https://doi.org/10.1016/j.cie.2020.107096

Effects of Imperfect IoT-enabled Diagnostics on Maintenance Services: A System Design Perspective

Mingyao Sun, Feng Wu*, Chi To Ng and T.C.E. Cheng

Mingyao Sun is with the Department of Industrial Engineering, School of Management, Xi'an Jiaotong University, Xi'an, P.R. China. Email: sunmingyao@stu.xjtu.edu.cn

Feng Wu is with the Department of Industrial Engineering, School of Management, Xi'an Jiaotong University, Xi'an, P.R. China.* ***Corresponding author**. *Email: fengwu830@126.com*

Chi To Ng is with the Logistics Research Centre, Department of Logistics and Maritime Studies, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. Email: daniel.ng@polyu.edu.hk.

T.C.E. Cheng is with the Logistics Research Centre, Department of Logistics and Maritime Studies, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. Email:lgtcheng@polyu.edu.hk.

Acknowledgements

This work was supported by the National Key R&D Program of the Ministry of Science and Technology under grant numbers 2018YFB1703000 and 2018YFB1703001, and the National Natural Science Foundation under grant numbers 71871177 and 71471144. It was also supported in part by the Research Grants Council of Hong Kong under Project Number 15505320.

Effects of Imperfect IoT-based Diagnostics on Maintenance Services:

A System Design Perspective

Although many firms have deployed Internet of Things (IoT)-based diagnostics to predict equipment failures and maintenance requirements, they receive imperfect demand signals due to inadequate data quality and the use of ineffective predictive tools. A strategic queuing model was developed in this study to investigate a maintenance service provider's optimal capacity allocation and pricing decisions in the presence of imperfect IoT-based diagnostics. In addition, we considered the heterogeneous service rates resulting from the accelerating effect of IoT-based diagnostics. The results of the study revealed that, on the one hand, the optimal service rate always increases with the diagnostic quality; on the other hand, if customer error cost is higher than that of the service provider, the optimal price increases with the diagnostic quality; otherwise, it decreases with the diagnostic quality. Furthermore, we found that the error costs for the two major stakeholders—the service provider and the customer—may affect the equilibrium in different ways. Finally, in the presence of the accelerating effect, we found that although the accelerating effect can improve the average service rate, system congestion actually increases if the effect becomes increasingly obvious; however, the service provider's profit improves in this case.

Keywords: Strategic queueing; Capacity allocation; Imperfect IoT-based diagnostics; Accelerating effect; Maintenance services

1. Introduction

The cost reduction (from $22.0 in 1992 to $1.4 in 2014) of sensors, their increased speed (from 29 million Hz in 1992 to 28751 million Hz in 2014), and advances in information communication technologies and predictive tools have enabled the development of the Internet of Things (IoT). It is estimated that IoT operational applications in the areas of inventory, maintenance, managers'/workers' productivity, and process optimization will contribute trillions of dollars to the economy by 2025 (Olsen and Tomlin 2020). Given this trend, many original equipment manufacturers have adopted IoT-based diagnostics to enhance their maintenance efficiency and effectiveness, especially those operating in high-precision and capital-intensive sectors (Wang et al. 2020; Manavalan and Jayakrishna 2019).

In IoT-based maintenance service systems, numerous equipment condition indicators such as vibration, temperature, pressure, and acoustic data are continuously monitored by various equipment-mounted sensors. These systems can issue warning signals before an actual failure occurs, which thus helps to optimize equipment maintenance. However, IoT systems are vulnerable to the external environment, and many hazardous elements affect the quality of the data that such systems collect (Karkouch et al. 2016). As a result, IoT-based diagnostics may produce false warning signals. This imperfect feature of IoT-based diagnostics ultimately affects the service provider operations. Nguyen et al. (2019) emphasized that condition monitoring is often imperfect and produces errors in the real state of equipment. Moreover, diagnostic quality can be controlled by adjusting IoT-associated costs (i.e., by incurring increased costs to implement more reliable sensors or to perform more thorough data analysis). This observation has also been made in practical settings. For example, the firm ASML has employed an IoT platform to predict its customers' global maintenance requirements. The demand information produced by its IoT diagnostic system helps the company maintain its inventory of spare parts at an efficient level without jeopardizing its availability commitments. However, the company also faces imperfect prediction signals such as false positives and false negatives (Topan et al. 2018). Another example of this problem is the maintenance services for wind turbines installed worldwide (Raza and Ulansky 2019).

This imperfect monitoring considerably influences maintenance services (Topan et al. 2018; Nguyen et al. 2019); however, few studies have investigated the problem. Furthermore, the few relevant studies have focused on optimization of either the maintenance policy (Nguyen et al. 2019) or spare parts inventory (Topan et al. 2018), but they have ignored service system design problems with imperfect monitoring considerations (e.g., capacity allocation and pricing). Moreover, the relevant literature has not focused on the relationship between IoT-based diagnostics and maintenance services. We believe that the information collected by IoT systems might be useful for actual maintenance services. For example, IoT-based diagnostics can signal equipment failure before it happens so that the service provider can prepare the required resources for repair work (e.g., important spare parts and technicians) in advance. This accelerating effect of IoT-based diagnostics may thus increase the average maintenance service rate, which has rarely been studied previously.

In practice, maintenance services are often accompanied by high traffic intensity (Maddah et al. 2017). Hence, long service times enable the service provider to perform thorough equipment maintenance and improve service value; however, long service times also lead to long average equipment downtimes for customers. This "quality-speed tradeoff" in customer-intensive services has been examined in the literature on queuing theory (e.g., Anand et al. 2011; Dai et al. 2016). Nevertheless, different from common service systems, service value in our study is generated from both testing (IoT-based diagnostics) and processing (maintenance) tasks because the prediction of potential failures can prevent customers from producing nonconforming products, thereby generating a value for the customers. Analyzing the interplay between testing and processing is crucial for maintenance service managers to make optimal capacity allocation and pricing decisions in light of imperfect IoT-based diagnostics.

Motivated by industrial practice and seeking to fill this research gap, we developed a strategic queuing model to investigate the equilibrium decisions of a maintenance service provider with the consideration of imperfect IoT-based diagnostics. In maintenance services, the service price is an essential element for both the customers and service provider (Marttila et al. 2015). The customers undertake their join-or-balk decisions by weighing the service value against the waiting cost and service price, whereas the service provider maximizes its profit by making joint decisions on the service rate, service price, and diagnostic quality. We focused on the context that service value stems from both the testing and processing tasks (Levi et al. 2018), in which testing by IoT-based diagnostics uncovers the potential failures of equipment and processing by maintenance activities repairs the degraded equipment. Specifically, our study answers the following questions: What are the structural results of the service provider's equilibrium decisions? How does imperfect IoT-based diagnostics affect the equilibrium? How do the results differ from the classical quality-speed tradeoff in the queuing literature? Does the accelerating effect improve profit or ease system congestion?

The main contributions of the study are as follows. First, by jointly considering IoT-based diagnostics and the quality-speed tradeoff in maintenance services, we developed a general representation of the service value that integrates both testing and processing tasks. With this model, we provided a guidance for maintenance service providers on how they can use imperfect IoT-based diagnostics to design their service systems. Second, the accelerating effect of IoT-based diagnostics divides customers into two classes with heterogeneous service rates. To achieve the same level of service value, accurately diagnosed customers experience faster service than misdiagnosed customers do. We characterized this feature in our model and study how the equilibrium changes when the accelerating effect becomes increasingly obvious. Third, we used a real case study to strengthen the practical and implementation implications of the results; furthermore, we provided useful managerial insights to help maintenance service providers improve the efficiency and effectiveness of their operations in the presence of imperfect IoT-based diagnostics.

We organized the rest of the paper as follows. Section 2 presents a review of the relevant literature. Section 3 describes the model. Section 4 characterizes the structural properties of the optimal strategies. Section 5 extends the model to consider the accelerating effect. Section 6 presents a case study from which this paper's managerial insights are generated. Finally, Section 7 presents the conclusions. To ease exposition, proofs of all the results are presented in the Appendix.

2. Literature Review

Three streams of the literature are related to our study: studies on maintenance service operations, remote monitoring and diagnostics, and strategic queuing.

First, we reviewed the maintenance service operations literature. Kurz (2016) developed a queuing model to investigate the capacity planning decisions of an aircraft engine maintenance service provider with the goal of minimizing the sum of maintenance and delay costs. As an extension, Kurz and Pibernik (2016) studied the effects of service rate heterogeneity on the queuing system. Considering the lack of quantitative methods for after-sales service network design, Liu (2012) studied group maintenance problems for an unreliable service system and proposed a specific class of maintenance policies by modeling the system as an M/M/N queuing system. Eruguz et al. (2018) developed an integrated maintenance and spare part optimization model for a single critical component of a moving asset. Toossi et al. (2015) conducted an empirical study to investigate the value dimensions of outsourced maintenance services. Service value in our model is consistent with that in the study of Toossi et al. (2015). Eltoukhy et al. (2018) used a leader-follower Stackelberg game model to conduct joint optimization of aircraft routing and maintenance staffing. Our research differs from these studies because we considered strategic customers that make their own join-or-balk decisions. Moreover, we captured the imperfect feature of IoT diagnostics in our optimization model to study its effect on the equilibrium outcome.

Second, operational management of remote monitoring and diagnostics has attracted research attention in recent years (Guillén et al. 2016; Raza and Ulansky 2019; Raja et al. 2017). Kurz (2016) and Kurz and Pibernik (2016) analyzed the effects of advanced information created and collected by real time condition monitoring for maintenance services. Grubic (2018) used a case study to demonstrate the importance of remote monitoring technology on servitization. With the consideration of imperfect advanced demand information, Topan et al. (2018) investigated the optimal inventory policies related to lost-sales inventory systems with the option of returning inventory. Ding et al. (2018) developed a game theoretic model to investigate the effects of IoT technology on the gray market. Nguyen et al. (2019) studied the joint optimization of monitoring quality and replacement decisions in condition-based maintenance, in which the monitoring quality determines the accuracy. They proposed a dynamic maintenance and inspection policy. Our study differs from theirs in that we focused on the service system design problem (e.g., capacity allocation and service pricing) from the queuing perspective rather than on maintenance policies or spare parts inventories.

Finally, our study is also related to the literature on strategic queuing in customer-intensive services (Anand et al. 2011; Li et al. 2016; Ni et al. 2013). Anand et al. (2011) modeled the dependence of waiting time and service value on service time for both single- and multiple-service providers. They determined the equilibrium service price and service rate in each scenario. Considering resource redundancy and network externality, Kuo et al. (2017) studied the joint pricing and resource allocation decisions by proposing two types of contracts with different service level agreements. In maintenance services, Sun et al. (2020) developed a multiple service provider queuing model to investigate the optimal expert skill level and service rate decisions in a diagnostic service center design problem. Wang et al. (2019) characterized the relationship between service quality and medical consumption cost in a hospital service setting and studied how medical consumption affects the quality-speed tradeoff in an equilibrium. Dobson and Sainathan (2011) divided the services into sorting and processing and investigated how customer information analysis (in sorting) and prioritization affect the service system's stable state. Alizamir et al. (2013) studied diagnostic services with the consideration of accuracy/congestion tradeoff. Our study differs from theirs in that our concept of service value depends on both service rate and IoT diagnostic quality. We analyzed the interplay between the two service steps. In addition, we investigated the influence of service rate heterogeneity resulting from the accelerating effect of IoT-based diagnostics.

3. Methodology

3.1 Model Setup

Maintenance services include inspection, preventive repair, and replacement of defective items. We classified these tasks as testing and processing in our model. Specifically, testing by IoT-based diagnostics analyzes the equipment status and reveals its potential failure, which enables customers to eliminate the cost of using degraded equipment. Subsequently, processing by maintenance provides the necessary repair or replacement of defective parts. Both testing and processing generate value for customers, who determine whether to join the service by weighing up the service value, waiting cost, and price. We considered IoT-based diagnostics to be imperfect because it may produce false signals. When a misdiagnosis occurs, both the corresponding customer and service provider incur error costs. In particular, if the misdiagnosis is a false positive (i.e., warning without failure), the customer must bear unnecessary downtime cost and the service provider incurs extra inventory cost. If the misdiagnosis is a false negative (i.e., failure without warning), the customer bears the cost of using unreliable equipment and the service provider incurs the cost of jeopardizing its spare parts availability commitment (Topan et al. 2018). Table 1 lists the main notations used in this study.

Table 1 Notations

Notation	Description
Decision variable	

q	Diagnostic quality
μ	Average service rate
p	Service price
Parameter	
c_1	Customer error cost
c_2	Service provider error cost
γ	Rate of IoT-based diagnostics accuracy improvement to the increase in q
θ	Diagnostic accuracy
Q_0	Baseline service value of processing
μ_0	Baseline service rate of processing
α	Rate of change of the service value as a function of the service rate μ
Λ	Potential (maximum) demand rate
c_w	Customers' waiting cost per unit time
k	Cost coefficient of Diagnostic quality
b	Accelerating Factor
U	Customers' net utility
λ	Effective demand rate
Objective	
R	Revenue of the maintenance service provider

3.2 Model of Maintenance Service Value

Value of testing: The accuracy of IoT-based diagnostics depends on diagnostic quality q. In our model, the accuracy of IoT-based diagnostics is reflected in the function $\theta(q)$, which increases with q. Furthermore, the marginal accuracy from an increase in diagnostic quality diminishes as it becomes increasingly difficult and costly to improve accuracy. This result is consistent with that of Nguyen et al. (2019). Therefore, we modeled the accuracy of IoT-based diagnostics by constructing the function $\theta(q)$ as a concave increasing function of q. Specifically, we let $\theta(q) = (1 - e^{-\gamma q})$,where γ measures the rate at which the diagnostic accuracy improves when q increases. We denoted c_1 as the customer error cost if a misdiagnosis occurs. In other words, customers obtain a value of c_1 if a potential failure is accurately diagnosed; otherwise, this value is zero. Hence, customers' expected value of testing equals

$$T(q) = \theta(q)c_1. \quad (1)$$

Value of processing: Maintenance services are often accompanied by high traffic intensity (Maddah et al. 2017; Sun et al. 2020). Thus, we modeled maintenance service as a customer-intensive service, in which the service value and average waiting time both increase along with the service time (Anand et al. 2011; Li et al. 2016). If the service time is long, the service provider can perform thorough inspection and maintenance of the equipment, which results in high customer satisfaction, as indicated in Kostami and Rajagopalan (2013) and Alizamir et al. (2013). Following Anand et al. (2011), we defined the value of processing as follows:

$$Q(\mu) = Q_0 + \alpha(\mu_0 - \mu), \quad (2)$$

where Q_0 is the baseline service value when the service provider works at the baseline service rate μ_0, and α measures the sensitivity of service value to service speed (see Anand et al. 2011 for illustration). Customers may not necessarily benefit from low service rates because low service rates also lead to long average waiting times. We illustrated customers' net utility and join-or-balk decisions in the following subsection.

3.3 Customer Net Utility and Join-or-Balk Decisions

Customer net utility depends on three crucial factors: service value, waiting cost, and price. Potential customers arrive at the service system according to a Poisson distribution at a rate of Λ. An arriving customer observes diagnostic quality q, service rate μ, and price p as set by the service provider. The customer then engages in the join-or-balk decisions by weighing the service value against the waiting cost and service price. Without loss of generality, we assumed that customers obtain zero utility from balking. A unit-time waiting cost c_w is incurred for each customer in the system. We assumed that the service time follows an exponential distribution, so the service system corresponds to an *M*/*M*/*1* queue, as has been widely adopted in the service operations literature. In this study, we assumed that potential demand Λ is sufficiently large such that it avoids situations of full or zero participation (i.e., $\lambda(q,\mu,P) = \Lambda \ or \ 0$). Thus, the expected waiting time of customers is $W(\mu,\lambda) = 1/(\mu - \lambda(q,\mu,p))$. Consequently, a customer's expected net utility is

$$U(q,\mu,p) = T(q) + Q(\mu) - \frac{c_w}{(\mu - \lambda(q,\mu,p))} - p. \quad (3)$$

The induced effective demand rate can be derived by setting $U(q,\mu,p)=0$. In particular, the effective demand is equal to

$$\lambda_e(q,\mu,p)=\mu-\frac{c_w}{T(q)+Q(\mu)-p}, \tag{4}$$

where subscript *e* represents this as an equilibrium decision.

4. Service Provider Decisions

We modeled the service provider's profit function in both the short-run model, where the diagnostic quality is exogenously given, and the long-run model, where the diagnostic quality becomes a part of the service provider's strategic decisions. We began our analysis with the short-run model and then moved to the long-run model.

4.1 Short-Run Model

The service provider incurs an error cost c_2 if IoT-based diagnostics produces a false outcome, which may result in increased spare parts inventory costs (Topan et al. 2018). Thus, the service provider's expected profit from each customer served equals $(p-c_2)\big(1-\theta(q)\big)+p\theta(q)$. In addition, the cost of diagnostic quality is ignored because q is fixed in the short run problem. Therefore, the service provider's profit maximization problem is

$$\max_{(\mu,p)} R(\mu,p)=\big[(p-c_2)\big(1-\theta(q)\big)+p\theta(q)\big]\lambda_e(\mu,p), \tag{5}$$

in which $\lambda_e(\mu,p)$ is the effective arrival rate illustrated in Equation (4). The variable q is ignored because it is fixed in this model.

In the following analysis, we assumed $T(q)+Q(\mu)-e^{-\gamma q}\geq c_w/\mu$ to rule out the trivial case that a customer cannot expect a non-negative utility at the non-profit price ($(p-c_2)\big(1-\theta(q)\big)+p\theta(q)=0$), even when no other customer precedes him in the queue. We solved the profit maximization problem in two steps. First, we found the optimal price $p^*(\mu)$ for a given service rate, μ. Then, using $p^*(\mu)$, we found the optimal service rate in the operating region. Our analysis yields the following results.

Proposition 1: In the short-run problem with an exogenous diagnostic quality level, a unique market equilibrium exists in which *(i) the optimal service rate is* $\mu^* = \frac{c_1-(c_1+c_2)e^{-\gamma q}+Q_0+\alpha\mu_0}{2\alpha}$ *and the optimal price is* $p^* = \frac{c_1+Q_0+\alpha\mu_0+(c_2-c_1)e^{-\gamma q}}{2} - \sqrt{\alpha c_w}$*;*

(ii) the expected service value that customers perceive in equilibrium is $T(q)+Q(\mu^*) = \frac{c_1+Q_0+\alpha\mu_0+(c_2-c_1)e^{-\gamma q}}{2}$*; and*

(iii) the induced arrival rate is $\lambda^* = \frac{c_1-(c_1+c_2)e^{-\gamma q}+Q_0+\alpha\mu_0}{2\alpha} - \sqrt{c_w/\alpha}$*, and the expected waiting time is* $W^* = \sqrt{\alpha/c_w}$.

Proposition 1 indicates that the optimal service price and service rate for exogenous diagnostic quality q. Although the customer's perceived service value decreases with the service rate, the expected waiting time increases with it. When the service rate is low (i.e., $\mu < \mu^*$), the increase in customer waiting cost dominates the decrease in the perceived service value that results from an increase in the service rate. Hence, the service provider can extract more surplus from the customer and earn more profit by improving the service rate. However, when the service rate is high (i.e., $\mu > \mu^*$), the perceived service value dominates. As a result, customers are less willing to purchase the service than they are when the service rate is low, so the service provider's profit decreases when the service rate increases. Regarding service price p, when $p < p^*$, the revenue from the increased price grows more proportionately than the decline in the demand rate does. Therefore, the service provider's profit increases with price until it reaches p^*. Once $p > p^*$, the customer's net utility is low, leading to a low effective demand rate and profit for the service provider.

From Proposition 1, we can readily obtain the following:

Corollary 1: Under the short-run model with exogenous diagnostic quality, the following results are obtained.

(i) When customer error cost c_1 *increases, the service provider improves the service rate* μ^* *and raises the service price* p^**, and the induced service value and effective demand rate both increase in equilibrium.*

(ii) When service provider error cost c_2 *increases, the service provider decreases the service rate* μ^* *but increases the service price* p^**, and the induced service value increases but the effective demand rate decreases in equilibrium.*

To visually illustrate the results in Corollary 1, we developed a numerical example in Figure 1. The selected parameters are $Q_0 = 1,\ q = 0.5,\ \gamma = 2,\ c_w = 1,\ \mu_0 = 2$, and $\alpha = 1$. In addition, in Figure 1(a), we let $c_2 = 3$; and in Figure 1(b), we let $c_1 = 1$. Corollary 1 indicates that the customer error cost c_1 and service provider error cost c_2 have different effects on equilibrium decisions. In particular, as c_1 increases, the customer's perceived service value increases [see the red line in Figure 1(a)], and thus the service provider increases the service rate [see the black line in Figure 1(a)] and raises the service price [see the bottle green line in Figure 1(a)] to extract additional customer surplus. In addition, an increased number of customers will be admitted to the system due to the increase in the service rate [see the orange line in Figure 1(a)]. However, as c_2 increases, the service provider will improve the service price [see the bottle green line in Figure 1(b)] to guarantee its profit. In order to keep the service attractable to customers, the service provider has to decrease the service rate [see the black line in Figure 1(b)] to enhance the service value. As a result, the effective demand decreases since the service rate declines [see the orange line in Figure 1(b)]. Moreover, according to Proposition 1(iii), the waiting time does not depend on c_1 or c_2 in equilibrium [see the blue lines in Figures 1(a) and 1(b)].

Please insert Figure 1 here.

Figure 1 Effects of c_1 (a) and c_2 (b) on service provider's equilibrium decisions.

Next, we investigated how exogenous diagnostic quality q affects the service provider's equilibrium decisions and profit. The results are derived in Corollary 2.

Corollary 2: Under the short-run model, as diagnostic quality q increases:

(i) both the optimal service rate μ^* and the effective demand rate λ^* increase in q;

(ii) if $c_1 > c_2$, both the service price p^* and the induced service value $T(q) + Q(\mu^*)$ increase in q; however, if $c_1 < c_2$, the result is the opposite; and

(iii) the service provider's profit increases in q.

To deepen our understanding and illustrate the results intuitively, we presented a numerical example in Figure 2. The parameters are selected as $Q_0 = 1,\ \gamma = 2,\ c_w = 1,\ \mu_0 = 2$, and $\alpha = 1$. In addition, we let $c_1 = 2$ and $c_2 = 1$ in Figure 2(a), whereas we let $c_1 = 1$ and $c_2 = 2$ in Figure 2(b). Corollary 2(i) shows that adding diagnostic quality may lead to a deterioration in the service value of processing $Q(\mu^*)$ because of the increased service rate [see the black lines in Figures 2(a) and 2(b)]. Consequently, additional customers may be attracted to the service [see the orange lines in Figures 2(a) and 2(b)].

The finding that the optimal service price does not necessarily increase along with diagnostic quality may appear to be counterintuitive. The intuition behind this result is that customers' perceived value depends on both the testing and processing tasks. In particular, when customers are primarily concerned with diagnostic quality (i.e., $c_1 > c_2$, Figure 2(a)), we found that as q increases, the increase in the service value of *testing* (i.e., $T(q)$) dominates the decline in the service value of *processing* (i.e., $Q(\mu^*)$). Therefore, customers' comprehensive service value is improved (i.e., $T(q) + Q(\mu^*)$). The service provider can increase the service price [see the bottle green line in Figure 2(a)] to extract all the customer surplus without reducing the effective demand rate [see the orange line in Figure 2(a)]. However, when service provider is primarily concerned with diagnostic quality (i.e., $c_1 < c_2$, Figure 2(b)), the increased service value of testing $T(q)$ cannot offset the loss in the service value of processing $Q(\mu^*)$. The service provider therefore reduces the service price [see the bottle green line in Figure 2(b)] to increase the attractiveness of the service to customers.

Corollary 2 also implies that for the short-run problem, although the service price may be reduced when $c_1 < c_2$, the service provider's revenue always rises as q increases [see the blue lines in Figures 2(a) and 2(b)]. This is because both the expected profit from each customer $(p - c_2)\big(1 - \theta(q)\big) + p\theta(q)$ and the equilibrium demand rate λ^* increase in q. However, this conclusion is not valid in the long-run problem when q is endogenous.

Please insert Figure 2 here.

Figure 2 Effects of q on service provider's equilibrium decisions.

Anand et al. (2011) studied the speed-quality tradeoff in a queuing system and showed that the equilibrium service rate decreases as the customer becomes more sensitive to the service time (i.e., α increases). That is, when service value $Q(\mu)$ depends heavily on the service time, the service provider always increases the service time to increase the service value. However, we showed the result may not be true in our model, which considered imperfect IoT-based diagnostics. The details are presented in Corollary 3.

Corollary 3: (i) The equilibrium service rate μ^* *decreases in the customers' sensitivity to service time* α *if* $Q_0 > (c_1 + c_2)e^{-\gamma q} - c_1$ *and increases otherwise. Moreover, the sensitivity of* μ^* *to the change of* α *increases when* c_1 *decreases or* c_2 *increases (*$\frac{\partial^2 \mu^*}{\partial\alpha\partial c_1} < 0$*;* $\frac{\partial^2 \mu^*}{\partial\alpha\partial c_2} > 0$*).*

(ii) If $Q_0 > (c_1 + c_2)e^{-\gamma q} - c_1$*, the relationship between* λ^* *and* α *is U-shaped. Specifically, effective demand rate* λ^* *decreases in* α *for* $\alpha < \frac{4[Q_0-(c_1+c_2)e^{-\gamma q}+c_1]^2}{{\mu_0}^2}$ *and increases in* α *for* $\alpha > \frac{4[Q_0-(c_1+c_2)e^{-\gamma q}+c_1]^2}{{\mu_0}^2}$*; however, the effective demand rate* λ^* *always increases in* α *if* $Q_0 < (c_1 + c_2)e^{-\gamma q} - c_1$.

(iii) The relationship between equilibrium service price p^* *and* α *is U-shaped. Specifically, when* $\alpha < \frac{c_w}{{\mu_0}^2}$*, service price* p^* *decreases in* α*; otherwise,* p^* *increases in* α*.*

To deepen our understanding and illustrate our results more intuitively, we presented a numerical example in Figure 3. Anand et al. (2011) demonstrated that the service rate always decreases as α increases in equilibrium. However, the results in Corollary 3(i) assert that the equilibrium service rate may decrease or increase in α when imperfect IoT-based diagnostics is being accounted for. That is, the service provider may have an incentive to increase the service rate even as customers become more concerned of the service time. The underlying intuition is as follows. Note that the benchmark service value Q_0 is generated irrespective of the service rate, along with a revenue loss $(c_1 + c_2)e^{-\gamma q} - c_1$ incurred by imperfect nature of IoT-based diagnostics. Therefore, $Q_0 - (c_1 + c_2)e^{-\gamma q} + c_1$ represents either an exogenous surplus or loss that the service provider has to consider when the service is performed. In particular, when the value is positive, as α increases, the service provider must decrease the service rate to keep the service attractable (see the upper left panel in Figure 3). However, when the value is negative (i.e., $Q_0 < (c_1 + c_2)e^{-\gamma q} - c_1$), it enables flexibility for the service provider to increase the service rate (see the lower left panel in Figure 3). The left panels of Figure 3 also show that μ^* becomes more sensitive to the change of α when c_1 decreases or c_2 increases ($\frac{\partial^2 \mu^*}{\partial\alpha\partial c_1} < 0$; $\frac{\partial^2 \mu^*}{\partial\alpha\partial c_2} > 0$), i.e., c_1 changes from 1(solid)→0.5(dashed), c_2 changes from 5(solid)→7(dotted).

Corollary 3(ii) shows that the U-shaped relationship between effective demand and α remains robust in our setting if $Q_0 > (c_1 + c_2)e^{-\gamma q} - c_1$. In this case, for low levels of α, the service value improvement $\alpha(\mu_0 - \mu)$ is low, and thus the waiting cost is more sensitive to the change of α than the service value. Hence, congestion dominates in this scenario and the effective demand decreases because of the decline in the customer's net utility. For high levels of α, the increase in service value can compensate for the increase in waiting cost, and thus the effective demand increases as α increases (see the upper middle panel in Figure 3). However, this U-shaped relationship vanishes if $Q_0 < (c_1 + c_2)e^{-\gamma q} - c_1$. Now, the effective demand always increases in α (see the lower middle panel in Figure 3).

Please insert Figure 3 here.

Figure 3 Effects of α on service provider's equilibrium decisions

4.2 Long-Run Model

Under the long-run model, diagnostic quality q becomes a decision variable for the service provider. In addition, imperfect IoT-based diagnostics not only incurs an error cost for the service provider, but also gives rise to an additional investment cost kq^2 under the long-run model, where k is the cost coefficient of diagnostic quality. Therefore, the service provider's objective function becomes

$$\max_{(q,\mu,p)} R(q,\mu,p) = \big[(p - c_2)\big(1 - \theta(q)\big) + p\theta(q)\big]\lambda_e(q,\mu,p) - kq^2. \tag{6}$$

In the long-run problem, the service provider has to make joint decisions among the diagnostic quality q, average service rate μ, and service price p. We obtained the following results by solving the optimization problem in a three-step scheme (see the online appendix).

Proposition 2: In the long-run problem with an endogenous diagnostic quality level, a unique market equilibrium exists in which

(i) the optimal diagnostic quality q^* for the service provider is obtained by solving the following equation

$$\left[c_1 - (c_1 + c_2)e^{-\gamma q} + Q_0 + \alpha\mu_0 - 2\sqrt{c_w\alpha}\right]\frac{\gamma(c_1+c_2)}{2\alpha}e^{-\gamma q} = 2kq;$$

(ii) the optimal service rate is $\mu^* = \frac{c_1-(c_1+c_2)e^{-\gamma q^*}+Q_0+\alpha\mu_0}{2\alpha}$ *and the optimal service price is* $p^* = \frac{c_1+Q_0+\alpha\mu_0+(c_2-c_1)e^{-\gamma q^*}}{2} - \sqrt{\alpha c_w}$*; and*

(iii) the equilibrium demand at the optimal diagnostic quality, service rate, and price is $\lambda^* = \frac{c_1-(c_1+c_2)e^{-\gamma q^*}+Q_0+\alpha\mu_0}{2\alpha} - \sqrt{c_w/\alpha}$.

Proposition 2 shows that the finding from the short-run problem in Corollary 2(iii) is not valid in the long-run problem when the diagnostic quality becomes a part of the service provider's strategic decision. In the short-run problem, our result indicates that the service provider can always benefit from a high value of the diagnostic quality q. However, in the long-run problem, Proposition 2 reveals that profit function R is unimodal with respect to q. This is because the benefit created by diagnostic accuracy diminishes as q increases; however, the investment cost of diagnostic quality gradually increases. Consequently, an equilibrium interior point q^* that maximizes the service provider's profit exists, which means that a high value of the diagnostic quality q is not necessarily beneficial to the service provider in the long-run problem. Regarding the optimal service rate μ^* and service price p^*, the illustrations are similar to those given in Proposition 1.

The following corollary about the effects of k on the service provider's optimal decisions are obtained immediately from Proposition 2.

Corollary 4: Under the long-run model, as the cost coefficient of diagnostic quality k increases, the following results are obtained:

(i) the relationship between the cost coefficient k and the diagnostic quality q^* is non-monotonic. Specifically, when $\frac{4\alpha k}{[\gamma(c_1+c_2)]^2} < \frac{e^{-2\gamma q}}{1+\gamma q}$, the optimal diagnostic quality q^* increases in k; otherwise, q^* decreases in k;

(ii) the corresponding optimal service rate μ^* increases in k when $\frac{4\alpha k}{[\gamma(c_1+c_2)]^2} < \frac{e^{-2\gamma q}}{1+\gamma q}$; otherwise, the corresponding optimal service rate μ^* decreases in k; and

(iii) if $c_1 > c_2$, the corresponding optimal service price p^* increases in k when $\frac{4\alpha k}{[\gamma(c_1+c_2)]^2} < \frac{e^{-2\gamma q}}{1+\gamma q}$; otherwise, it decreases in k. However, if $c_1 < c_2$, the result is the opposite.

Corollary 4 indicates that the cost coefficient of diagnostic quality k affects the service provider's strategy profoundly. Corollary 4(i) may initially seem surprising because it reveals that the service provider may increase the diagnostic quality q^* even as k increases. The underlying intuition is as follows. When q is relatively low, the revenue obtained from decreased error cost is higher than the investment cost. Hence, increasing q is beneficial for the service provider. However, when q is relatively high, the benefit of improving q is limited but the related cost is high. Thus, it is beneficial for the service provider to decrease the diagnostic quality when k increases.

5. Influence of the Accelerating Effect

The model developed in the previous section only considers IoT-based diagnostics as a process to identify equipment status. However, the information collected by IoT-based diagnostics may also be beneficial for subsequent maintenance. For example, the forecasted service requirements (e.g., spare parts and technicians) revealed by IoT-based diagnostics can help providers to prepare sufficient resources for the service in advance. Thus, the service rate can be increased if maintenance requirements can be accurately predicted.

5.1 Market Equilibrium and Expected Waiting Time

Given imperfect IoT-based diagnostics, customers are divided into two classes with heterogeneous service rates. These classifications are indexed by $i = A, B$, where A comprises customers whose potential failures are accurately diagnosed and B comprises customers whose potential failures are misdiagnosed. Upon engaging the service, class A customers experience faster service than class B customers do, because the service process for class A customers can be well prepared before equipment failures occur. The service provider attempts to achieve a common level of service value $Q(\mu)$ across both types of customers. Hence, the workload size for class B customers follows the exponential distribution with mean $\frac{1}{\mu}$; and the workload size for class A customers follows the exponential distribution with mean $\frac{1}{b\mu}$, where $b > 1$ is a measure of the useful information collected by IoT-based diagnostics. We illustrated this process pictorially in Figure 4.

Please insert Figure 4 here

Figure 4 Illustration of the service process with the accelerating effect

Let p_i be the probability that a customer is in class i. In this study, we assumed that the service provider is "fair", meaning that customer discrimination by adopting different admission policies or service prices does not occur. Therefore, we can obtain $p_A = \theta(q)$ and $p_B = 1 - \theta(q)$. Customers in class i arrive from the Poisson process with rate $\lambda_i = p_i \cdot \lambda$, where λ is the total arrival rate and is a decision variable for the service provider. The service provider determines the service price p, the basic service rate μ, the scheduling policy ϕ, and the admission arrival rate λ to maximize its profit. Because customers do not know their status *ex ante*, both classes of customers have the same joining probability and expected waiting time. Let $W^{\Phi}(\lambda, \mu_A, \mu_B)$ denote the expected waiting time. The following market-clearing condition is satisfied in equilibrium:

$$T(q) + Q(\mu) - c_w W^{\Phi}(\lambda, \mu_A, \mu_B) - p = 0. \tag{7}$$

We first examined the expected waiting time $W^{\Phi}(\lambda, \mu_A, \mu_B)$. We assumed that the first-come-first-served policy applies within the same customer class but that the service provider can determine the service sequence of the two customer classes. For this type of service system, Dai et al. (2016) discovered that the shortest expected processing time policy maximizes the service provider's expected total profit. Thus, for customers in the queue, class A customers have strict priority (non-preemptive) over class B customers (see Dai et al. 2016 for the proof). Thus, given the accelerating factor b, the resulting waiting times of classes A and B are respectively given by

$$W_A(\lambda, \mu_A, \mu_B) = \frac{\lambda_A/{\mu_A}^2 + \lambda_B/{\mu_B}^2}{(1-\lambda_A/\mu_A)(1-\lambda_B/\mu_B)} + \frac{1}{\mu_A}; \tag{8}$$

$$W_B(\lambda, \mu_A, \mu_B) = \frac{\lambda_A/{\mu_A}^2 + \lambda_B/{\mu_B}^2}{(1-\lambda_A/\mu_A)(1-\lambda_B/\mu_B-\lambda_A/\mu_A)} + \frac{1}{\mu_B}. \tag{9}$$

The proof of these results was presented by Dai et al. (2016), so we need not to provide it in our study. Equations (8) and (9) represent the true waiting time of customer classes A and B, respectively; however, when making the join-or-balk decision, a customer does not know his status *ex ante* and make the decision based on the expected waiting time given by $W(\lambda, \mu_A, \mu_B) = p_A W_A + p_B W_B$. Integrating $\lambda_i = p_i \cdot \lambda$, $\mu_A = b\mu$, and $\mu_B = \mu$ into this formulation, we obtained the following:

$$W(\lambda, \mu) = \left[\frac{\frac{p_A}{(b\mu)^2} + \frac{p_B}{\mu^2}}{\frac{1}{\lambda} - \frac{p_A}{b\mu}}\right]\left[p_A + \frac{p_B}{1-\lambda\left(\frac{p_A}{b\mu} + \frac{p_B}{\mu}\right)}\right] + \frac{p_A}{b\mu} + \frac{p_B}{\mu}. \tag{10}$$

where p_A and p_B are illustrated previously.

On the basis of the formulation of $W(\lambda, \mu)$, we derived the following properties of $W(\lambda, \mu)$, which are crucial to the subsequent analysis.

Proposition 3: The expected waiting time increases with the effective demand and decreases with the service rate, (i.e., $(\partial/\partial\lambda)W(\lambda,\mu) \geq 0$ *and* $(\partial/\partial\mu)W(\lambda,\mu) \leq 0$*). Moreover,* $W(\lambda,\mu)$ *becomes more sensitive to changes in* λ *and* μ *when* λ *and* μ *are high, (i.e.,* $(\partial^2/\partial\lambda^2)W(\lambda,\mu) > 0$ *and* $(\partial^2/\partial\mu^2)W(\lambda,\mu) > 0$).

Intuitively, the expected waiting time of the queue increases with customer total arrival rate λ and decreases with basic service rate μ. In addition, Proposition 3 indicates that as total arrival rate λ increases, the expected waiting time increases convexly (i.e., $(\partial^2 / \partial\lambda^2)W(\lambda,\mu) > 0$). This is because an increase of class B customers leads to an increase in system congestion. This result also reveals that when basic service rate μ is relatively low, the expected waiting time decreases rapidly in μ; while as μ increases, the declining rate of the expected waiting time decreases.

5.2 Effects of the Accelerating Factor

We explored how accelerating factor b affects service provider's strategy in the short-run problem. Analytical results for the long-run problem are not available due to the complexity of the expressions, so we used numerical studies to reveal the results in the next section. Proposition 3 indicates that the expected waiting time strictly increases with total arrival rate λ, so the market-clearing condition uniquely defines arrival rate $\lambda(\mu)$ for each level of service price p. That is, for a given service price p, a unique demand rate $\lambda(\mu)$ exists that can be obtained from Equation (7). Substituting the market-clearing condition (Equation (7)) for p, we found that the service provider's objective function in the short-run problem can be stated as choosing λ and μ to maximize:

$$R(\lambda,\mu) = \lambda\big[T(q) + Q(\mu) - c_w W(\lambda,\mu) - c_2\big(1-\theta(q)\big)\big]. \tag{11}$$

The following proposition summarizes the effect of accelerating factor b on the service provider's decisions in the short-run problem.

Proposition 4: When service rate differentiation is allowed under the accelerating effect of IoT-based diagnostics, as the value of accelerating factor b increases, the following results are obtained:

(i) the service provider admits an increased number of customers to the service system, (i.e., λ increases in b), but operates slower (i.e., μ decreases in b);

(ii) expected waiting time W increases in b; and

(iii) equilibrium price p decreases in b.

Proposition 4 indicates that the service provider tends to provide a high-quality-low-price service to customers in the presence of the accelerating effect. Proposition 4(i) shows that when the accelerating effect becomes increasingly obvious, (i.e., b increases), the service provider will reduce the service rate and admit more customers into the service system. This is rather intuitive: as b increases, the service provider's capacity is relaxed. To increase customer's willingness to engage the service, the service provider will reduce the service rate (i.e., μ decreases) to improve the service quality, and customers are hence more willing to join the service (i.e., λ increases) than they would be otherwise.

Proposition 4(ii) indicates that customers encounter an increasingly congested service system as accelerating factor b increases. This is rather counterintuitive because one might think that the service system congestion could be eased if the accelerating effect becomes increasingly obvious. An explanation is as follows: although the average service rate improves as the accelerating effect increases, effective demand λ increases and basic service rate μ decreases in equilibrium, in accordance with Proposition 4(i). The reduced service rate and increased effective demand result in increased system congestion.

Proposition 4(iii) may initially seem surprising in that although the service quality increases, the service provider must charge a lower price as b increases. The underlying intuition is as follows: from the market-clearing condition, we can know that the service price is determined by the relative value of the service quality and expected waiting cost. Although the service quality increases in b, the induced net service value declines because of the corresponding increase in waiting cost. Therefore, the service provider must lower the price so that customers' individual benefits are balanced between the waiting cost and service price in equilibrium.

5.3 Numerical Studies

In this subsection, we conducted numerical studies to generate practical insights about Proposition 4. Analytical results for the long-run problem are not available because of the complexity of the revenue function, which involves three variables. Thus, the numerical studies also provide insights into the effect of accelerating factor b on the equilibrium in the long-run problem.

In the numerical studies, we used the following parameters: $c_1 = 3$, $c_2 = 1$, $\gamma = 1$, $c_w = 1$, $k = 2$, $\mu_0 = 1$, $Q_0 = 2$, and $\alpha = 1$. We took the case where the value of b is 1 as a benchmark, and we then varied b to examine the effects of the accelerating factor on the equilibrium outcomes. Also, we maintained diagnostic quality q as a constant 2 in the short-run problem, and treated q as a decision variable in the long-run problem. Figures 5 and 6 illustrate the numerical results of the short-run and long-run problems, respectively.

Please insert Figure 5 here

Figure 5 Effects of b on the equilibrium decisions in the short-run problem

Please insert Figure 6 here

Figure 6 Effects of b on equilibrium decisions in the long-run problem

We obtained the following observations from the results of the numerical studies:

(i) The effect of accelerating factor b has the following in common in both short- and long-run problems: (1) the service provider admits additional customers in equilibrium [see Figures 5(c) and 6(d)]; (2) the optimal service rate decreases [see Figures 5(a) and 6(a)]; (3) the optimal price decreases [see Figures 5(b) and 6(b)]; and (4) the total expected waiting time increases [see Figures 5(d) and 6(e)]. These results indicate that in both the short- and long-run problems, accelerating factor b has similar effects on the effective demand, service rate, expected waiting time, and price. In addition, the numerical results also reveal that although the optimal price decreases in b, the service provider's net revenue increases in both settings [see Figures 5(e) and 6(f)]. (ii) In addition to the above results, b also affects the equilibrium diagnostic quality in the long-run problem. In particular, we observed that the optimal diagnostic quality q concavely decreases in b [see Figure 6(c)]. That is, as the accelerating effect becomes increasingly obvious, (i.e., b increases), the service provider decreases the diagnostic quality at a declining rate.

6. Case Study and Managerial Implications

6.1 Case Study of ASML

In this section, we presented a case study on ASML, a globally prominent original equipment manufacturer that produces lithography systems, which are essential for integrated circuit production in the semiconductor industry. The data for this case study are from Topan et al. (2018); they sourced it from ASML. We considered four representative failure modes (denoted by P, T, X, and W) to reflect the different characteristics of the maintenance services that ASML provides to its customers. ASML has deployed an IoT platform to continuously monitor numerous equipment condition indicators such as vibration, temperature, pressure, and acoustic signals. Hence, this system can predict potential equipment failures, which increases maintenance efficiency and effectiveness. However, ASML also faces imperfect demand signals, which induce error costs for both ASML and its customers. The imperfect demand signals consist of false positives (i.e., warning without failure) and false negatives (i.e., failure without warning).

For example, if the demand signal is a false positive, the corresponding customer incurs unnecessary downtime cost and ASML incurs the extra cost of spare parts inventory; if the demand signal is a false negative, the customer incurs the cost of producing nonconforming chips and ASML bears the cost of jeopardizing its spare parts availability commitment. In this paper, we denoted the error costs of the customers and ASML as c_1 and c_2, respectively. The average lead time of the spare parts is L. We summarized the data on the four failure modes in Table 2. The time unit is the number of weeks and the cost unit is measured in euros (€). Incidentally, we took $\mu_0 = \frac{1}{L}$ as the baseline service rate, as in previous models. Moreover, we set $Q_0 = 5{,}000$, $\alpha = 10^4$, and $k = 5000$.

Using the analytical results in Sections 3 and 4, we derived the optimal decisions on the service price and service rate for both the short- and long-run problems based on the real diagnostic quality level q provided by ASML. Furthermore, we analyzed the optimal decision on diagnostic quality q^* in the long-run problem and compare it with the q values provided by ASML to explore whether the company has deployed IoT-based diagnostics appropriately. Table 3 presents the results.

Table 2. ASML case data

Failures	c_1(€)	c_2(€)	$L(week)$	q	c_w(€)
P	5500	2720	2	0.44	7500
T	325	112	2	0.90	7500
X	3400	152	2	0.43	7500
W	1400	646	2	0.50	7500

Table 3. Optimal decisions for ASML in the short-run and long-run problems

	Short-run Problem				Long-run Problem			
Failures	q	λ^*	p^*	μ^*	q^*	λ^*	p^*	μ^*
P	0.44	0.1709	7416	0.3209	0.74	0.2301	6908	0.3028
T	0.90	0.1781	5008	0.3281	0.82	0.2086	4476	0.2914
X	0.43	0.1690	5426	0.3190	0.79	0.1845	5038	0.2876
W	0.50	0.1565	2878	0.3065	0.86	0.1802	2496	0.2568

The results in Table 3 demonstrate that ASML has not achieved the optimal diagnostic accuracy.

By comparing the q value provided by ASML with the optimal q^*, we revealed that the diagnostic quality levels of the failure types P, X, and W are all lower than the corresponding optimal values. This means that ASML should use more reliable sensors or perform more thorough data analytics than it currently does to improve the diagnostic quality of these three failure types. However, the diagnostic accuracy of failure type T provided by ASML is higher than the optimal value, indicating that nearly all the T type failures can be diagnosed. In summary, the company should pay more attention to the failure types P, X, and W than it does to the failure type T. Furthermore, in this paper, we presented the optimal service price and service rate decisions for the company in both the short- and long-run problems.

6.2 Managerial Implications

This section presents the implications of our results for maintenance service decision-makers.

First, it is necessary for service managers to take account of the misdiagnosis costs (error costs) seriously when making capacity allocation and pricing decisions. Customer and service provider error costs affect the equilibrium decisions in different ways (as illustrated by Corollary 1). Service managers can often increase the service price if the diagnostic quality increases; however, we discovered that doing so may not be optimal because the effect of diagnostic quality on the service price depends on whose error cost dominates (Corollary 2).

Second, it is not always optimal for service managers to decrease the service rate even when customers become more sensitive to the service time (Corollary 3). In common service systems that do not involve imperfect diagnostics, the service rate always decreases as customers become more sensitive to the service time. However, when the imperfect diagnosis of IoT systems in maintenance service systems is considered, the service rate decreases with customer sensitivity to the service time only when the service value provided to customers is sufficiently high.

Third, when service managers want to ascertain a suitable diagnostic quality level, they should seek to strike a balance between revenue and investment cost to maximize profit rather than improve the diagnostic quality persistently (Proposition 2).

Finally, service managers must expand the buffer zone if the service rate can be improved with IoT-based diagnostics because the system becomes increasingly congested as the importance of the accelerating effect increases. Service managers should also fully use the information collected by IoT systems to improve the average service rate and thereby increase profit (Proposition 3 and the numerical studies).

7. Conclusions

This study investigates the effects of imperfect IoT-based diagnostics on equipment maintenance services. We examined both the short-run problem, in which the diagnostic quality is fixed, and the long-run problem, in which diagnostic quality is a strategic decision variable. In addition, we explored the influence of the accelerating effect on service provider decisions. The following assertions were detected.

First, Corollary 2(iii) and Proposition 2 reveal that in the short-run problem, the service provider always benefits from high diagnostic quality. However, in the long-run problem, the service provider's profit is unimodal in the diagnostic quality. Second, Corollary 3 indicates that the results in the classic quality-speed tradeoff literature (Anand et al. 2011) may not hold when imperfect IoT-based diagnostics is taken account in our setting. We found that the equilibrium service rate may increase even as customers become increasingly sensitive to the service time. Third, Corollaries 2(i) and 2(ii) reveal that when customers are more concerned with the diagnostic accuracy than the service provider is, the equilibrium price and service value increase with diagnostic quality; otherwise, the equilibrium price and service value decrease. Moreover, Corollary 1 also shows that the error costs for the two major stakeholders—the service provider and the customer—may affect the equilibrium in different ways. Finally, conventional wisdom would suggest that an increase in the average service rate is beneficial because it should ease system congestion; however, in our setting, Proposition 4 suggests that the accelerating effect increases congestion because the service provider improves the service quality and attracts an increased number of customers into the system as the accelerating effect becomes increasingly crucial. In this case, the service provider offers a high-quality-low-price service to its customers, and thereby increases its profit.

There are two main limitations of our study. First, our study assumes the customers are homogenous. However, considerable heterogeneity exists among customers in the real world. Future research can focus on how heterogeneity affects service provider's decisions in equilibrium. Second, we analyzed the imperfect nature of IoT-based diagnostics as accurate and inaccurate. This method helps us obtain the analytical results in our study. However, in practice, imperfect conditions can be divided into more than two categories (e.g., failure without warning, warning without failure, and the uncertainty of the exact time of failure). Future researchers can consider examining these additional elements into their models.

References

Ali-Marttila, M., Tynninen, L., Marttonen-Arola, S., Kärri, T. (2015). Value elements of industrial maintenance: Verifying the views of the customer and the service provider. *International Journal of Strategic Engineering Asset Management,* 2(2), 136-158.

Alizamir, S., De Véricourt, F., Sun, P. (2013). Diagnostic accuracy under congestion. *Management Science,* 59(1), 157-171.

Anand, KS, Paç, MF, Veeraraghavan, S. (2011). Quality-speed conundrum: Trade-offs in customer-intensive services. *Management Science,* 57(1), 40-56.

Dai, T., Akan, M., Tayur, S. (2016). Imaging room and beyond: The underlying economics behind physicians' test-ordering behavior in outpatient services. *Manufacturing & Service Operations Management,* 19(1), 99-113.

Ding, L., Hu, B., Ke, C., Wang, T., Chang, S. (2019). Effects of IoT technology on gray market: An analysis based on traceability system design. *Computers & Industrial Engineering,* 136, 80-94.

Dobson, G., Sainathan, A. (2011). On the impact of analyzing customer information and prioritizing in a service system. *Decision Support Systems,* 51(4), 875-883.

Eltoukhy, A. E. E., Wang, Z.X., Chan, F, Chuang, S.H. (2018). Joint optimization using a leader–follower Stackelberg game for coordinated configuration of stochastic operational aircraft maintenance routing and maintenance staffing. *Computers & Industrial Engineering,* 125, 46-68.

Eruguz, A S., Tan, T., van Houtum, G J. (2018). Integrated maintenance and spare part optimization for moving assets. *IISE Transactions*, 50(3), 230-245.

Grubic, T. (2018). Remote monitoring technology and servitization: Exploring the relationship. *Computers in Industry* 100, 148-158.

Guillén, A. J., Crespo, A., Gómez, J. F., Sanz, M. D. (2016). A framework for effective management of condition-based maintenance programs in the context of industrial development of E-Maintenance strategies. *Computers in Industry*, 82, 170-185.

Karkouch A, Mousannif H, Al Moatassime H, et al. (2016). Data quality in internet of things: A state-of-the-art survey. *Journal of Network and Computer Applications*, 73, 57-81.

Kostami, V., Rajagopalan, S. (2013). Speed–quality trade-offs in a dynamic model. *Manufacturing & Service Operations Management,* 16(1), 104-118.

Kurz, J., Pibernik, R. (2016). Flexible capacity management with future information. *Working Paper.*

Kurz, J. (2016). Capacity planning for a maintenance service provider with advanced information. *European Journal of Operational Research,* 251(2), 466-477.

Kuo, CW., Huang, KL., Yang, CL. (2017). Optimal contract design for cloud computing service with resource service guarantee. *Journal of the Operational Research Society,* 68, 1030–1044.

Levi, R., Magnanti, T., Shaposhnik, Y. (2019). Scheduling with testing. *Management Science,* 65(2), 776-793.

Li, X., Guo, P., Lian, Z. (2016). Quality-speed competition in customer-intensive services with boundedly rational customers. *Production and Operations Management,* 25(11), 1885-1901.

Liu, GS. (2012). Three m-failure group maintenance models for M/M/N unreliable queuing service systems. *Computers & Industrial Engineering,* 62 (4), 1011-1024.

Maddah, B., W. W. Nasr., and A. Charanek. (2017). A Multi-station System for Reducing Congestion in High-variability Queues. *European Journal of Operational Research*, 262 (2), 602–619.

Manavalan, E., Jayakrishna K. (2019). A review of Internet of Things (IoT) embedded sustainable supply chain for industry 4.0 requirements. *Computers & Industrial Engineering,* 127, 925-953.

Ni, G., Xu, Y., Dong, Y. (2013). Price and speed decisions in customer-intensive services with two classes of customers. *European Journal of Operational Research,* 228(2), 427-436.

Nguyen, K.T.P., Do, P., Huynh, K.T., Bérenguer, C., Grall, A. (2019). Joint optimization of monitoring quality and replacement decisions in condition-based maintenance. *Reliability Engineering and System Safety* 189, 177-195.

Olsen, T.L., Tomlin, B. (2020). Industry 4.0: Opportunities and Challenges for Operations Management. *Manufacturing & Service Operations Management,* 22(1),113-122.

Raja, J.Z., Frandsen, T. and Mouritsen, J. (2017). Exploring the managerial dilemmas encountered by advanced analytical equipment providers in developing service-led growth strategies. *International Journal of Production Economics*, 192, 120-132.

Raza A, Ulansky V. (2019). Optimal preventive maintenance of wind turbine components with imperfect continuous condition monitoring. *Energies*, 12(19), 1-24.

Sun, M., Wu, F., Zhao, S. (2020). Machine diagnostic service center design under imperfect diagnosis with uncertain error cost consideration. *International Journal of Production Research,* 58(10), 3015-3035.

Toossi, A., Lockett, H. L., Raja, J. Z., Martinez, V. (2013). Assessing the value dimensions of outsourced maintenance services. *Journal of Quality in Maintenance Engineering,* 19(4), 348-363.

Topan, E., Tan, T., Van Houtum, GJ., Dekker, R. (2018). Using imperfect advance demand information in Lost-

Sales inventory systems with the option of returning inventory. *IISE Transactions*, 50(3), 246-264.

Wang, X., Wu, Q., Lai, G., Scheller-Wolf, A. (2019). Offering Discretionary Healthcare Services with Medical Consumption. *Production and Operations Management*, 28 (9), 2291-2304.

Wang, Y., Liu, Y., Chen, J., Li, X. (2020). Reliability and condition-based maintenance modeling for systems operating under performance-based contracting. *Computers & Industrial Engineering,* 42, Article 106344.

Effects of Imperfect IoT-enabled Diagnostics on Maintenance Services: A System Design Perspective

Online Appendix

Proof of Proposition 1

We solve the service provider's profit maximization problem in two steps. First, we find the optimal price $p^*(\mu)$ for a given service rate μ. Then we solve the optimal service rate μ^* by plugging $p^*(\mu)$ into the objective function.

First, we show that $R(\mu,p)$ is concave in the price p. Taking the first and second derivatives of R with respect to p, we have

$$\frac{\partial R}{\partial p}=\mu-\frac{c_w}{T(q)+Q(\mu)-p}-(p-c_2e^{-\gamma q})\frac{c_w}{(T(q)+Q(\mu)-p)^2},$$

$$\frac{\partial^2 R}{\partial p^2}=-\frac{2c_w}{(T(q)+Q(\mu)-p)^2}-\frac{2c_w(p-c_2e^{-\gamma q})}{(T(q)+Q(\mu)-p)^3}<0.$$

So $R(\mu,p)$ is concave in p. By solving the first-order condition, we find the optimal service price p for any value of the service rate μ: $p^*(\mu)=T(q)+Q(\mu)-\sqrt{c_w(T(q)+Q(\mu)-c_2e^{-\gamma q})/\mu}$. As a result, the service provider's objective function becomes

$$R\big(\mu,p^*(\mu)\big)=\left[\mu-\sqrt{\frac{c_w\mu}{g(\mu)}}\right]\left[g(\mu)-\sqrt{\frac{c_w}{\mu}g(\mu)}\right],$$

where $g(\mu)=T(q)+Q(\mu)-c_2e^{-\gamma q}$. Taking the first derivative of R with respect to μ, we have

$$\frac{\partial R}{\partial \mu}=g(\mu)+\mu\frac{\partial\big(g(\mu)\big)}{d\mu}-\frac{1}{2}\cdot 2\sqrt{c_w}[\mu g(\mu)]^{-\frac{1}{2}}\left(\mu+\frac{\partial\big(g(\mu)\big)}{d\mu}\right).$$

As a result,

$$\frac{\partial R}{\partial \mu}=(g(\mu)-\alpha\mu)\left[1-\sqrt{\frac{c_w}{\mu g(\mu)}}\right].$$

When $1-\sqrt{\frac{c_w}{\mu g(\mu)}}\leq 0$, the equilibrium demand is equal to $\lambda\leq 0$. Thus, we only consider the effective case where $1-\sqrt{\frac{c_w}{\mu g(\mu)}}>0$, which indicates that the sign of $\frac{\partial R}{\partial \mu}$ is the same as that of $g(\mu)-\alpha\mu$. Because the sign of $g(\mu)-\alpha\mu$ is first positive and then negative in μ, $R\big(\mu,p(\mu)\big)$ is unimodal in μ. Solving the first-order condition, we find the optimal service rate as

$$\mu^*=\frac{c_1-(c_1+c_2)e^{-\gamma q}+Q_0+\alpha\mu_0}{2\alpha},$$

which, in turn, yields

$$p^* = \frac{c_1-(c_1+c_2)e^{-\gamma q}+Q_0+\alpha\mu_0}{2} - \sqrt{\alpha c_w},$$

$$T(q)+Q(\mu^*) = \frac{c_1+Q_0+\alpha\mu_0+(c_2-c_1)e^{-\gamma q}}{2},$$

$$\lambda^* = \frac{c_1-(c_1+c_2)e^{-\gamma q}+Q_0+\alpha\mu_0}{2\alpha} - \sqrt{c_w/\alpha}\ .$$

Proof of Corollary 1

Taking the derivatives of μ^*, p^*, $T(q)+Q(\mu^*)$, and λ^* with respect to c_1 and c_2, respectively, we have

$$\frac{\partial\mu^*}{\partial c_1} = \frac{1-e^{-\gamma q}}{2\alpha} > 0;\ \frac{\partial p^*}{\partial c_1} = \frac{1-e^{-\gamma q}}{2} > 0,$$

$$\frac{\partial[T(q)+Q(\mu^*)]}{\partial c_1} = \frac{1-e^{-\gamma q}}{2} > 0;\ \frac{\partial\lambda^*}{\partial c_1} = \frac{1-e^{-\gamma q}}{2\alpha} > 0.$$

$$\frac{\partial\mu^*}{\partial c_2} = \frac{-e^{-\gamma q}}{2\alpha} < 0; \frac{\partial p^*}{\partial c_2} = \frac{e^{-\gamma q}}{2} > 0,$$

$$\frac{\partial[T(q)+Q(\mu^*)]}{\partial c_2} = \frac{e^{-\gamma q}}{2\alpha} > 0; \frac{\partial\lambda^e}{\partial c_2} = \frac{-e^{-\gamma q}}{2\alpha} < 0.$$

Therefore, we derive the conclusions in Corollary 1.

Proof of Corollary 2

(i) Taking the derivatives of μ^* and λ^* with respect to q, respectively, we have

$$\frac{\partial\mu^*}{\partial q} = \frac{\gamma(c_1+c_2)}{2\alpha}e^{-\gamma q} > 0 \text{ and } \frac{\partial\lambda^*}{\partial q} = \frac{\gamma(c_1+c_2)}{2\alpha}e^{-\gamma q} > 0.$$

Therefore, we obtain the conclusion that both the optimal service rate μ^* and the equilibrium demand rate λ^* increase in q.

(ii) Taking the derivatives of p^* and $T(q)+Q(\mu^*)$ with respect to q, respectively, we have

$$\frac{\partial p^*}{\partial q} = \frac{\gamma(c_1-c_2)}{2}e^{-\gamma q} \text{ and } \frac{\partial[T(q)+Q(\mu^*)]}{\partial q} = \frac{\gamma(c_1-c_2)}{2}e^{-\gamma q}.$$

Therefore, if $c_1 > c_2$, $\frac{\partial p^*}{\partial q} > 0$ and $\frac{\partial[T(q)+Q(\mu^*)]}{\partial q} > 0$, indicating that both the optimal service price and induced service value increase in the diagnostic quality q. However, if $c_1 < c_2$, the result is the opposite.

(iii) The induced service provider's revenue function in the short-run problem is

$$R(\mu,p) = (p - c_2 e^{-\gamma q})\lambda^*(\mu,p).$$

The expected profit from each customer is $p - c_2 e^{-\gamma q} = \frac{c_1+Q_0+\alpha\mu_0-(c_2+c_1)e^{-\gamma q}}{2} - \sqrt{\alpha c_w}$, which evidently increases in q. In addition, we have confirmed in Corollary 1(i) that the equilibrium demand λ^* increases in q. Since both $p - c_2e^{-\gamma q}$ and λ^* both are positive and increase in q, we confirm that the profit increases in q.

Proof of Corollary 3

(i) Taking the first order derivative of μ^* with respect to α, we have

$$\frac{\partial \mu^*}{\partial \alpha} = \frac{-[Q_0 - (c_1 + c_2)e^{-\gamma q} + c_1]}{4\alpha^2}.$$

Thus, it is obvious that when $Q_0 - (c_1 + c_2)e^{-\gamma q} + c_1 > 0$, $\frac{\partial \mu^*}{\partial \alpha} < 0$; and when $Q_0 - (c_1 + c_2)e^{-\gamma q} + c_1 < 0$, $\frac{\partial \mu^*}{\partial \alpha} > 0$. In addition, with the results in Corollary 1, we can easily obtain $\frac{\partial^2 \mu^*}{\partial \alpha \partial c_1} < 0$; $\frac{\partial^2 \mu^*}{\partial \alpha \partial c_2} > 0$.

(ii) Similarly, we proof

$$\frac{\partial \lambda^*}{\partial \alpha} = \frac{-[Q_0 - (c_1 + c_2)e^{-\gamma q} + c_1] + \sqrt{\alpha c_w}}{2\alpha^2}.$$

Thus, when $[Q_0 - (c_1 + c_2)e^{-\gamma q} + c_1] < 0$, it is easy to know $\frac{\partial \lambda^*}{\partial \alpha} > 0$. When $[Q_0 - (c_1 + c_2)e^{-\gamma q} + c_1] > 0$, there exists a threshold $\hat{\alpha} = \frac{4[Q_0 - (c_1 + c_2)e^{-\gamma q} + c_1]^2}{{\mu_0}^2}$ such that when $\alpha < \hat{\alpha}$, $\frac{\partial \lambda^*}{\partial \alpha} < 0$, and $\frac{\partial \lambda^*}{\partial \alpha} > 0$ otherwise. The value $\hat{\alpha}$ can be obtained by setting $-[Q_0 - (c_1 + c_2)e^{-\gamma q} + c_1] + \sqrt{\alpha c_w} = 0$.

(iii) The result can be obtained with the same method above.

Proof of Proposition 2

To obtain the results in Proposition 2, we proceed with the solution method in three steps:

Step 1: Find the optimal price $p^*(\mu, q)$ for a given service rate μ and diagnostic quality q.

The service provider's objective function in the long-run problem is

$$\max_{(q,\mu,p)} R(q, \mu, p) = (p - c_2e^{-\gamma q})\left(\mu - \frac{c_w}{(1 - e^{-\gamma q})c_1 + Q_0 + \alpha(\mu_0 - \mu) - p}\right) - kq^2.$$

Taking the first and second order derivatives of R with respect to p, we have

$$\frac{\partial R}{\partial p} = \mu - \frac{c_w}{(1-e^{-\gamma q})c_1 + Q_0 + \alpha(\mu_0 - \mu) - p}$$

$$- [P - c_2 e^{-\gamma q}] \frac{c_w}{((1-e^{-\gamma q})c_1 + Q_0 + \alpha(\mu_0 - \mu) - p)^2},$$

$$\frac{\partial^2 R}{\partial p^2} = -\frac{2c_w}{((1-e^{-\gamma q})c_1 + Q_0 + \alpha(\mu_0 - \mu) - p)^2} - \frac{2c_w[p - c_2 e^{-\gamma q}]}{((1-e^{-\gamma q})c_1 + Q_0 + \alpha(\mu_0 - \mu) - p)^3} < 0.$$

As a result, for given μ and q, the revenue function is concave in p. Solving the first-order condition, we find the optimal price and the corresponding equilibrium demand

$$p^*(\mu, q) = (1-e^{-\gamma q})c_1 + Q_0 + \alpha(\mu_0 - \mu) - \sqrt{\frac{c_w((1-e^{-\gamma q})c_1 + Q_0 + \alpha(\mu_0 - \mu) - c_2 e^{-\gamma q})}{\mu}},$$

$$\lambda^e(\mu, q) = \mu - \sqrt{\frac{c_w \mu}{c_w\ [(1-e^{-\gamma q})c_1 + Q_0 + \alpha(\mu_0 - \mu) - c_2 e^{-\gamma q}]}}.$$

Step 2: Using $p^*(\mu, q)$ and $\lambda^e(\mu, q)$, we find the optimal service rate $\mu^*(q)$ for a given q.

Plugging $p^*(\mu, q)$ and $\lambda^e(\mu, q)$ into the objective function, the service provider's revenue function is

$$R(q, \mu) = \max_{(q,\mu)} \left[T(q) + Q(\mu) - \sqrt{\frac{c_w(T(q) + Q(\mu) - c_2 e^{-\gamma q})}{\mu}} - c_2 e^{-\gamma q} \right] \left[\mu - \sqrt{\frac{c_w \mu}{T(q) + Q(\mu) - c_2 e^{-\gamma q}}} \right] - kq^2.$$

Taking the first derivative of *R* with respect to μ, we have

$$\frac{\partial R}{\partial \mu} = (T(q) + Q(\mu) - c_2 e^{-\gamma q} - \alpha\mu) \left[1 - \sqrt{\frac{c_w}{\mu[T(q) + Q(\mu) - c_2 e^{-\gamma q}]}} \right].$$

Similar to Proposition 1, we find the optimal service rate by solving $T(q) + Q(\mu) - c_2 e^{-\gamma q} - \alpha\mu = 0$. As a result,

$$\mu^*(q) = \frac{c_1 - (c_1 + c_2)e^{-\gamma q} + Q_0 + \alpha\mu_0}{2\alpha}.$$

Step 3: Using $p^*(\mu^*(q), q)$ and $\mu^*(q)$, we analyze the optimal diagnostic quality q.

Note that $T(q) + Q(\mu) - c_2 e^{-\gamma q} = \frac{c_1 - (c_1 + c_2)e^{-\gamma q} + Q_0 + \alpha\mu_0}{2} = \alpha\mu^*(q)$, so the service provider's objective function becomes

$$R(q) = \left[\alpha\mu^*(q) - \sqrt{\frac{c_w \cdot \alpha\mu^*(q)}{\mu^*(q)}} \right] \left[\mu^*(q) - \sqrt{\frac{c_w \mu^*(q)}{\alpha\mu^*(q)}} \right] - kq^2.$$

As a result,

$$R(q) = \alpha[\mu^*(q)]^2 - 2\sqrt{c_w\alpha}\mu^*(q) + c_w - kq^2.$$

Taking the first and second derivatives of R with respect to q, we have

$$\frac{\partial R}{\partial q} = 2\alpha\mu^*(q)\frac{\partial \mu^*(q)}{\partial q} - 2\sqrt{c_w\alpha}\frac{\partial \mu^*(q)}{\partial q} - 2kq,$$

$$\begin{aligned}\frac{\partial^2 R}{\partial q^2} &= \frac{(c_1+c_2)}{2\alpha}\gamma^2 e^{-\gamma q}\left[2(c_1+c_2)e^{-\gamma q} - \left(c_1 + Q_0 + \alpha\mu_0 - 2\sqrt{\alpha c_w}\right)\right] - 2k \\ &= \frac{(c_1+c_2)}{2\alpha}\gamma^2 e^{-\gamma q}\cdot(-2)\left[\frac{c_1 + Q_0 + \alpha\mu_0 - (c_1+c_2)e^{-\gamma q}}{2} - \sqrt{\alpha c_w}\right] - 2k.\end{aligned}$$

As we only consider the case $\lambda^e \neq 0$ and $\frac{c_1+Q_0+\alpha\mu_0-(c_1+c_2)e^{-\gamma q}}{2} - \sqrt{\alpha c_w} > 0$, we can obtain the revenue function R is concave in q as $\frac{\partial^2 R}{\partial q^2} < 0$.

In addition,

$$\begin{aligned}\lim_{q\to 0}\frac{\partial R}{\partial q} &= \lim_{q\to 0}\frac{\gamma(c_1+c_2)}{2\alpha}\left[\left(c_1+Q_0+\alpha\mu_0-2\sqrt{\alpha c_w}\right)e^{-\gamma q} - (c_1+c_2)e^{-2\gamma q}\right] - 2kq \\ &= \frac{\gamma(c_1+c_2)}{2\alpha}\left[\left(c_1+Q_0+\alpha\mu_0-2\sqrt{\alpha c_w}\right) - (c_1+c_2)\right] > 0\ ,\end{aligned}$$

$$\lim_{q\to\infty}\frac{\partial R}{\partial q} = \lim_{q\to\infty}\frac{\gamma(c_1+c_2)}{2\alpha}\left[\left(c_1+Q_0+\alpha\mu_0-2\sqrt{\alpha c_w}\right)e^{-\gamma q} - (c_1+c_2)e^{-\gamma q}\right] - 2kq < 0.$$

Thus, there exists a unique solution q^* that maximizes the service provider's revenue, which can be solved by the first order condition of R:

$$\left[c_1 - (c_1+c_2)e^{-\gamma q} + Q_0 + \alpha\mu_0 - 2\sqrt{c_w\alpha}\right]\frac{\gamma(c_1+c_2)}{2\alpha}e^{-\gamma q} = 2kq.$$

With the expression of $p^*(\mu, q)$ and $\mu^*(q)$, we obtain the following results

The optimal service rate is $\mu^* = \frac{c_1-(c_1+c_2)e^{-\gamma q^*}+Q_0+\alpha\mu_0}{2\alpha}$ and the optimal price is $p^* = \frac{c_1+Q_0+\alpha\mu_0+(c_2-c_1)e^{-\gamma q^*}}{2} - \sqrt{\alpha c_w}$.

Proof of Corollary 4

In order to examine how the optimal strategies $q^*, \mu^*(q^*)$, and $p\left(q^*, \mu^*(q^*)\right)$ vary with *k*, we first find the derivative of q^* with respect to k by utilizing the implicit function theory, and then focus on $\mu^*(q^*)$ and $p\left(q^*, \mu^*(q^*)\right)$, respectively.

(i) Taking the derivative of q with respect to k by utilizing the implicit function theory, we have

$$\alpha(c_1+c_2)e^{-\gamma q^*}\frac{\partial q^*}{\partial k} = \frac{4\alpha}{\gamma(c_1+c_2)}\left[qe^{\gamma q^*} + k\left(e^{\gamma q^*} + \gamma q^* e^{\gamma q^*}\right)\frac{\partial q^*}{\partial k}\right].$$

As a result,

$$\frac{\partial q^*}{\partial k} = \frac{\frac{4\alpha}{\gamma(c_1+c_2)} q^*}{\gamma(c_1+c_2)e^{-2\gamma q^*} - k(1+\gamma q^*)\frac{4\alpha}{\gamma(c_1+c_2)}}.$$

Therefore, the sign of $\frac{\partial q^*}{\partial k}$ is the same as that of $\gamma(c_1+c_2)e^{-2\gamma q^*} - k(1+\gamma q^*)\frac{4\alpha}{\gamma(c_1+c_2)}$. Dealing with the above problem, we obtain the result: When $\frac{4\alpha k}{[\gamma(c_1+c_2)]^2} < \frac{e^{-2\gamma q^*}}{1+\gamma q^*}$, $\frac{\partial q^*}{\partial k} > 0$, so the optimal diagnostic quality q^* increases in k; otherwise, $\frac{\partial q^*}{\partial k} < 0$ and the optimal diagnostic quality q^* decreases in k.

(ii) $\frac{\partial \mu^*}{\partial k} = \frac{\partial \mu^*}{\partial q}\frac{\partial q^*}{\partial k}$. Because $\frac{\partial \mu^*}{\partial q} = \frac{\gamma(c_1+c_2)}{2\alpha}e^{-\gamma q} > 0$, we can know that the sign of $\frac{\partial \mu^*}{\partial q}$ is the same as that of $\frac{\partial q^*}{\partial k}$, so k has the same impacts on q^* and μ^*.

(iii) $\frac{\partial p^*}{\partial k} = \frac{\partial p^*}{\partial q}\frac{\partial q^*}{\partial k}$. Because $\frac{\partial p^*}{\partial q} = \frac{\gamma(c_1-c_2)}{2}e^{-\gamma q}$ and $\frac{\partial q^*}{\partial k} > 0$, we find that the sign of $\frac{\partial p^*}{\partial k}$ depends on the value of $c_1 - c_2$. If $c_1 > c_2$, $\frac{\partial p^*}{\partial q} > 0$, so when $\frac{4\alpha k}{[\gamma(c_1+c_2)]^2} < \frac{e^{-2\gamma q^*}}{1+\gamma q^*}$, the optimal price p^* increases in k; otherwise, the optimal price p^* decreases in k. However, if $c_1 < c_2$, the corresponding result is the opposite.

Proof of Proposition 3

The formulation of the waiting time is

$$W(\lambda,\mu) = \left[\frac{\frac{p_A}{(b\mu)^2} + \frac{p_B}{\mu^2}}{\frac{1}{\lambda} - \frac{p_A}{b\mu}}\right]\left[p_A + \frac{p_B}{1-\lambda\left(\frac{p_A}{b\mu} + \frac{p_B}{\mu}\right)}\right] + \frac{p_A}{b\mu} + \frac{p_B}{\mu}$$

and we define

$$\varphi_1(\lambda,\mu) = \frac{\frac{p_A}{(b\mu)^2} + \frac{p_B}{\mu^2}}{\frac{1}{\lambda} - \frac{p_A}{b\mu}} \quad \text{and} \quad \varphi_2(\lambda,\mu) = p_A + \frac{p_B}{1-\lambda\left(\frac{p_A}{b\mu} + \frac{p_B}{\mu}\right)}.$$

It is evident that $\frac{\partial \varphi_1}{\partial \mu} < 0$, $\frac{\partial \varphi_2}{\partial \mu} < 0$ and $\frac{\partial \varphi_1}{\partial \lambda} > 0$, $\frac{\partial \varphi_2}{\partial \lambda} > 0$.

(i) Taking the first order derivative of W with respect to λ, we have

$$\frac{\partial W}{\partial \lambda} = \frac{\partial \varphi_1}{\partial \lambda}\varphi_2 + \varphi_1\frac{\partial \varphi_2}{\partial \lambda} > 0.$$

Taking the first order derivative of W with respect to μ, we have

$$\frac{\partial W}{\partial \mu} = \frac{\partial \varphi_1}{\partial \mu}\varphi_2 + \varphi_1\frac{\partial \varphi_2}{\partial \mu} < 0.$$

(ii) Taking the second order derivative of W with respect to λ, we have

$$\frac{\partial^2 W}{\partial \lambda^2} = \frac{\partial^2 \varphi_1}{\partial \lambda^2}\varphi_2 + 2\frac{\partial \varphi_1}{\partial \lambda}\frac{\partial \varphi_2}{\partial \lambda} + \frac{\partial^2 \varphi_2}{\partial \lambda^2}\varphi_1,$$

$$\frac{\partial^2 \varphi_1}{\partial \lambda^2} = \left[\frac{p_A}{(b\mu)^2} + \frac{p_B}{\mu^2}\right]\left[\frac{-2\lambda^{-3}\left(\frac{1}{\lambda} - \frac{p_A}{b\mu}\right)^2 - 2\lambda^{-2}\left(\frac{1}{\lambda} - \frac{p_A}{b\mu}\right)(-\lambda^{-2})}{\left(\frac{1}{\lambda} - \frac{p_A}{b\mu}\right)^4}\right]$$

$$= \left[\frac{p_A}{(b\mu)^2} + \frac{p_B}{\mu^2}\right]\left[\frac{-2\lambda^{-4}\left(\frac{1}{\lambda} - \frac{p_A}{b\mu}\right)\left[\lambda\left(\frac{1}{\lambda} - \frac{p_A}{b\mu}\right) - 1\right]}{\left(\frac{1}{\lambda} - \frac{p_A}{b\mu}\right)^4}\right] > 0.$$

Similarly, we have

$$\frac{\partial^2 \varphi_2}{\partial \lambda^2} = p_B\left(\frac{p_A}{b\mu} + \frac{p_B}{\mu}\right)(-2)\left[1 - \lambda\left(\frac{p_A}{b\mu} + \frac{p_B}{\mu}\right)\right]^{-3}\left(-\left(\frac{p_A}{b\mu} + \frac{p_B}{\mu}\right)\right) > 0.$$

As a result, $\frac{\partial^2 W}{\partial \lambda^2} > 0$.

Taking the second order derivative of W with respect to μ, we have

$$\frac{\partial^2 W}{\partial \mu^2} = \frac{\partial^2 \varphi_1}{\partial \mu^2}\varphi_2 + 2\frac{\partial \varphi_1}{\partial \mu}\frac{\partial \varphi_2}{\partial \mu} + \frac{\partial^2 \varphi_2}{\partial \mu^2}\varphi_1,$$

$$\frac{\partial^2 \varphi_1}{\partial \mu^2} = -(p_A + b^2 p_B)\frac{\frac{1}{\lambda}2b^2\left(\frac{1}{\lambda}b^2\mu^2 - p_A b\mu\right)^2 - \left(\frac{1}{\lambda}2b^2\mu - p_A b\right)^2 2\left(\frac{1}{\lambda}b^2\mu^2 - p_A b\mu\right)}{\left(\frac{1}{\lambda}b^2\mu^2 - p_A b\mu\right)^4}$$

$$= -(p_A + b^2 p_B)\frac{3\frac{1}{\lambda}b^3\mu\left(p_A - \frac{b\mu}{\lambda}\right) - {p_A}^2 b^2}{\left(\frac{1}{\lambda}b^2\mu^2 - p_A b\mu\right)^4}.$$

As $\lambda_A = p_A\lambda < b\mu$ (the queue is limited), we obtain $\frac{\partial^2 \varphi_1}{\partial \mu^2} > 0$.

Similarly, we have

$$\frac{\partial^2 \varphi_2}{\partial \mu^2} = -p_B\left[\lambda\left(\frac{p_A}{b} + p_B\right)\right]\frac{-1}{\left[\mu - \lambda\left(\frac{p_A}{b} + p_B\right)\right]^4} > 0.$$

As a result, $\frac{\partial^2 W}{\partial \mu^2} > 0$.

Because $\frac{\partial W}{\partial \mu} = \frac{\partial \varphi_1}{\partial \mu}\varphi_2 + \varphi_1\frac{\partial \varphi_2}{\partial \mu}$, we have

$$\frac{\partial W(\lambda, \mu)}{\partial \mu \partial \lambda} = \frac{\partial^2 \varphi_1}{\partial \mu \partial \lambda}\varphi_2 + \frac{\partial \varphi_1}{\partial \mu}\frac{\partial \varphi_2}{\partial \lambda} + \frac{\partial \varphi_1}{\partial \lambda}\frac{\partial \varphi_2}{\partial \mu} + \frac{\partial^2 \varphi_2}{\partial \mu \partial \lambda}\varphi_1.$$

Furthermore, we can get

$$\frac{\partial^2 \varphi_1}{\partial \mu \partial \lambda} = -(p_A + b^2 p_B)\frac{(2b^2\mu - 2\lambda b p_A)(b^2\mu^2 - \lambda p_A b\mu)^2 - (2\lambda b^2\mu - \lambda^2 b p_A)2(b^2\mu^2 - \lambda p_A b\mu)(-p_A b\mu)}{(b^2\mu^2 - \lambda p_A b\mu)^4} < 0;$$

similarly, we can also confirm that $\frac{\partial^2 \varphi_2}{\partial \mu \partial \lambda} < 0$.

Because $\frac{\partial \varphi_1}{\partial \mu} < 0$, $\frac{\partial \varphi_2}{\partial \lambda} > 0$, $\frac{\partial \varphi_1}{\partial \lambda} > 0$, and $\frac{\partial \varphi_2}{\partial \mu} < 0$, we have $\frac{\partial^2 W(\lambda,\mu)}{\partial \mu \partial \lambda} < 0$.

Proof of Proposition 4

(i) In the short-run problem, the service provider's revenue function is

$$R(\lambda, \mu) = \lambda[T(q) + Q(\mu) - c_w W(\lambda, \mu)] - \lambda c_2 e^{-\gamma q}.$$

For a given value of λ,

$$\frac{\partial R}{\partial \mu} = \lambda\left(-\alpha - c_w \cdot \frac{\partial W}{\partial \mu}\right)$$

and

$$\frac{\partial^2 R}{\partial \mu^2} = -\lambda c_w \frac{\partial^2 W}{\partial \mu^2} < 0.$$

Therefore, we can find the optimal service rate $\mu^*(\lambda)$ by solving

$$-\alpha - c_w \cdot \frac{\partial W}{\partial \mu} = 0.$$

Let $h(\lambda)$ denote the optimal objective value of $T(q) + Q(\mu) - c_w W(\lambda, \mu) - c_2 e^{-\gamma q}$, i.e.,

$$h(\lambda) = \max_{\mu} \quad T(q) + Q(\mu) - c_w W(\lambda, \mu) - c_2 e^{-\gamma q}.$$

Because $W(\lambda, \mu)$ is convex in μ, $h(\lambda)$ exists and is continuous.

Next, consider the optimization problem of λ, which we write as follows:

$$R = \max_{\lambda} \quad g(\lambda, b) = \max_{\lambda} \quad \lambda h(\lambda, b).$$

In the above optimization problem, we treat the arrival rate λ as a decision variable while b as a parameter. We want to compare the optimal solutions for different values of b. Specifically, let λ_1 maximize $g(\cdot, b_1)$ and λ_2 maximize $g(\cdot, b_2)$, where b_1 and b_2 are randomly chosen with $b_1 < b_2$. Therefore, we have the following results

$$g(\lambda_1, b_1) \geq g(\lambda_2, b_1) \quad \text{and} \quad g(\lambda_2, b_2) \geq g(\lambda_1, b_2).$$

Subtracting the first inequality from the second, we obtain

$$g(\lambda_2, b_2) - g(\lambda_2, b_1) \geq g(\lambda_1, b_2) - g(\lambda_1, b_1).$$

Then, we define the function $l(\lambda) = g(\lambda, b_2) - g(\lambda, b_1)$ and the above inequality becomes

$$l(\lambda_1) \leq l(\lambda_2).$$

Then we have

$$0 \le l(\lambda_2) - l(\lambda_1) = \int_{\lambda_1}^{\lambda_2} l'(\lambda) d\lambda = \int_{\lambda_1}^{\lambda_2} \left(\frac{\partial g(\lambda, b_2)}{\partial \lambda} - \frac{\partial g(\lambda, b_1)}{\partial \lambda} \right) d\lambda.$$

Next, we write the integrand in the above inequality as an integral over b, i.e.,

$$\int_{\lambda_1}^{\lambda_2} \left(\frac{\partial g(\lambda, b_2)}{\partial \lambda} - \frac{\partial g(\lambda, b_1)}{\partial \lambda} \right) d\lambda = \int_{\lambda_1}^{\lambda_2} \int_{b_1}^{b_2} \frac{\partial^2 g(\lambda, b)}{\partial \lambda \partial b} db d\lambda \ .$$

As a result,

$$l(\lambda_2) - l(\lambda_1) = \int_{\lambda_1}^{\lambda_2} \int_{b_1}^{b_2} \frac{\partial^2 g(\lambda, b)}{\partial \lambda \partial b} db d\lambda \ge 0.$$

Next, we want to determine the sign of $\frac{\partial^2 g(\lambda,b)}{\partial \lambda \partial b}$.

As $g(\lambda, b) = \lambda h(\lambda, b)$, the second derivative of g is

$$\frac{\partial^2 g(\lambda,b)}{\partial \lambda \partial b} = \frac{\partial h(\lambda,b)}{\partial b} + \lambda \frac{\partial^2 h(\lambda,b)}{\partial \lambda \partial b}.$$

And we have

$$\frac{\partial h(\lambda,b)}{\partial b} = -\alpha \frac{\partial \mu^*}{\partial b} - c_w \left(\frac{\partial W}{\partial b} + \frac{\partial W}{\partial \mu} \frac{\partial \mu^*}{\partial b} \right),$$

$$\frac{\partial^2 h(\lambda,b)}{\partial \lambda \partial b} = -\alpha \frac{\partial^2 \mu^*}{\partial \lambda \partial b} - c_w \left(\frac{\partial^2 W}{\partial \lambda \partial b} + \frac{\partial^2 W}{\partial \mu \partial b} \frac{\partial \mu^*}{\partial \lambda} + \frac{\partial W}{\partial \mu} \frac{\partial^2 \mu^*}{\partial \lambda \partial b} \right).$$

Because the optimal μ^* satisfies $-\alpha - c_w \cdot \frac{\partial W}{\partial \mu} = 0$, the above equations become

$$\frac{\partial h(\lambda,b)}{\partial b} = -c_w \frac{\partial W}{\partial b},$$

$$\frac{\partial^2 h(\lambda,b)}{\partial \lambda \partial b} = -c_w \left(\frac{\partial^2 W}{\partial \lambda \partial b} + \frac{\partial^2 W}{\partial \mu \partial b} \frac{\partial \mu^*}{\partial \lambda} \right).$$

Then we confirm the following results are valid.

Remark 1: $\frac{d\mu^*}{d\lambda} < 0$.

Proof: Taking the derivative of μ with respect to λ by the implicit function theory, we have

$$c_w \left[\frac{\partial W(\lambda, \mu)}{\partial \lambda} + \frac{\partial^2 W(\lambda, \mu)}{\partial \mu^2} \frac{\partial \mu}{\partial \lambda} \right] = 0.$$

By Proposition 3 , $\frac{\partial W(\lambda,\mu)}{\partial \mu \partial \lambda} > 0$ and $\frac{\partial^2 W(\lambda,\mu)}{\partial \mu^2} > 0$, so we have $\frac{\partial \mu^*}{\partial \lambda} < 0$.

Remark 2: $\frac{\partial^2 W}{\partial \lambda \partial b} < 0$ and $\frac{\partial^2 W}{\partial \mu \partial b} > 0$

Proof:

$$\frac{\partial^2 W}{\partial \lambda \partial b} = \frac{\partial^2 \varphi_1}{\partial \lambda \partial b} \varphi_2 + \frac{\partial \varphi_1}{\partial \lambda} \frac{\partial \varphi_2}{\partial b} + \frac{\partial^2 \varphi_2}{\partial \lambda \partial b} \varphi_1 + \frac{\partial \varphi_2}{\partial \lambda} \frac{\partial \varphi_1}{\partial b},$$

$$\frac{\partial \varphi_1}{\partial \lambda} = \left(\frac{p_A}{b^2\mu^2} + \frac{p_B}{\mu^2}\right)\left[\frac{\frac{1}{\lambda^2}}{\left(\frac{1}{\lambda} - \frac{p_A}{b\mu}\right)^2}\right].$$

Since both $\left(\frac{p_A}{b^2\mu^2} + \frac{p_B}{\mu^2}\right)$ and $\frac{\frac{1}{\lambda^2}}{\left(\frac{1}{\lambda} - \frac{p_A}{b\mu}\right)^2}$ are positive and decrease in b, we have $\frac{\partial^2 \varphi_1}{\partial \lambda \partial b} < 0$. Similarly, we have $\frac{\partial^2 \varphi_2}{\partial \lambda \partial b} < 0$. Becasue $\frac{\partial \varphi_2}{\partial b} < 0$ and $\frac{\partial \varphi_1}{\partial b} < 0$ can be easily obtained, we finally conclude that $\frac{\partial^2 W}{\partial \lambda \partial b} < 0$. The result $\frac{\partial^2 W}{\partial \mu \partial b} < 0$ can be found using the same method.

Combining the results in Remarks 1 and 2, we have $\frac{\partial^2 h(\lambda,b)}{\partial \lambda \partial b} > 0$.

Because we have already confirmed that

$$l(\lambda_2) - l(\lambda_1) = \int_{\lambda_1}^{\lambda_2}\int_{b_1}^{b_2} \frac{\partial^2 g(\lambda, b)}{\partial \lambda \partial b} db d\lambda \geq 0$$

and $b_1 < b_2$, we can confirm that $\lambda_1 < \lambda_2$.

Having shown $\lambda_1 < \lambda_2$, we confirm that $\mu_1 > \mu_2$ by solving the sub-problem

$$\max_{\mu} \quad h(\mu, \lambda) = T(q) + Q(\mu) - c_w \cdot W(\mu, \lambda) - c_2 e^{-\gamma q}.$$

In the above optimization problem, we treat μ as a decision variable and λ as a parameter, where $\lambda_1 < \lambda_2$. Let μ_1 maximize $h(\cdot, \lambda_1)$ and μ_2 maximize $h(\cdot, \lambda_2)$. Because

$$\frac{\partial^2 h(\mu, \lambda)}{\partial \mu \partial \lambda} = -c_w \frac{\partial^2 W(\mu, \lambda)}{\partial \mu \partial \lambda} > 0,$$

by repeating a similar process as above, we confirm that $\mu_1 > \mu_2$.

Therefore, we derive that the arrival rate λ increases but the service rate μ decreases with the accelerating factor b.

(ii) The first-order condition is

$$-\alpha - c_w \cdot \frac{\partial W(\mu(\lambda), \lambda)}{\partial \mu} = 0.$$

As a result, $\frac{\partial W(\mu,\lambda)}{\partial \mu} = -\frac{\alpha}{c_w}$.

We differentiate the above first-order condition implicitly with respect to λ to obtain

$$0 = \frac{\partial^2 W(\mu, \lambda)}{\partial \mu^2}\frac{d\mu(\lambda)}{d\lambda} + \frac{\partial W(\mu, \lambda)}{\partial \lambda}.$$

Thus,

$$\frac{d\mu(\lambda)}{d\lambda} = -\frac{\frac{\partial W(\mu,\lambda)}{\partial \lambda}}{\frac{\partial^2 W(\mu,\lambda)}{\partial \mu^2}}.$$

Considering $\frac{d(\mu(\lambda),\lambda)}{d\lambda}$, we write

$$\frac{dW(\mu(\lambda),\lambda)}{d\lambda} = \frac{\partial W(\mu(\lambda),\lambda)}{\partial \mu}\frac{d\mu(\lambda)}{d\lambda} + \frac{\partial W(\mu(\lambda),\lambda)}{\partial \lambda}.$$

Substituting $\frac{\partial W(\mu,\lambda)}{\partial \mu}$ and $\frac{d\mu(\lambda)}{d\lambda}$ into the above equation, we have

$$\frac{dW(\mu(\lambda),\lambda)}{d\lambda} = \frac{\alpha}{c_w}\frac{\frac{\partial W(\mu,\lambda)}{\partial \lambda}}{\frac{\partial^2 W(\mu,\lambda)}{\partial \mu^2}} + \frac{\partial W(\mu(\lambda),\lambda)}{\partial \lambda} = \frac{\partial W(\mu(\lambda),\lambda)}{\partial \lambda}\left(\frac{\alpha}{c_w}\frac{1}{\frac{\partial^2 W(\mu,\lambda)}{\partial \mu^2}} + 1\right) > 0.$$

We can derive that the expected waiting time increases in the arrival rate from the above result. In addition, the expected waiting time W is continuous and differentiable in b. Since the arrival rate λ increases in b, the above results show that the expected waiting time W also increases in b.

(iii) The market equilibrium price is $p = T(q) + Q\big(\mu^*(\lambda)\big) - c_w \cdot W\big(\lambda, \mu^*(\lambda)\big) - c_2 e^{-\gamma q}$. We differentiate p with respect to λ and obtain

$$\frac{\partial p}{\partial \lambda} = -\alpha\frac{d\mu^*(\lambda)}{d\lambda} - c_w\left(\frac{\partial W}{\partial \lambda} + \frac{\partial W}{\partial \mu}\frac{\partial \mu^*}{\partial \lambda}\right).$$

Because $-\alpha - c_w \cdot \frac{\partial W}{\partial \mu} = 0$ at optimality, we have

$$\frac{\partial p}{\partial \lambda} = -c_w\frac{\partial W}{\partial \lambda} < 0.$$

Since the arrival rate λ increases in b, the above result shows that the optimal price p decreases in b.

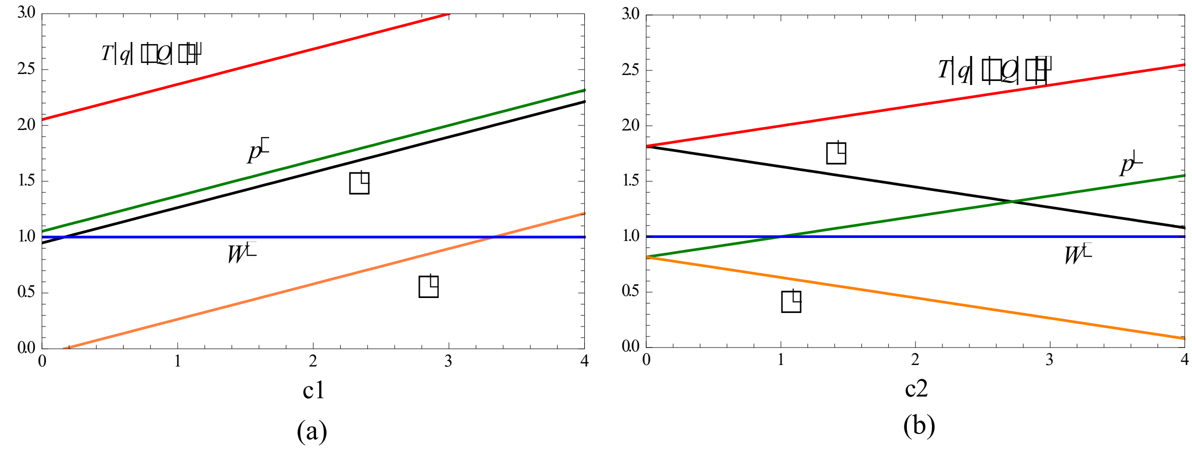
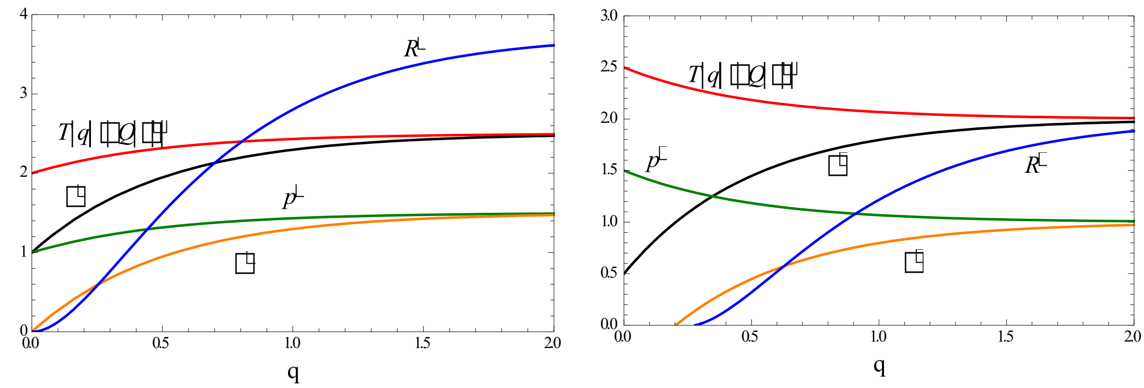


Figure 1 Effects of c_1 (a) and c_2 (b) on the service provider's equilibrium decisions.



(a) $c_1 = 2$ and $c_2 = 1$ (b) $c_1 = 1$ and $c_2 = 2$

Figure 2 Effects of q on the service provider's equilibrium decisions.

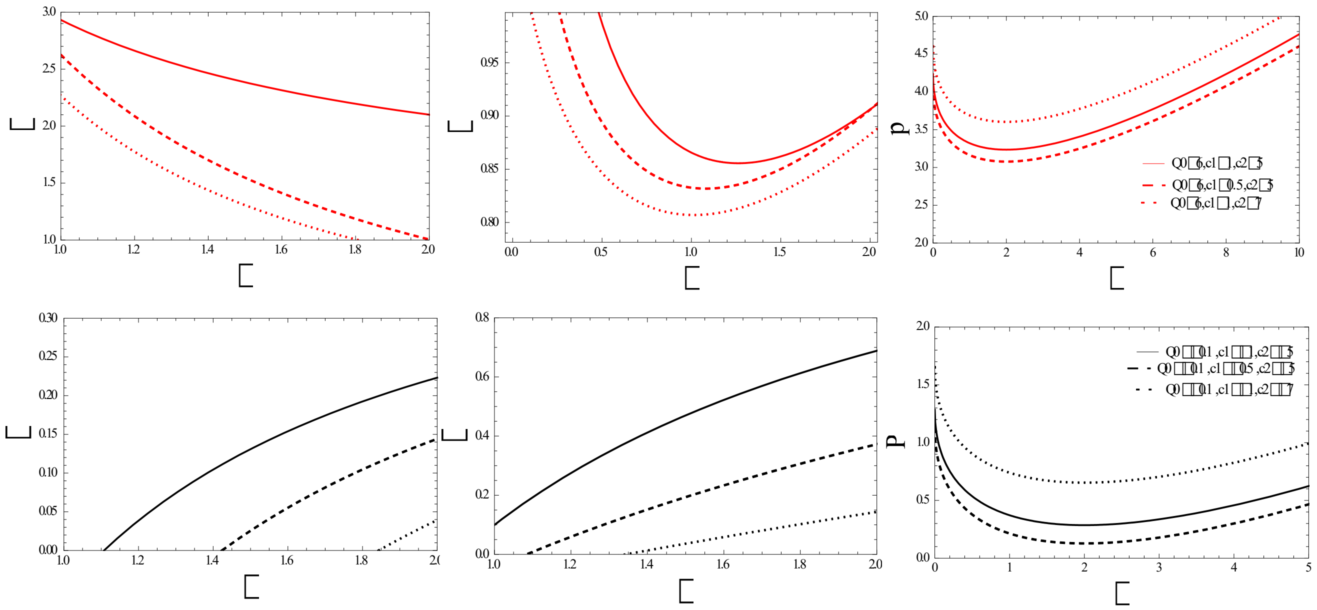


Figure 3 Effects of α on the service provider's equilibrium decisions

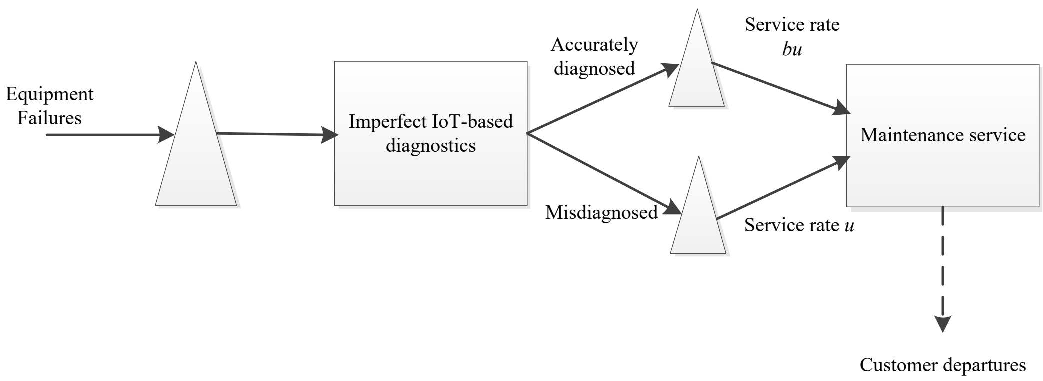


Figure 4 Illustration of the service process with the accelerating effect

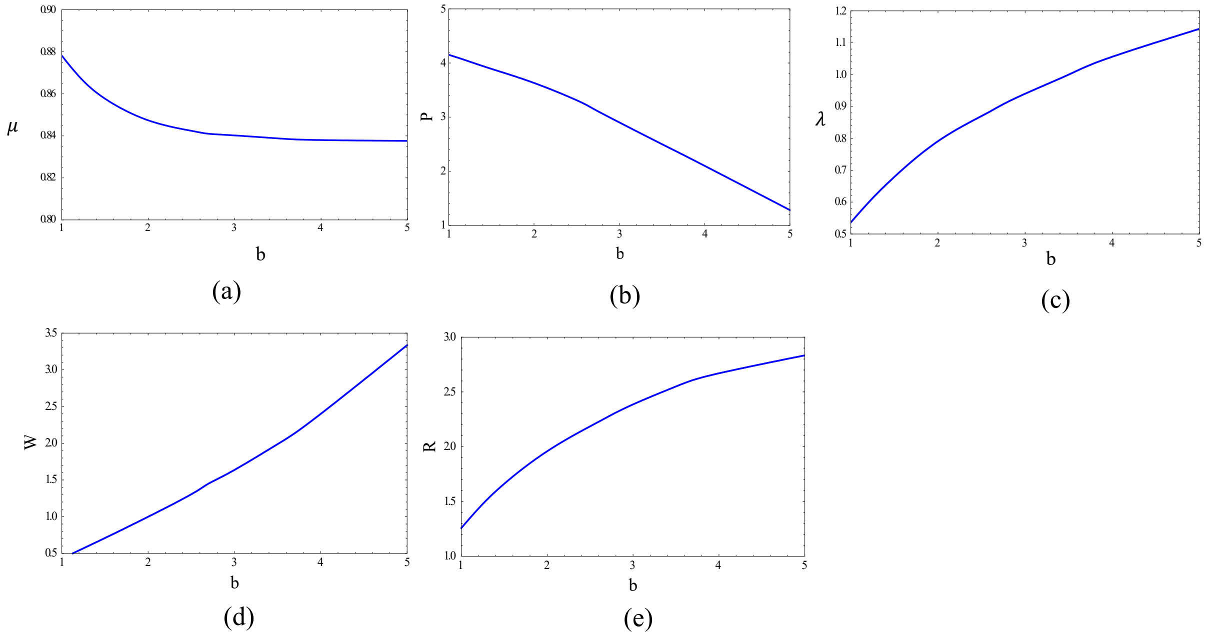


Figure 5 Effects of *b* on the equilibrium decisions in the short-run problem

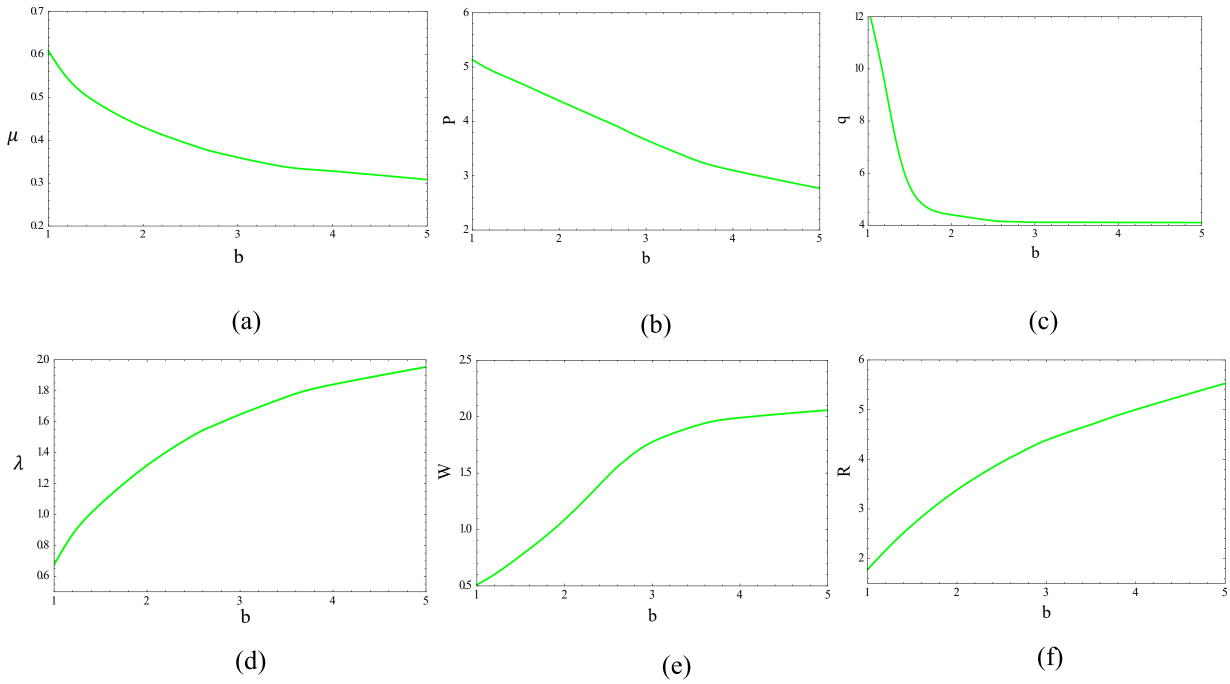


Figure 6 Effects of *b* on the equilibrium decisions in the long-run problem