Joint Capacity and Contract Management for Operating Service Facilities

Yabing Jiang  
Graduate School of Business Administration  
Fordham University  
yajiang@fordham.edu

Abraham Seidmann  
Simon Graduate School of Business Administration  
University of Rochester  
seidmannav@simon.rochester.edu

Abstract

Capacity planning and incentive contract design have always been challenging strategic decisions, especially for companies operating in a stochastic service demand and delivery environment. Such service facilities often must make capacity decisions long before observing demand, so a practical approach is to make the decision based on the expected demand distribution. Due to stochastic demand and capacity constraints, service centers often incur delays and delay-related costs when the actual demand is high. When the actual demand is low, however, centers have to bear high idle capacity costs. Economics literature on principal-agent issue has mostly focused on designing salesforce compensation plans so as to motivate the agent and manage risk-sharing. However, in a typical principal-agent model setting, the principal faces no capacity limit and incurs no delay-related costs. Furthermore, the capacity decision and the compensation decision are usually treated as two separate decision processes both in practice and in research. This suboptimal decision process leads to suboptimal results: the principal either incurs high delay costs or high idle capacity costs. To tackle service facility capacity planning and incentive contract issue, we propose instead an integrated approach. That is, we integrate a firm’s decision regarding capacity investment with its decision regarding the design of a compensation contract. We outline this integrated decision approach and illustrate its benefits with numerical examples. We show that following our decision methodology the firm can achieve significantly higher profits. Several cases in the paper depict that the firm can achieve profit increments ranging from 10% to 42% by properly integrating contract design with its capacity decision process early on.

1. Introduction and overview

Capacity planning is a critical economic decision and a challenge for service facilities that often deal with stochastic demands. In many cases, management faces capacity constraints and may incur significant delay and capacity costs. Yet with few exceptions existing research on incentive contract design does not consider these issues. In theoretical research on incentive contract design, see [3, 7, 8, 10], the principal offers the optimal compensation plan to motivate the agent and manage risk-sharing. These studies assume that the principal faces no capacity limit and incurs no delay-related costs.

In the service industries waiting time is commonly used as a measure for both internal and external performance evaluations; firms often commit to certain service levels as measured by the expected waiting time [1]. How the principal should compensate its agent, knowing that the agent does not consider delay costs when deciding on his effort, is still an open research topic.

Banker et al. [2] study the relevant costs of capacity by explicitly modeling delay costs in a stochastic manufacturing environment. Using an M/G/1 queuing model, they evaluate and estimate the expected delay costs, specifically WIP inventory costs, which are shown to be enormous. Although recognizing that capacity-related delay costs are relevant for firms’ product mix decisions, they ignore the issues of management incentives, capacity planning, and agency costs.

In this paper, we address the problem by integrating the principal’s incentive contract design with its capacity decision process and introducing a delay costs term in the principal’s profit function. Our study extends the results of Basu et al. [3], who focus on the compensation contract design of a sales force selling a commodity good with an unlimited supply. In contrast, we model the capacity management and contract design issue of a service organization with a finite capacity. By incorporating delay costs and capacity constraints in the model, we extend existing economics literature on principal-agent problems to a new setting and contribute to current research on incentive contract design and resource management. Our study provides guidelines for firms that deal with congestion-prone systems and sheds light on how to
balance agency, capacity, and delay costs with the design of an incentive contract, which provides correct management incentives for different market conditions.

The rest of the paper is organized as follows. We first present our basic model where the capacity level is given and the expected market demand is known. We then extend the model to the case where the capacity decision is endogenous and the expected market demand is a random variable at the time of capacity decision making. We next analyze three linear contracts and illustrate the advantage of the integrated capacity approach over the exogenous capacity case with some numerical examples in section 3. We conclude the paper in section 4.

2. The model

We address the issue of incentive contract design within a service center operating under capacity constraints. Consider, for example, computerized medical imaging centers. These facilities are usually capital intensive and can be profitable only with enough patients. For example, typical costs for modern MRI or CT units range from one to three million dollars [4], and the costs for a new proton therapy center are usually above one hundred million dollars [11]. In addition, such centers’ capacity levels are usually fixed in the short run; they can process only a limited number of tests per unit time. Random variations in test time per patient, the stochastic arrival of test requests, and the normal scheduling approach can give rise to queuing delays at the imaging center for patients [6]. The centers incur delay-related costs in the form of patient dissatisfaction, delayed diagnoses, and possible losses in reputation or future demand from referring physicians.

We study this service center incentive contract design problem using a principal-agent framework. Consider a computerized medical imaging center with one principal (the investor) and one agent (the executive running the center). We assume that the arrival of diagnostic test requests is a random process. From time to time, patients have to wait to take a diagnostic test. The principal must first determine how much capacity to install, and the agent running the center is held responsible for marketing and for getting imaging referrals from the local medical community. Because the agent’s effort is not directly verifiable, the principal can only offer an incentive contract that rewards the agent based on the realized demand. That is, the principal must decide on the capacity and the compensation contract well before observing the number of patients to be processed at the center.

Next, we study the principal’s incentive contract design problem in a basic model and an extended model, depending on whether the center’s capacity level is given and the expected base demand is known at the time of capacity decision making.

2.1. The basic model

Let $N$, a random variable, be the demand of the center. The arrival of patients ($N$) follows a stochastic process with rate $\lambda$. Suppose the center capacity $\mu$ is predetermined and the expected base demand $\theta$ is known in this basic model. This base demand is the expected demand without the agent’s marketing input, and it is affected by factors such as pricing, advertising, and other decisions made by the principal. When the agent invests marketing effort $\alpha$, $\alpha \in \Lambda \subseteq R^+$, the expected demand will increase by $\phi(\alpha)$, with $\phi'(\alpha) > 0$ and $\phi''(\alpha) \leq 0$. That is, the expected demand increases with the agent’s marketing effort, but at a decreasing rate. Given the expected base demand $\theta$ and the agent’s effort $\alpha$, the expected demand $\lambda$ is characterized by $\lambda = \theta + \phi(\alpha)$. The capacity level $\mu$ is measured by the expected number of tests that can be processed per unit time. Similar to [3], we assume that both the principal and the agent know the effort response function. The principal cannot observe and has no means of directly verifying the agent’s effort level, which is known only to the agent, but both parties can observe the realized demand. The principal would like the agent to invest marketing effort to maximize the principal’s net benefit, which is given by revenue minus delay costs and capacity costs. The agent’s interest differs, because he incurs a disutility of effort. Hence, the principal must design a compensation scheme based on observable variables, such as the realized number of tests, to encourage the agent to exert a desired level of effort. Here we focus on a simple linear contract in the form of $s(N) = c_0 + c_1 N$, which is commonly used in practice.

Following general agency theory assumptions, the principal is risk-neutral, and the agent is both risk- and effort-averse. We assume that the agent’s utility function for compensation and effort is additively separable and can be expressed as $U(s) = V(\alpha)$. The agent realizes utility $U(s)$ from the compensation $s$, and this utility function satisfies $U(s) \geq 0$, $U'(s) > 0$, and $U''(s) < 0$; i.e., the agent has diminishing marginal utility for income. At an effort level $\alpha$, the agent incurs

---

1 Because $N$ is a random variable, the principal cannot infer the value of $\alpha$ even after observing the expected base demand $\theta$ and the realized value of $N$. 
We assume that $V(\alpha) \geq 0$, $V'(\alpha) > 0$, and $V''(\alpha) > 0$; i.e., the agent’s marginal disutility from effort increases with effort.

Because of demand uncertainty and the stochastic arrival process, given the fixed capacity level $\mu$, from time to time patients have to wait for the diagnostic test. Let $W$ be the expected waiting time for a random patient and $\tau$ be the principal’s per patient, per unit time delay cost, which represents the cost of a delayed diagnosis, an extended hospital stay, a patient’s opportunity cost, etc. Because patients’ waiting times and waiting costs directly affect patient satisfaction, the center’s operating costs and perceived service quality, the center’s reputation, and future demand, $\tau W$ times the expected number of patients in the system reveals the principal’s total expected delay costs.

Knowing the expected base demand $\theta$, the principal’s problem is characterized by the following optimization problem:

$$\max_{\alpha \in A} (\theta + \phi(\alpha)) p - (\theta + \phi(\alpha)) \tau W - E_N[s(N)] \quad (1)$$

subject to: $E_N[U(s(N)) - V(\alpha)] \geq M \quad (IR)$

$$\alpha \in \arg \max_{\alpha \in A} E_N[U(s(N)) - V(\alpha^*)]. \quad (IC)$$

The principal will decide the optimal effort $\alpha$ and compensation contract terms $c_0$ and $c_1$ to maximize its expected profit. The first term in equation (1) is the principal’s expected revenue, which is given by price ($p$) times the expected demand $\lambda = \theta + \phi(\alpha)$. We assume that the marginal cost of processing a test is constant, and it is normalized to zero. The second term of equation (1) is the principal’s expected delay cost. Implicitly, we assume that $\mu > \lambda$, because otherwise the delay costs will explode. We model the operation at the facility as an M/M/1 queue; it has a Poisson arrival rate and exponential service time with rate $\mu$. For the M/M/1 queue, in steady state we have $W = W(\mu - \theta - \phi(\alpha))$. The third term, which can be expressed as $E_N[s(N)] = c_0 + c_1 \lambda$, is the expected payment to the agent.

Here we do not directly model patients’ choices, although it is true that extensive delay at an MRI center may drive patients away. Instead, we model the principal’s decision in the steady state since an individual referral, in the short run, will tend to ignore the excessive waiting time. On the long run, excessive waiting times will reduce the reputation of the imaging center, and will limit its ability to accommodate urgent studies. Excessive delay will also lead to additional costs to the principal, such as costs for the patients’ extended hospital stays and treatments while waiting for the diagnostic tests and the costs incurred for comforting the anxious and unhappy patients and controlling the effect of negative word of mouth. Accounting for this loss of reputation and increased operation costs, the principal needs to endogenize the negative impact of the waiting time externality (Mendelson [9]). At a given capacity level, the principal decides the optimal contract terms by balancing delay costs and agency costs. The effect of a long waiting time as well as the negative externality created by a patient joining the queue on operation cost and future demand is modeled as delay-related costs incurred by the principal in our model. The principal thus endogenously determines the optimal expected demand, considering both agency issues and the effect of waiting times.

Next, consider the constraints. The first is the agent’s individual rationality constraint. He will accept the compensation contract $s(N)$ only if his expected net benefit, given by the difference between his expected utility from compensation $E_N[U(s(N))]$ and his disutility of effort $V(\alpha)$, is at least as much as $M$, his gain from an outside alternative. The second constraint is the agent’s incentive compatibility constraint. That is, given a compensation contract $s(N)$, the agent will choose an optimal effort $\alpha$ to maximize his expected net benefit. Note that $p$, $\tau$, and $M$ are exogenous in this model, and $\alpha$ belongs to the set of the agent’s optimal effort choices. Because the agent’s net-utility function is concave in effort $\alpha$, we can replace the agent’s (IC) constraint with the first-order condition of the agent’s net-utility function.$^2$

This basic model setting differs from the normal principal-agent moral hazard setting in that we include a delay cost term in the model. Given the capacity level $\mu$ and observing the base demand $\theta$, the principal will decide on the optimal effort $\alpha$ and compensation contract terms $(c_0, c_1)$ that maximize equation (1) subject to the agent’s (IR) and (IC) constraints. The Lagrange form of the problem then is given by equation (2), where $g$ and $h$ are the Lagrange multipliers of the (IR) and (IC) constraints, respectively:

$$L = (\theta + \phi(\alpha)) p - (\theta + \phi(\alpha)) \tau W - E_N[s(N)] + g(E_N[U(s(N))])$$

$$- V(\alpha) - M + h[\phi(\alpha) E_N[U(s(N) + 1)] - U(s(N))] - V(\alpha). \quad (2)$$

$^2$ Consider the executive’s expected net-utility function $E_N[U(s(N))]$.

The second-order derivative with respect to $\alpha$ is given by

$$-V''(\alpha) + \phi(\alpha)^2 E_N[U(s(N) + 2)] + U(s(N)) - 2 U(s(N + 1)) + \phi'(\alpha) E_N[U(s(N) + 1) - U(s(N))] - V'(\alpha).$$

Since the executive’s net-utility function is concave in $\alpha$, we can replace the executive’s (IC) constraint with the first-order condition of the executive’s net-utility function, given by

$$-V'(\alpha) + \phi(\alpha) E_N[U(s(N) + 1) - U(s(N))].$$
Given capacity $\mu$ and the base demand $\theta$, our analysis can be reduced to solving the five Lagrange first-order conditions (FOCs) jointly for the optimal effort level and contract terms.

### 2.2. The extended model

We relax the assumption in the basic model that the center capacity $\mu$ and base demand $\theta$ are given. Instead, we consider the case in which the principal will optimally decide the center's capacity level, and at the time of capacity decision the expected base demand is only known to follow certain distribution. We assume that the expected base demand $\theta$ is a random variable drawn from a set $\Theta = \{\theta_1, \theta_2, ..., \theta_j\}$ with a probability function $\psi(\theta)$. As we discussed above, this base demand is exogenous, without the agent’s marketing input, and it is affected by decisions made by the principal, such as pricing and advertising.

Because of demand uncertainty, the principal needs to decide how much capacity to install and design a set of compensation contracts $S = \{s_1(\cdot), \ldots, s_k(\cdot)\}$, which maps each base demand $\theta_i$ to a compensation contract $s_i(\cdot)$. Because of the long lead time of computerized medical imaging equipment as well as the normal budgeting process, we assume that the principal has to make the capacity decision before observing the base demand $\theta_i$, and it does not have the flexibility to adjust the capacity level afterward. We also assume that after the capacity is installed and the principal’s pricing and advertising decisions are made, both the principal and the agent possess the same information about the realized base demand. That is, in our model the agent does not possess private information regarding the base demand. We focus on the moral-hazard problem; however, it is also interesting to investigate the adverse selection problem by allowing the agent to have private information about the base demand. Here, the principal has to balance agency, capacity, and delay costs when making the capacity and contracting decisions.

The extended model differs from the basic model in that here the principal has to make capacity decision and the expected base demand $\theta$, is not observable at the time of capacity decision making. Thus the principal has to design a set of optimal contracts $S(N)$ which maps each base demand $\theta_i$ with a contract $s_i(N)$, and the principal’s capacity decision is based on the distribution of $\theta_i$ and the corresponding agent effort in response to the optimal contract $s_i(N)$.

For expositional purpose we can solve the principal’s problem in two steps. First, for each expected base demand $\theta_i$, we can solve for the optimal contract $s_i(N)$ and effort $\alpha_i$ with $\mu$ as a parameter, i.e., design an optimal contract $s_i(N)$ for each base demand case. Then we can solve for the optimal capacity $\mu$, given the optimal $s_i(N)$ and $\alpha_i$ derived in the first step.

First, given the optimal capacity and observing the expected base demand $\theta_i$, the principal’s problem is the same as the basic model studied in section 2.1, except that we need to add a subscript $i$ to all corresponding terms.

\[
\max_{\alpha_i, \theta_i} (\theta_i + \phi(\alpha_i)) p - (\theta_i + \phi(\alpha_i)) \tau \psi - E_s(s_i(N)) \quad (1')
\]

subject to:

\[
E_s[U(s_i(N))] - V(\alpha_i) \geq M \quad (IR')
\]

\[
\alpha_i = \arg \max_{\alpha_i \in \Theta_s} E_s[U(s_i(N))] - V(\alpha_i) \quad (IC')
\]

Second, consider the capacity decision. The principal needs to determine the optimal capacity, given the set of optimal contracts $S(N)$ and anticipating the agent’s effort levels in response to the optimal contracts in the future. The principal’s problem is described in equation (3), where $F(\mu)$ is the capacity cost function satisfying $F'(\mu) > 0$ and $F''(\mu) \geq 0$:

\[
\max_{\mu} \sum_{i=1}^k \psi(\theta_i) \left[p \lambda_i - \tau \psi - E_N[s_i(N)] - F(\mu) \right] \quad (3)
\]

Here the capacity cost $F(\mu)$ includes costs of resources needed to supply the capacity level $\mu$. The resources include diagnostic equipment, technologists and radiologists, front and back office staff, and real-estate space, etc. which jointly determine the capacity of the image center.

For given values of $\alpha$ and $\theta_i$, the first-order condition of equation (3) with respect to $\mu$ is given by equation (4), where $\partial F(\mu) / \partial \mu = \left(\mu - \lambda^2 \right)^{-1}$:

\[
\sum_{i=1}^k \psi(\theta_i) \lambda_i \left[\frac{m_i}{\mu} - \tau \psi - E_N[s_i(N)] - F(\mu) \right] = 0 \quad (4)
\]

The first term of equation (4) represents the marginal benefit of capacity in terms of reduced delay costs. The second term is the marginal cost of capacity. We see that at the optimal capacity level marginal benefit equals marginal cost of capacity.

Considering equations (1’) and (3) together, we can reduce the principal’s problem to solving a group of first-order conditions jointly for the optimal capacity, contract terms, and effort levels.

### 3. Numerical example

We illustrate the benefits of the optimal linear contract over other forms of linear contracts with numerical examples. Consider the case in which the agent’s utility from income $U_s(N)$ is in the form of $s(N)^{1/\delta}$, with $\delta < 1$ (a commonly used Cobb-Douglas utility function). With this utility function, the agent has constant relative risk aversion. We focus on the
case in which the base demand could be either $\theta_1$ or $\theta_2$ ($\theta_1 < \theta_2$) with probability $\psi(\theta_1)$ and $\psi(\theta_2)$, respectively.

Using the technique in Basu et al. [3], we transform the agent’s effort function so that it is linear in order to simplify our analysis. Let $\Lambda$ be the agent’s actual effort. Assume $\phi(\Lambda) = k\Lambda^n$, with $k > 0$ and $t < 1$, and $V(\Lambda) = m\Lambda^\alpha$, with $m > 0$ and $t > 1$, so that the effort response function increases with the agent’s effort at a decreasing rate and the agent’s disutility function increases with effort at an increasing rate. Let $t = t_2/t_1$. By making the transformation $\alpha = \Lambda^n$, we have $\phi(\alpha) = k\alpha$ and $V(\alpha) = m\alpha^\alpha$, with $t > 1$.

To simplify the analysis, we assume that the capacity cost function follows $F(\mu) = b\mu$. The linear capacity cost assumption is consistent with prior research that models service facilities and is used as a good approximation in cost accounting for stepwise cost function; see Zimmerman [13], Mendelson [9], Whang [12], and Dewan and Mendelson [5], among others.

Consider a computerized imaging center operated in an environment described by the following parameters: $b = 50$, $k = 2$, $m = 1$, $M = 70$, $p = 150$, $t = 3/2$, $\tau = 100$, $\theta_1 = 20$/day, $\theta_2 = 30$/day, and $\psi(\theta_1) = \psi(\theta_2) = \frac{1}{2}$. In this example, the principal expects that the daily base demand of the year could be either high or low with equal probability. We break market demand down to the daily level, but this does not indicate that the capacity and the compensation contract are adjusted on a daily basis.

As a benchmark, we first consider a straight-salary contract. If the principal offers a straight-salary contract, the compensation is not connected to the facility’s performance so the agent will not invest any marketing effort to increase demand. In this example, the straight-salary contract is given by $s = 1225$. With this contract the principal will install a capacity that can process 35.7 tests per day. The expected waiting time for a patient is about 0.06 days in the low-demand market and 0.18 days in the high-demand market. Moreover, utilization $\rho(0)$ is only 56% in the low-demand market.

Alternatively, if adopting the optimal linear contract the principal will offer $s(\bar{N}) = 54.77N$ when the realized base demand is $\theta_1$ and $s(\bar{N}) = 39.29N$ for the case of $\theta_2$. With this contingent linear contract, the principal is able to achieve a higher demand at either $\theta_1$ or $\theta_2$ by inducing marketing effort from the agent. We find that the expected demand is about 27 tests per day for $\theta_1$ and 34 tests per day for $\theta_2$. Accordingly, the principal installs a higher capacity ($\mu(\bar{\alpha})$) that can process about 40.5 tests per day. In addition, for $\theta_2$, the expected waiting time $W_\mu(\bar{\alpha})$ is only 0.16 days, which is shorter than under the straight-salary contract. Table 1 contains a detailed comparison of the results.

<table>
<thead>
<tr>
<th></th>
<th>Optimal Linear Contract</th>
<th>Straight-salary Contract</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected Demand</td>
<td>$\lambda_1(\alpha)$ = 27.22</td>
<td>$\lambda_0(\alpha)$ = 20</td>
</tr>
<tr>
<td></td>
<td>$\lambda_2(\alpha)$ = 123.13</td>
<td>$\lambda_0(\alpha)$ = 30</td>
</tr>
<tr>
<td>Capacity</td>
<td>$\mu_1(\alpha)$ = 40.48</td>
<td>$\mu_0(\alpha)$ = 35.7135</td>
</tr>
<tr>
<td>Utilization</td>
<td>$\rho_1(\alpha)$ = 0.6725</td>
<td>$\rho_0(\alpha)$ = 0.5600</td>
</tr>
<tr>
<td>Expected Waiting Time</td>
<td>$W_1(\alpha)$ = 0.0754</td>
<td>$W_0(\alpha)$ = 0.0636</td>
</tr>
<tr>
<td></td>
<td>$W_2(\alpha)$ = 0.1574</td>
<td>$W_0(\alpha)$ = 0.1750</td>
</tr>
<tr>
<td>Expected Profit</td>
<td>$E_1(\alpha)$ = 790.07</td>
<td>$E_0(\alpha)$ = 413.15</td>
</tr>
</tbody>
</table>

We see that with the optimal contingent linear contract the principal installs 13.3% more capacity, and the facility’s performance is improved. Specifically, the principal realizes 91.2% more profits, and the facility has 10.1% reduced waiting times when the base demand is high and 20.1% more capacity utilization when it is low.

While companies may not have the flexibility to change capacity once the investment is made, they can manage capacity utilization by offering proper incentives to facility managers. This numerical example vividly demonstrates that when the realized market condition is weak, to reduce idle capacity a company can offer a high commission term in the compensation contract and thus induce a higher level of marketing effort from the site manager. When the realized market condition is good, to avoid a congested system, a company can offer a low commission term in the contract so that the site manager will not over-invest in marketing effort and further clog the system. Thus firms can efficiently manage capacity and significantly improve performance by installing the optimal capacity and strategically adjusting the compensation contract contingent on the realized market condition.

Next, consider the same set of parameter values described above, except that $\theta_1$ varies from 23 to 33. Here, each $(\theta_1, \theta_2)$ pair corresponds to a possible scenario faced by the center. Suppose the principal naively offers a simple universal linear contract that does not vary with the base demand. We find that the optimal universal linear contract is given by $s_1(N) = 54.77N$. Because this universal contract overcompensates the agent when the base demand is high, the agent will invest more effort than he would under the contingent linear contract. As a result, the
principal installs more capacity, and the agent’s (IR) constraint is not binding when the base demand is high. Without contract adjustments, the expected profit first increases and then decreases in $\theta_h$ because the increase in compensation and delay costs is higher than the increase in gross profits when the parameter $\theta_h$ is high. Figure 1 shows the expected profit levels under the contingent-linear, universal-linear, and straight-salary contracts with $\theta_l = 20$ and $\theta_h$ varying from 23 to 33.

\[ \begin{align*} &\text{Figure 1: Expected profits under different contracts, with } \theta_l = 20 \text{ and } \theta_h \text{ varying from } 23 \text{ to } 33. \\
&\text{Without contract adjustment, the universal linear contract overcompensates the agent in high-demand markets, and the increase in revenue is mostly used to cover the increase in delay costs and agent compensation. Thus, the profit under the universal linear contract is very flat. Because the contingent linear contract induces desired levels of marketing efforts from the agent it outperforms the universal linear and straight-salary contracts.}
\end{align*} \]

Next, we compare our extended model of an integrated capacity and contract decision with the case in which the capacity is exogenously given. To be comparable with the extended model, here the exogenous capacity case differs from the basic model in that the expected base demand could be either $\theta_l$ or $\theta_h$. Consider the same set of parameter values described above. When the principal integrates the capacity and contract decision the optimal endogenous capacity level, which corresponds to each $(\theta_l, \theta_h)$ pair, ranges from 35.84 to 42.94 tests per day. Consider two scenarios with the exogenous capacity level fixed at $\mu = 40$, which is within the set of the optimal capacity levels, and $\mu = 45$, which is above the optimal capacity level. Figure 2 compares the expected profits at $\mu = 40$, $\mu = 45$, and the optimal capacity.

\[ \begin{align*} &\text{When the capacity is exogenously given, the optimal capacity could be higher or lower than the given capacity. In the former case, because the delay costs are higher the expected profit is reduced. In the latter case, the principal incurs too much idle capacity so the profit decreases as well. The profit curves at } \mu = 40 \text{ and } \mu = 45 \text{ in Figure 2 clearly indicate the disadvantage of separating the capacity and contract decision making.}
\end{align*} \]

\[ \begin{align*} &\text{Figure 2: Expected profits at } \mu=40, \mu=45, \\
&\text{and the optimal capacity, with } \theta_l = 20 \text{ and } \\
&\theta_h \text{ varying from 23 to 33.}
\end{align*} \]

When the exogenous capacity is close to the optimal capacity, the expected profit in the exogenous capacity case is very close to that of the optimal capacity case (over 95%). However, when the exogenous capacity is well above or below the optimal capacity, the expected profit deviates significantly from the integrated capacity decision case. In this example, when $(\theta_l, \theta_h) = (20, 23)$ the optimal capacity is 35.84. If the capacity is exogenous given at $\mu = 45$, the profit is only 58% of the integrated capacity case. When $(\theta_l, \theta_h) = (20, 33)$ the optimal capacity is 42.94. If the capacity is exogenous given at $\mu = 40$, the profit is only 87% of the integrated capacity case. Hence the example illustrates the advantage of integrating a firm’s contract design with its capacity decision process.

4. Conclusions

Capacity planning and incentive contract design decisions present a major challenge for managers and principals operating service facilities such as computerized medical imaging centers. In many cases, firms have to make capacity decisions before observing the demand for their services; moreover, demand may fluctuate substantially. This paper introduces delay costs and the capacity decision into a principal-agent model and extends current research on incentive contract design.

In our model, the delay cost term captures the principal’s costs for keeping patients in line to take the diagnostic tests. This term can be modified to represent general costs caused by limited capacity. Thus our study is not restricted to computerized medical imaging centers and can be applied to other service facilities that face demand uncertainty, have a fixed capacity,
incur delay costs or degradation of service, and have to deal with agency issues.

Our main insights are as follows. First, because of demand uncertainty and an inflexible capacity investment process, companies should make their capacity decisions based on demand distribution and, most importantly, integrate their capacity decision with their compensation contract design so as to maximize total profit. When the capacity is exogenously given, the optimal capacity could be higher or lower than the given capacity. This means the principal either incurs high delay costs or high idle capacity costs, hence has to compromise in either case. By incorporating delay costs and capacity decision in a general principal-agent setting, our study provides guidelines for firms that deal with congestion-prone systems.

Second, companies should design a set of compensation contracts and offer an appropriate contract contingent on the realized market condition. With such contingent compensation contracts, they can provide correct management incentives and therefore induce desired levels of marketing effort from the on-site agent.

We compare the optimal contingent linear contract with the benchmark straight-salary contract and the universal linear contract in numerical examples. Comparing the numerical results of these three contract forms, we see that the principal can realize better capacity utilization in a poor market but reduce waiting times in a good market with the contingent linear contract.

One natural extension of our work is to consider an agent with several responsibilities. When a center’s agent also invests operational effort to reduce service time, the facility can process more patients without further investment in capacity. How the principal should compensate such a multi-tasking agent in an environment with constrained capacity is an interesting research question. One could also introduce information asymmetry into the model by allowing the agent to possess private information about the state of base demand.

References