2kr_t^*} \quad (14)$$

For expositional ease, we let $k(r_t^*)^2 = \theta$, where $\theta \in (\sqrt{2}-1, 1)$ as shown in the proof of Proposition 8. Recall that (10) can be re-expressed as $\beta = \frac{r_t^*(1-\theta)^2}{2[(1+\theta)^2-2]}$. Plugging this into equation (13) and (14), we get the following after some algebra:

$$W_t^* \geq W_f^* \quad \text{if and only if} \quad g_1(\theta) := \frac{(1-\theta)^2}{4k\theta[(1+\theta)^2-2]}(\theta^2 - 4\theta + 1 + 2\theta\sqrt{\theta\gamma}) \leq 0; \quad (15)$$

$$W_t^* \geq W^* \quad \text{if and only if} \quad g_2(\theta) := \frac{(1-\theta)^2}{4k\theta[(1+\theta)^2-2]}(\theta^2 - 4\theta + 1 + 2\theta\sqrt{\theta}) \leq 0. \quad (16)$$

Now, denote $q_1(\theta) := \theta^2 - 4\theta + 1 + 2\theta\sqrt{\theta\gamma}$, and $q_2(\theta) := \theta^2 - 4\theta + 1 + 2\theta\sqrt{\theta}$. It is easy to

check that both $q_1(\theta)$ and $q_2(\theta)$ are convex functions. Also $q_1(1) = 2\sqrt{\gamma} - 2 \leq 0$ as $\gamma \in (0, 1]$, $q_2(1) = 0$ and $q_2(\sqrt{2}-1) = 0.0479 > 0$. Plugging $\theta = \sqrt{2}-1$ into $q_1(\theta)$, it is easy to verify that when $\gamma \leq 0.828427$, $q_1(\theta) \leq 0$ for all $\theta \in (\sqrt{2}-1, 1)$.

Based on the convexity of $q_1(.)$ and $q_2(.)$, we denote $\bar{\theta_1}$ as the unique solution, if any, of $q_1(\theta) = 0$ for $\theta \in (\sqrt{2}-1, 1)$, $\bar{\theta_2}$ as the unique solution of $q_2(\theta) = 0$ for $\theta \in (\sqrt{2}-1, 1)$. We know $\bar{\theta_1} \leq \bar{\theta_2}$ because $q_2(\theta) \geq q_1(\theta)$. Now, we have the following possible cases (we do not differentiate between $>$ and $\geq$ for expositional ease in the following)

case 1): When $\gamma < 0.828427$, and $\theta \in (\sqrt{2}-1, \bar{\theta_2})$, we have $q_1(\theta) < 0$ and $q_2(\theta) > 0$, thus $W^* > W_t^* > W_f^*$;

case 2): When $\gamma < 0.828427$, and $\theta \in (\bar{\theta_2}, 1)$, we have , thus $W_t^* > W^* > W_f^*$;

case 3): When $\gamma > 0.828427$, and $\theta \in (\sqrt{2}-1, \bar{\theta_1})$, we have $q_1(\theta) > 0$ and $q_2(\theta) > 0$, thus$W^* > W_f^* > W_t^*$;

case4): When $\gamma > 0.828427$, and $\theta \in (\bar{\theta_1}, \bar{\theta_2})$, we have $q_1(\theta) < 0$ and $q_2(\theta) > 0$, thus $W^* > W_t^* > W_f^*$;

case 5): When $\gamma > 0.828427$, and $\theta \in (\bar{\theta_2}, 1)$, we have $q_1(\theta) < 0$ and $q_2(\theta) < 0$, thus $W_t^* > W^* > W_f^*$.

Summarizing the cases listed above, we have:

1)$W_t^* > W^* > W_f^*$, when $\theta \in (\bar{\theta_2}, 1)$;

2)$W^* > W_f^* > W_t^*$, when $\gamma > 0.828427$ and $\theta \in (\sqrt{2}-1, \bar{\theta_1})$;

3)$W^* > W_t^* > W_f^*$, otherwise.

As shown in Proposition 3, r_t^* decreases in β , so does $\theta = k(r_t^*)^2$. Further, θ decreases in k as also shown in Proposition 8, which implies θ decreases in τ_0. So, when $\theta \in (\bar{\theta_2}, 1)$, which represents the scenarios such that β is less than some threshold $\bar{\beta_2}$ given any τ_0, and/or τ_0 is less than some threshold $\bar{\tau_2}$ given any β, we have $W_t^* > W^* > W_f^*$. Similar rationales hold for the rest of the conditions (note $\bar{\theta_1} \leq \bar{\theta_2}$ as shown above) and thus are skipped. □

Proof of Proposition 10. Analogous to the analysis of the fixed fee scheme in Section 4.2, the equilibrium condition would be such that customers' expected net utility is zero, that is:

$$E(v_{f2}) - f2 = \beta \frac{\lambda_{f2}}{(\mu_{f2} - \lambda_{f2})\mu_{f2}}. \tag{17}$$

Defining $r_{f2} = \frac{f2}{E(\tau_{f2})}$ as before and manipulating the equilibrium condition to solve for λ_{f2}, the revenue optimization problem is:

$$\max_{\{r_{f2}, v_{f2}, \bar{\tau}\}} R_{f2} = f2 * \lambda_{f2} = r_{f2} \frac{1}{1 + \frac{\beta}{\mu_{f2} E(v_{f2}) - r_{f2}}}$$

Compared with the revenue optimization problem in Section 4.2, it is easy to observe that the optimization problem shares the same structure as the one in the fixed pricing scheme without a

maximum service time $\bar{\tau}$ imposed, except that we need to maximize $\mu_{f2}E(v_{f2})$ over two decision variables $(v_{f2}, \bar{\tau})$, rather than just maximize $\mu_f v_f$ over a single decision variable v_f as in the case of the fixed pricing scheme studied in Section 4. Specifically, the optimal committed service value and the maximum service time $(v_{f2}^*, \bar{\tau})$ are derived by solving:

$$\begin{aligned} \max_{(v_{f2},\tau_1)} \quad & \mu_{f2}E(v_{f2}) = \frac{qv_{f2} + (1-q)\sqrt{\frac{\tau_1}{\alpha_H}}}{(\tau_0 + q\alpha_L v_{f2}^2) + (1-q)\tau_1} \\ s.t. \quad & (17). \end{aligned}$$

Assuming that the parameters are such that the joint concavity conditions are satisfied (we have used extensive numerical experiments to verify it), we have the following first order conditions to characterize the optimal choice of (v_{f2}, τ_1):

$$\frac{\partial[\mu_{f2}E(v_{f2})]}{\partial v_{f2}} = 0 \qquad \Rightarrow \tau_0 + (1-q)\tau_1 - 2(1-q)\sqrt{\frac{\tau_1}{\alpha_H}}\alpha_L v_{f2} - q\alpha_L v_{f2}^2 = 0; \tag{18}$$

$$\frac{\partial[\mu_{f2}E(v_{f2})]}{\partial \tau_1} = 0 \qquad \Rightarrow \tau_0 - (1-q)\tau_1 - 2\sqrt{\alpha_H}qv_{f2}\sqrt{\tau_1} + q\alpha_L v_{f2}^2 = 0. \tag{19}$$

Adding (18) and (19) and simplifying it, we have:

$$\tau_0 - \frac{q\alpha_H + (1-q)\alpha_L}{\sqrt{\alpha_H}} v_{f2}\ \sqrt{\tau_1} = 0. \tag{20}$$

For expositional ease, we define $\alpha_2 = q\alpha_H + (1-q)\alpha_L$ (note that $\hat{\alpha} = q\alpha_L + (1-q)\alpha_H$). Plugging (20) into (18), we get the following:

$$(1-q)\alpha_2^2\tau_1^2 - \tau_0(\alpha_2^2 - 2q\alpha_H\alpha_2)\tau_1 - q\alpha_L\alpha_H\tau_0^2 = 0.$$

Solving this quadratic equation and discarding the negative root, we get $\tau_1^* = \frac{\alpha_L}{\alpha_2}\tau_0$. The optimal committed service value is $v_{f2}^* = \sqrt{\tau_0 \frac{\alpha_H}{\alpha_L\alpha_2}}$ by plugging τ_1^* back into (20). Since

$$\alpha_H\hat{\alpha} = \alpha_H[q\alpha_L + (1-q)\alpha_H] \geq \alpha_L\alpha_2 = \alpha_L[q\alpha_H + (1-q)\alpha_L],$$

we have that $v_{f2}^* = \sqrt{\tau_0 \frac{\alpha_H}{\alpha_L\alpha_2}} \geq v_f^* = \sqrt{\frac{\tau_0}{\hat{\alpha}}}$ and $\mu_{f2}^* = \frac{1}{\tau_0 + q\alpha_L \frac{\tau_0\alpha_H}{\alpha_L\alpha_2} + (1-q)\frac{\alpha_L}{\alpha_2}\tau_0} = \frac{1}{2\tau_0} = \mu_f^*$.

The average service across joining customers is $E(v_{f2}^*) = qv_{f2}^* + (1-q)\sqrt{\frac{\tau_1^*}{\alpha_H}} = \sqrt{\tau_0}[q\sqrt{\frac{\alpha_H}{\alpha_L\alpha_2}} + (1-q)\sqrt{\frac{\alpha_L}{\alpha_H\alpha_2}}] = \sqrt{\tau_0}\sqrt{\frac{\alpha_2}{\alpha_L\alpha_H}}$ after some algebra. To show $E(v_{f2}^*) \geq v_f^* = \sqrt{\frac{\tau_0}{\hat{\alpha}}}$, it suffices to show that $\sqrt{\frac{\alpha_2}{\alpha_L\alpha_H}} \geq \sqrt{\frac{1}{\hat{\alpha}}}$. Since

$$\begin{aligned} \alpha_2\hat{\alpha} \quad & = [q\alpha_H + (1-q)\alpha_L][q\alpha_L + (1-q)\alpha_H] = (1-q)q(\alpha_H^2 + \alpha_L^2) + [q^2 + (1-q)^2]\alpha_H\alpha_L \\ & \geq 2(1-q)q\alpha_H\alpha_L + [q^2 + (1-q)^2]\alpha_H\alpha_L = \alpha_H\alpha_L, \end{aligned}$$

we have that $E(v_{f2}^*) \geq v_f^* = \sqrt{\frac{\tau_0}{\hat{\alpha}}}$.

Define $b_2 = \mu_{f2}^* E(v_{f2}^*)$, it is easy to show that $b_2 = \frac{1}{2\sqrt{\tau_0}}\sqrt{\frac{\alpha_H}{\alpha_L \alpha_2}} \geq b = \frac{1}{2\sqrt{2\tau_0\hat{\alpha}}}$ (see the proof of Proposition 2). Exactly as in the proofs of Propositions 2 and 8, we can show that the utilization $\rho_{f2}^* = 1 - \sqrt{\frac{\beta}{b_2+\beta}} \geq \rho_f^* = 1 - \sqrt{\frac{\beta}{b+\beta}}$.

Since $\lambda_{f2}^* = \rho_{f2}^* \mu_f^*$, and $\mu_{f2}^* = \mu_f^*$, we conclude that $\lambda_{f2}^* \geq \lambda_f^*$.

As shown in Proposition 1, the utilization in the optimal scheme is $\rho^* = 1 - \sqrt{\frac{\beta}{B+\beta}}$, where B is the upper bound of all $\mu E(v)$, specifically $B \geq b_{f2}$, therefor $\rho^* \geq \rho_{f2}^*$. Based on the result of $\mu_{f2}^* = \mu_f^* = \mu^*$, we see $\lambda^* \geq \lambda_{f2}^*$.

Now we show the average service value in the fixed scheme with a maximum service time can render the same average service value as the optimal scheme: $E(v_{f2}^*) = E(v^*)$. As shown above, $E(v_{f2}^*) = \sqrt{\tau_0}\sqrt{\frac{\alpha_2}{\alpha_L \alpha_H}}$, the average service time in the optimal scheme is $E(v^*) = \sqrt{\frac{1}{\gamma}}\sqrt{\frac{\tau_0}{\hat{\alpha}}} = \sqrt{\tau_0}\sqrt{E(\frac{1}{\alpha})}$. Now,

$$\frac{\alpha_2}{\alpha_L \alpha_H} = \frac{q\alpha_H + (1-q)\alpha_L}{\alpha_H \alpha_L} = q\frac{1}{\alpha_L} + (1-q)\frac{1}{\alpha_H} = E(\frac{1}{\alpha}),$$

which leads to $E(v_{f2}^*) = E(v^*)$.

□

Proof of Proposition 11. After the first phase of diagnosis $(0, \tau_0)$, a customer of type α would choose an additional service time $\frac{1}{4\alpha(r_2^*)^2} = \tau_0 \frac{\gamma\hat{\alpha}}{\alpha}$, and pay an amount $\frac{1}{4\alpha r_2^*} = \frac{\sqrt{\tau_0\gamma\hat{\alpha}}}{2\alpha}$ in addition to the fixed payment of F_2^*, given the rate $r_2^* = \frac{1}{2\sqrt{\tau_0\gamma\hat{\alpha}}}$ per unit of service time. This chosen service time exactly mimics the optimal allocation of service time among heterogenous customers as shown in Proposition 1, so does the service value delivered, and service rate.

The expected payment from each customer is $F_2^* + E(\frac{\sqrt{\tau_0\gamma\hat{\alpha}}}{2\alpha}) = F_2^* + \frac{\sqrt{\tau_0}}{2\sqrt{\gamma\hat{\alpha}}} = 2\tau_0[B + \beta - \sqrt{(B+\beta)\beta}]$, which is equal to the expected payment in the optimal scheme $E[P^*(\alpha)]$ (see the proof of Proposition 1).

Since $B + \beta - \sqrt{(B+\beta)\beta} = \frac{B}{1+\sqrt{\frac{\beta}{B+\beta}}} \in (\frac{B}{2}, B)$, we get that

$$F_2^* = 2\tau_0[B + \beta - \sqrt{(B+\beta)\beta}] - \frac{\sqrt{\tau_0}}{2\sqrt{\gamma\hat{\alpha}}} \geq \tau_0 B - \frac{\sqrt{\tau_0}}{2\sqrt{\gamma\hat{\alpha}}} = 0.$$

So, in this two-part tariff scheme, customers are charged the same expected payment, obtain the same service value and have the same service time as the optimal scheme. This concludes the proof. □

Appendix B: The Approximation Approach

The analysis of various pricing schemes discussed in the paper made the assumption that the term $(1 + CV^2(\tau))$ varies very little with the pricing decision. We now show that this assumption is justifiable in many reasonable scenarios using both analytical and numerical results. Specifically, the next result provides upper and lower bounds on the potential value of $(1 + CV^2(\tau))$ in the fixed fee and time-based schemes and shows that the gap between these bounds is typically very small.

Lemma 1.

$$\text{In the fixed fee scheme: } \frac{2 + \frac{2}{\sqrt{\xi}}}{1 + \frac{1}{\xi} + \frac{2}{\sqrt{\xi}}} \leq (1 + CV_f^2) \leq \frac{3 + \xi}{4}, \text{ where } \xi := \frac{E(\alpha^2)}{\hat{\alpha}^2} \geq 1$$

$$\text{In the time-based scheme: } \frac{3 + \xi'}{4} \leq (1 + CV_t^2) \leq \frac{1 + \xi'}{2}, \text{ where } \xi' := \frac{E(\frac{1}{\alpha^2})}{[E(\frac{1}{\alpha})]^2} \geq 1$$

Proof. In the full model of the fixed-scheme, $\tau_f = \tau_0 + \alpha v_f^2$, where v_f is the service value to which SP commits. The revenue function is:

$$R_f = r_f \frac{1}{1 + \frac{\frac{\beta'}{2}(1+CV_f^2)}{\mu_f v_f - r_f}}$$

where $1 + CV_f^2 = \frac{\tau_0^2 + 2\tau_0\hat{\alpha}v_f^2 + v_f^4 E(\alpha^2)}{\tau_0^2 + 2\tau_0\hat{\alpha}v_f^2 + v_f^4\hat{\alpha}^2}$.

For any given v_f, we define $\beta = \frac{\beta'}{2}(1 + CV_f^2)$; $b = \mu_f v_f = \frac{v_f}{\tau_0 + \alpha v_f^2}$. The revenue function is re-written as:

$$R_f = r_f \frac{1}{1 + \frac{\beta}{b - r_f}}.$$

For any given v_f, thus β and b, it can be verified that the optimal $r_f^* = b + \beta - \sqrt{(b+\beta)\beta}$ to maintain the stability of the queueing system.

Plugging r_f^* into the revenue function, we have:

$$R_f = (\sqrt{(b+\beta)} - \sqrt{\beta})^2 = \Big(\frac{1}{\sqrt{\frac{1}{b} + \frac{\beta}{b^2}} + \sqrt{\frac{\beta}{b^2}}}\Big)^2.$$

Now the upper bound of v_f, denoted as $UB(v_f)$ is the one such that $UB^2(v_f) = \frac{\tau_0}{\hat{\alpha}}$. This is so because $UB(v_f)$ maximizes b, and for any $v_f \geq UB(v_f)$, $\beta(v_f) \geq \beta(UB(v_f))$ (note β monotonically increases in v_f).

Now we identify the lower bound of v_f, denoted as $LB(v_f)$. Since

$$\frac{\beta}{b^2} = \frac{\frac{\beta'}{2}E(\tilde{\tau})^2\mu_f^2}{\mu_f^2 v_f^2} = \frac{\beta'}{2}\frac{E[\tau_0^2 + 2\tau_0\hat{\alpha}v_f^2 + v_f^4\hat{\alpha}^2]}{v_f^2} = \frac{\beta'}{2}(\frac{\tau_0^2}{v_f^2} + v_f^2\hat{\alpha}^2\xi + 2\tau_0\hat{\alpha})$$

where $\xi := \frac{E(\alpha^2)}{\hat{\alpha}^2} \geq 1$

$LB(v_f)$ is the one such that $LB^2(v_f) = \frac{\tau_0}{\hat{\alpha}\sqrt{\xi}}$. This is so because: for any $v_f \leq LB(V_f)$, $\frac{\beta}{b^2}(v_f) \geq \frac{\beta}{b^2}(LB(v_f))$, and $\frac{1}{b}(v_f) \geq \frac{1}{b}(LB(v_f))$.

So, the optimal v_f^* for the full model is located between $LB(v_f), UB(v_f)$.

Plugging $LB(v_f)$ and $UB(v_f)$ into the expression of $1 + CV_f^2$, we have:

$$\frac{2 + \frac{2}{\sqrt{\xi}}}{1 + \frac{1}{\xi} + \frac{2}{\sqrt{\xi}}} \leq (1 + CV_f^2) \leq \frac{3 + \xi}{4}$$

Now we show the bounds for $(1 + CV_t^2)$ for the time-based scheme:

Following the proof of Proposition 3 shown in Appendix A, we show in the following that the range of r_t^* in the full model is (r_1, r_2), where r_1, r_2 is also defined in Proposition 3..

In the time-based scheme, the service time for customers of type α, when charged a rate r_t, is $\tau_t = \tau_0 + \frac{1}{4r_t^2\alpha}$. Now in the full model, the first order condition is changed to $\phi(r_t^*)^2 + \beta(r_t)[\phi(r_t^*) + r\phi'(r_t^*)] = r_t\phi(r_t)\beta'(r_t)$.

Since β decreases in r_t by viewing:

$$1 + CV_t^2 = \frac{E(\tau_t)^2}{[E(\tau_t)]^2} = \frac{\tau_0^2 + \frac{2\tau_0}{4r_t^2}E(\frac{1}{\alpha}) + \frac{1}{16r_t^4}E[\frac{1}{\alpha}]^2}{\tau_0^2 + \frac{2\tau_0}{4r_t^2}E(\frac{1}{\alpha}) + \frac{1}{16r_t^4}[E(\frac{1}{\alpha})]^2} \tag{21}$$

we need $\phi(r_t^*) + r\phi'(r_t^*) < 0$ at optimality, which derives the lower bound of r_t as r_1. The upper bound is still r_2 to maintain the system stability, see the proof of Proposition 3 for more details.

Plugging r_1 and r_2 into the (21), also using the fact that $\gamma = \frac{1}{\hat{\alpha}E(\frac{1}{\alpha})}$ to simplify expressions, we have that: $1 + CV_t^2 \in (l_t, u_t)$, where,

$$\begin{aligned} l_t &= \frac{1 + 2\gamma\hat{\alpha}E(\frac{1}{\alpha}) + \gamma^2\hat{\alpha}^2E[(\frac{1}{\alpha})]^2}{1 + 2\gamma\hat{\alpha}E(\frac{1}{\hat{\alpha}}) + \gamma^2\hat{\alpha}^2[E(\frac{1}{\alpha})]^2} = \frac{3 + \xi'}{4} \\ u_t &= \frac{1 + \frac{2}{\sqrt{2}-1}\gamma\hat{\alpha}E(\frac{1}{\alpha}) + \frac{1}{(\sqrt{2}-1)^2}\gamma^2\hat{\alpha}^2E[(\frac{1}{\alpha})]^2}{1 + \frac{2}{\sqrt{2}-1}\gamma\hat{\alpha}E(\frac{1}{\alpha}) + \frac{1}{(\sqrt{2}-1)^2}\gamma^2\hat{\alpha}^2[E(\frac{1}{\alpha})]^2} = \frac{1 + \xi'}{2} \end{aligned}$$

where $\xi' = \frac{E(\frac{1}{\alpha})^2}{[E(\frac{1}{\alpha})]^2} \geq 1$. □

Note that the bounds depend only on the distribution of α and are independent of any other model parameters. Taking the example of $\alpha \sim U(1,3)$, where $U(1,3)$ indicates that the value-adding service time for the slowest customers is 3 times as much as that for the quickest customers for any given service value, it can be verified that $\xi := 1.083$ and $\xi' = 1.102$. Correspondingly, $(1 + CV_f^2) \in (1.019, 1.0208)$ and $(1 + CV_t^2) \in (1.025, 1.051)$. This illustrates the negligible impact of the pricing decisions on $(1 + CV^2(\tau))$ in the fixed fee and time-based schemes. For those distributions that have less spread and variability, the accuracy will be even

greater. While the bounds in the above result are tight for the fixed fee scheme, the bounds for the time-based scheme are not. This suggests that the term $(1 + CV_t^2)$ will actually vary over an even tighter range as a function of the potential rates charged in the time-based scheme.

For the optimal scheme, we are unable to obtain reasonably tight bounds analytically due to the complex nature of the optimal prices and service times. So, we use a numerical study to show that the approximation used earlier is reasonable. This numerical study was also used to establish that the bounds established in Lemma 1 are indeed very tight. As shown in Table 3, the variability of the total system changes negligibly with the pricing schemes adopted: the deviations in $(1 + CV^2)$ across pricing schemes are within 1% for a variety of parameters. We also compared the revenues from the full and approximate model for the optimal pricing scheme, as well as for the time-based and fixed fee schemes. These results can be found in Table 3, which shows that the optimal revenues in the approximate or simplified model are very close to those in the full or original model in all three pricing schemes. In the case of the optimal pricing scheme, the revenues in the approximate model are within 0.7% of the full model and the gaps are even smaller for the other two pricing schemes. Thus, both the analytical and numerical results confirm the robustness of the approximation.

Table 3: Relative change of system variability

Minimum Service-time	Congestion Penalty	Uniform (1,2)			Triangular (1,2)		
		Fixed	Time-based	Approx	Fixed	Time-based	Approx
τ_0	β'	Y_f/Y^*	Y_t/Y^*	Y_{approx}/Y^*	Y_f/Y^*	Y_t/Y^*	Y_{approx}/Y^*
0.1	1	1.001767	1.004955	1.003081	1.0011	1.00331	1.002051
	3	1.004828	1.00776	1.00446	1.001996	1.004986	1.002955
	7	1.004259	1.009462	1.005632	1.002745	1.006875	1.003718
0.3	1	1.002402	1.006274	1.003739	1.001526	1.00414	1.002483
	3	1.003855	1.00947	1.005217	1.002481	1.006217	1.003448
	7	1.004985	1.010444	1.006379	1.003219	1.004201	1.004201
0.7	1	1.002946	1.007399	1.004290	1.001884	1.004874	1.002844
	3	1.004427	1.0109111	1.005806	1.002845	1.006881	1.003830
	7	1.005509	1.010994	1.006921	1.003567	1.009162	1.004551

Note: Y_f, Y_t, Y^* represent the system variability in the fixed scheme, time-based scheme and the optimal scheme for the *full* model, respectively. Y_{approx} represents the system variability in the *approximation* model of the optimal scheme. Specifically, $Y_f = 1 + CV_f^2$, $Y_t = 1 + CV_t^2$,$Y_{approx} = 1 + CV_{approx}^2$, and $Y^* = 1 + CV_*^2$.

Table 4: Optimality Using the Approximation Approach

Minimum Service-time	Congestion Penalty	α Uniform Distributed (0,1)			α Triangular Distributed (0,1)		
		Fixed	Time-based	Optimal Scheme	Fixed	Time-based	Optimal Scheme
τ_0	β'	$\frac{R_f^*-R_f}{R_f^*}$	$\frac{R_t^*-R_t}{R_t^*}$	$\frac{R^*-R}{R^*}$	$\frac{R_f^*-R_f}{R_f^*}$	$\frac{R_t^*-R_t}{R_t^*}$	$\frac{R^*-R}{R^*}$
0.1	1	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%
	3	0.2%	0.1%	0.2%	0.1%	0.1%	0.1%
	7	0.2%	0.2%	0.4%	0.1%	0.1%	0.3%
0.3	1	0.1%	0.1%	0.2%	0.1%	0.1%	0.1%
	3	0.2%	0.2%	0.3%	0.1%	0.1%	0.3%
	7	0.3%	0.2%	0.6%	0.2%	0.1%	0.4%
0.7	1	0.1%	0.1%	0.2%	0.09%	0.08%	0.11%
	3	0.2%	0.2%	0.4%	0.16%	0.13%	0.27%
	7	0.4%	0.3%	0.7%	0.2%	0.15%	0.6%

Note: R, R_f and R_t represent the optimal revenue of the optimal pricing scheme, the fixed pricing scheme and the time-based scheme using the approximation approach; the subscript $*$ represents the optimal revenue obtained from the full model