SERVICE OUTSOURCING: CAPACITY, QUALITY AND CORRELATED COSTS

Z. Justin Ren* Fuqiang Zhang†

March 2009

Abstract

This paper studies how to design service outsourcing contracts to ensure fast, quality services from an independent service provider. The outsourcer does not have perfect information about either the service provider’s capacity cost (i.e., cost for providing fast service), or her quality cost (i.e., cost of achieving a high quality level). Moreover, the two unknown costs may be correlated with each other. We solve for the outsourcer’s optimal outsourcing contract, and show that the structure of the optimal contract depends on the relationship between the costs. Specifically, we highlight the following observations when the two costs are negatively correlated: First, under certain conditions, the outsourcer may be able to squeeze the supplier’s profit (information rent) to zero for an intermediate range of cost realizations; second, it is possible that the service supply chain is coordinated by using the outsourcer’s optimal contract. We then examine the performance of two classes of commonly observed contracts that are relatively simple to implement. It has been found that these simple contracts generally perform well when the costs are positively correlated, but they could perform much worse when the costs are negatively correlated. Our results therefore caution outsourcing companies that the potential trade-off between capacity cost and quality cost may require a careful design of outsourcing contracts.

Keywords: outsourcing, service, capacity, quality, information asymmetry, contracting, queueing systems

*Boston University School of Management. Email: ren@bu.edu
†Olin Business School, Washington University in St. Louis. Email: fzhang22@wustl.edu
1 Introduction

Business process outsourcing has been rapidly growing during the past decade. However, many companies are just beginning to discover the true cost of outsourcing, and suddenly find themselves in a disadvantaged position. Service providers are frequently located thousands of miles away, often in a foreign country, and it can be very difficult to obtain perfect information on every aspect of their service operations. After companies engage in an outsourcing transaction, they are often surprised to find out the true costs of their vendors (Kharif 2003).

Speed and quality are both critical in outsourced service. Poor service performance by the outside service supplier reflects directly on its outsourcing company. Slow service wastes customers’ time and may cause frustration and anxiety. Quality is defined as the probability that an incoming job will be successfully done. In revenue-generating customer contact centers (e.g., a phone-order service), a well-trained and motivated sales agent can answer customers’ inquiries to their satisfaction and increase the likelihood of a sale. In non-revenue-generating service areas (e.g., a technical support center), a knowledgeable agent can solve technical problems in a timely manner, which can reduce the likelihood of customer complaints and enhance the outsourcing company’s image. Indeed, according to a recent survey of more than 200 large companies, quality, price, and delivery performance have been ranked the top three criteria used in selecting service providers (CAPS Research 2006).

To ensure fast, quality service, providers need to be well-staffed and make quality-enhancing investments. Moreover, there is often a correlation between service providers’ costs related to capacity and quality. In some situations, the two costs may be positively correlated. For example, successful quality improvement initiatives such as Total Quality Management (TQM) may simultaneously bring down the capacity and quality costs of a firm. Or a better managed organization may enjoy the advantage of both lower capacity and quality expenditures. On the other hand, a trade-off (or negative correlation) between the two costs may exist in many other situations. In the call center industry, for instance, labor costs are typically lower for call centers located overseas, compared to those located in the US. However, the service quality for domestic call centers is often much higher. For example, Dell employs call centers in both North America and overseas, but it charges customers an extra fee ($12.95 a month, or $99 a year for people who buy a new computer) for access to its American call center agents. This is mainly because the costs for domestic call centers are typically 50% to 70% higher than those overseas (Whoriskey 2008). Such a negative correlation also exists among employees within a firm. For example, more experienced salespeople
can close deals with higher probabilities, or better trained customer representatives can achieve higher customer satisfaction, but they are generally more expensive to acquire and retain.

We study a service outsourcing problem that involves all the issues discussed above: information asymmetry, capacity and quality considerations, and correlated costs. Our model consists of two independent firms: an outsourcer (referred to as “he”) and a supplier (referred to as “she”, and also called the service provider). The outsourcer procures a certain service from the supplier. The supplier is free to choose her capacity and quality levels in serving the customers of the outsourcer. The supplier’s capacity cost (denoted by $b\mu$) and quality cost (denoted by $bq$) are both private, so the outsourcer does not have perfect knowledge of this information. Moreover, there is a certain relation between the two unknowns. Specifically, we model such a relationship as a linear one: $b_q = a + kb_\mu$. We shall call $k$ as the “strength of correlation”, which can be negative, zero, or positive. As a result of this interaction, the supplier may not always have an incentive to bias her true private information in one direction. In fact, when there is a negative correlation between the two costs, i.e., $k$ is within a negative interval, the supplier may wish to either overstate or understate the capacity cost (understating the capacity cost is essentially overstating the quality cost). Such a phenomenon is called “countervailing incentives”. To the best of our knowledge, the impact of countervailing incentives on outsourcing contract design has not been studied in the operations literature.

The main results in this paper can be summarized as follows. First, we solve the theoretically optimal contract for the outsourcer under information asymmetry and cost correlation. We show that the form of the optimal outsourcing contract depends critically on the relationship between capacity and quality costs. For positive $k$ values, the structure of the optimal contract is similar to those in principal-agent models involving private information on a single cost parameter. That is, the supplier’s profit (also known as information rent) is monotonically decreasing in the capacity cost realization, and the supplier’s capacity and quality levels are both lower than the supply chain optimal solution. However, when $k$ is negative, the supplier’s profit could be either increasing or decreasing in the capacity cost realization, and the capacity and quality levels could be either higher or lower than the supply chain optimal solution. Two specific observations from negative $k$ values are worth highlighting: First, under certain conditions, there may be an intermediate interval of the capacity cost realizations where the supplier’s profit is zero. In other words, the outsourcer is able to extract all the surplus from the supply chain for a certain range of cost realizations. Second, there may be a non-boundary cost realization at which the optimal contract can actually coordinate the supply chain despite information asymmetry.
Second, we study service outsourcing contracts from a practical point of view. In particular, we are interested in the performance of simple contracts, especially the ones that are not contingent on the supplier’s cost information. We survey the literature and find that the service outsourcing contracts in practice are in two main categories: those that contract directly on desired capacity and quality levels, and those that provide financial incentives or penalties to induce the service provider to allocate adequate capacity and maintain high quality. Both categories of contracts are sub-optimal, but are easy to implement and have been widely observed in practice. They do not require the supplier’s cost as an input, and the contract terms only involve linear functions. We hence label them “simple” contracts in this paper.

We find that such simple contracts perform quite well, as long as \( k \), the strength of correlation between capacity and quality costs, is non-negative. In this case, the cost increase for the outsourcer by using the simple contracts relative to the optimal contract is negligible. However, when the strength of correlation is negative, simple contracts might perform poorly from the outsourcer’s perspective. Therefore, countervailing incentives play an important role in the selection of outsourcing strategies. Finally, we show that between the two simple contracts incentive contracts are always better for the outsourcer than direct contracting on capacity and quality levels.

The organization of this paper is as follows. After a literature review (§2), we introduce our model setting and characterize the optimal solution under centralized management (§3). In §4, we present the optimal contract for the outsourcer, which is non-linear and contingent on the supplier’s cost. We then turn our attention to the simpler contracts that are linear and non-contingent on the supplier’s cost information (§5). §6 compares the simple contracts to the optimal contract using numerical studies. Finally, we conclude with §7. All proofs are in Appendix A.

## 2 Literature Review

Supply chain contract design under asymmetric information has been widely studied in the literature. Several studies focus on settings where a supplier offers a contract to a retailer, who has private cost information: Corbett and de Groote (2000), Corbett (2001), Ha (2001), and Corbett, Zhou and Tang (2004), to name a few. In these papers, the supplier designs a contract to induce the retailer to truthfully report its cost information, so they consider the so-called screening games. Cachon and Lariviere (2001) formulate a capacity procurement model, where the manufacturer has private information about its market size and wishes to convey this information credibly to the supplier, who has to build capacity before demand is realized. Their game falls into the category
of signalling games, i.e., one of the entities intends to reveal its private information through some signalling device. Özer and Wei (2007) study a supply chain model similar to that of Cachon and Lariviere (2001), except that they derive the optimal screening contract for the supplier. Our model falls into the screening model category. However, it differs from these existing papers because our setting involves multiple decisions (i.e., service speed and quality), as well as asymmetric information on correlated costs.

Several recent papers study procurement mechanism design by taking operational characteristics, in addition to price, into consideration. Cachon and Zhang (2006) consider a procurement problem, where the buyer prefers a lower procurement price and a speedier delivery. The supplier is modelled as an M/M/1 queue, and the buyer does not have perfect information about the supplier’s capacity cost. The authors identify the optimal procurement mechanism that minimizes the buyer’s total costs. Moreover, recognizing that the optimal mechanism is complex to implement, they show that several simple mechanisms, such as one with a fixed late fee and another with a pre-specified lead time requirement, perform very well, both for the buyer and for the supply chain. Zhang (2008) considers a procurement problem in a two-echelon inventory system with price-sensitive market demand, where the buyer takes both delivery performance (i.e., the service level attribute) and procurement quantity (i.e., the quantity attribute) into consideration. He finds that imposing a simple term on the supplier’s service level attribute works well for the buyer, but the same does not hold for the quantity attribute. Our paper also studies multi-attribute outsourcing contracts, but it differs from the above papers in several important ways: First, we consider quality in service outsourcing, which is absent in the above models. Second, we model correlated asymmetric information on both quality and capacity costs, and characterize the optimal mechanism under the presence of countervailing incentives. Finally, this paper generates different qualitative insight regarding simple contracts: We find that simple contracts do not always perform well, in particular in situations with negative correlation between capacity and quality costs.

An important feature of our model is capturing information asymmetry on two correlated costs. With negatively correlated costs, the agent may have incentives to either overstate or understate its private information. Lewis and Sappington (1989) study a principal-agent model where the agent’s variable production cost is negatively associated with its reservation profit (or fixed opportunity cost). Such situations may arise because a more efficient agent usually faces more profitable outside opportunities. It can be shown that the agent has incentives to overstate its variable cost for low realizations of variable cost, and to understate its variable cost (or overstate its fixed opportunity cost) for high realizations of variable cost. The authors thus call such a phenomenon
“countervailing incentives”. Maggi and Rodriguez-Clare (1995) further demonstrate that the form of countervailing incentives plays a crucial role in determining the structure of optimal contract and the distribution of the agent’s information rent. In both papers, the agent needs to make a single decision about the production quantity. Here we apply the principle-agent framework to study a service outsourcing setting where the agent needs to make two decisions (capacity and quality levels), and the countervailing incentives arise from the cost coefficients associated with the two decisions (rather than variable and fixed costs). Moreover, we consider both positive and negative correlations between the two unknown costs, and identify conditions under which those commonly observed and relatively simple contracts perform well.

We define service quality as the probability of a customer’s job being successfully done. For example, when incoming calls generate revenue, the resolution of a call may mean the conversion of a customer inquiry into a sale. When the calls are for customer service, the resolution of a call may mean that the customer is satisfied and will not call back for the same question, thus reducing future service cost. The importance of call resolution in providing the agents with the appropriate economic incentive has been recognized in Shumsky and Pinker (2003). In that paper, it is shown that paying the agents a flat wage, plus a volume-based fee and a resolution-based fee (“pay for solve”), provides an incentive for the agents to take the right actions. Equivalently, the agents are penalized for not satisfying a customer, which is similar to the notion of quality penalty studied in our paper. However, in our paper, there are two dimensions of service performance: job resolution and waiting. The outsourcing company, therefore, has an incentive to induce a high level of capacity allocation by the outsourced service provider, as well as a high job-resolution rate. Lim (2001) studies a quality-control problem where a producer purchases parts from a supplier but is unsure about the supplier’s inherent quality level. The author shows that the producer can design an optimal contract to induce the true quality level by the supplier. The quality studied in Lim (2001) relates to production technology, and therefore is different from the service quality we study. Moreover, our focus in this paper is not limited to identifying an optimal contract. We are also interested in quantifying the effectiveness of simple-to-implement contracts that are observed in practice.

There has not been much work on addressing the costs related to information asymmetry and service quality in the context of service outsourcing. Ren and Zhou (2008) consider the service quality issue in call center outsourcing. In their model, the call center has to decide on both its staffing and effort level to achieve a certain quality level. They study the effectiveness of some commonly used contracts, such as piece-rate and pay-per-call-resolved contracts. When effort is not
directly observable, they show that a partnership-type contract can coordinate this supply chain. However, they assume perfect information in their model, whereas in this paper the outsourcer does not have perfect information about the supplier’s costs. Hasija, Pinker, and Shumsky (2008) consider a call center outsourcing setting where the service provider’s service rate is unknown to the outsourcer. Motivated by a list of contracts that are being used at a large call center, they study how to design different types of screening contracts to reveal the true service rate, as well as conditions under which supply chain coordination is achieved.

Similar to Ren and Zhou (2008), our paper considers complete outsourcing, i.e., when all of the service work for a particular business process is being outsourced. Gans and Zhou (2007) and Akşin, de Vericourt, and Karaesmen (2008) allow the user company to outsource some, but not all, of the service work. In particular, Gans and Zhou (2007) focus on the queueing control and capacity planning aspects, while Akşin, de Vericourt, and Karaesmen (2008) suppress queueing details to focus on the higher-level contract design. They also allow service requirements (call volumes) to vary over time, and the key question for them becomes how many calls to outsource in each period, i.e., to “outsource the peak” or “outsource the base”. Milner and Olsen (2008) consider a call center that serves both contract customers and noncontract customers, where a service-level agreement is exogenously given for the contract customers. They study the call center’s priority-setting policy for the two types of customers, and investigate the implication of the service-level agreement on call center management. However, neither of these models considers service quality. For a more comprehensive review of the literature on call center management and outsourcing, please see Gans, Koole, and Mandelbaum (2003) and Aksin, Armony, and Mehrotra (2007).

3 Model

Consider a supply chain where an outsourcer (“he”) obtains a certain service from a supplier (“she”). Demand for the service arrives at the outsourcer according to an arrival process with an average rate $\lambda$. The supplier is modeled as a queueing system with a general structure. Customers are patient and there is no abandonment in the basic model. We extend the model to consider abandonment in Appendix B. For jobs that are not immediately served, a waiting cost $c_w$ is incurred for each job per time unit. Let $\mu$ be the supplier’s capacity decision variable that determines the average waiting time for the jobs. For instance, $\mu$ could be the service rate (i.e., $1/\mu$ is the mean service time) if there is only a single server; or it could be the number of servers if the staffing level is the decision to make (e.g., for a large call center). Let $W(\mu)$ be the average waiting time and
\( L(\mu) = \lambda W(\mu) \) the average number of jobs in the system for a chosen \( \mu \). Assume that \( L(\mu) \) is convex and decreasing in \( \mu \). This has been proven to be true for any single-server queue (Weber 1983). For multi-server queuing systems, we apply a commonly used approximation of GI/G/m queues (Whitt 1993; Hopp and Spearman 1996) and show that this assumption is also satisfied (in this case, \( \mu \) is the total capacity the supplier chooses; see Section 6.2). Further, Armony, Plambeck, and Seshadri (2005) show that such a convexity property holds even for queues with reneging and balking.

As in any real-world situation, not all jobs are served to customers’ satisfaction. Service quality is therefore measured by a satisfaction probability \( q \). For example, in the context of call center outsourcing, it is the percentage of calls satisfactorily served and resolved (Shumsky and Pinker 2003, Ren and Zhou 2008). For the \( 1 - q \) proportion of the jobs that are not satisfactorily served, a loss of goodwill cost \( c_q \) is incurred to the outsourcer. For simplicity, we assume that dissatisfied customers are lost and never return. The situation with customer retrials will be discussed in Appendix B.

The supplier may choose her service rate \( \mu \) and service quality \( q \), both of which are costly. In achieving \( \mu \) and \( q \), the supplier has to exert effort \( g(\mu) \) and \( h(q) \), respectively. Such effort may include hiring human-resource consultants to improve the recruiting process, providing productivity-enhancing facilities (e.g., a better work environment) and amenities, or purchasing equipment and training to improve servers’ service speed and effectiveness. It is natural to assume that both \( g(\mu) \) and \( h(q) \) are convexly increasing: The marginal effort needed for improving service is increasing. The function forms of \( g(\cdot) \) and \( h(\cdot) \) are common knowledge, as they can be obtained by market intelligence.

In exerting effort, the supplier incurs a cost of \( b_\mu g(\mu) + b_q h(q) \), where \( b_\mu \) and \( b_q \) are the cost coefficients. A higher cost coefficient \( b_i \) \((i = \mu, q)\) indicates a more costly supplier; \( b_i \) can be interpreted as the cost of serving a customer, which the service provider may know better than the outsourcer company. More importantly, there can be a positive or negative correlation between the two costs. For example, in many environments there is a negative correlation between capacity cost \( b_\mu \) and quality cost \( b_q \). Such a relationship reflects the fact that more expensive capacity has a higher effectiveness in achieving a given level of quality. It includes better capital equipment that generates less waste or errors, or better educated workers who can produce a higher yield. On the other hand, a positive correlation is also permissible. Imagine that as a firm improves its overall capability through initiatives such as TQM, it may spend less on recruiting or other productivity-enhancing activities and still be able to achieve better quality and decrease its quality-related
costs. To maintain tractability and for ease of exposition, we model such a relationship using a linear function: \( b_q = v(b_\mu) = a + kb_\mu \), where \( k \) could be positive, negative, or zero. It captures the three exclusive types of correlation structures between the two costs.\(^1\) For notational convenience, in the rest of the paper, we will omit the subscript \( \mu \) in \( b_\mu \) and use \( v(b) = a + kb \) for \( b_q \). We assume that the values of \( a \) and \( k \) are common knowledge. Information on different capacity and quality costs is generally obtainable by market research, and the value of \( a \) and \( k \) can thus be estimated with a linear regression.

Finally, the outsourcer does not have perfect information about the supplier’s cost realization. In particular, the cost coefficient \( b \) is private information, and the outsourcer only has an unbiased belief about its distribution. Let \([b, \bar{b}]\) (\( 0 < b \leq \bar{b} < \infty \)) be the support of \( b \) and let \( F \) and \( f \) be the distribution and density functions, respectively. We assume that \( F \) is log-concave (i.e., \( F(x)/f(x) \) is increasing), and has an increasing failure rate (i.e., \( f(x)/(1-F(x)) \) is increasing). This assumption is satisfied by many commonly used distributions (e.g., normal, uniform, exponential, gamma, etc.) and is standard in the mechanism design literature. See Lewis and Sappington (1989) and Bagnoli and Bergstrom (2005) for more details.

Both the outsourcer and the supplier are risk neutral and aim to minimize/maximize their expected cost/profit. It is clear that an incentive conflict exists in the supply chain: The outsourcer prefers fast and high-quality service while the supplier has to incur the cost. The outsourcer’s challenge is to design an outsourcing contract to induce the supplier to provide desirable services. The sequence of events in our model is as follows: The supplier’s cost is realized and observed only by the supplier; then the outsourcer offers a take-it-or-leave-it contract to the supplier; after the contract is taken, the supplier takes actions to serve customers according to the contract terms; lastly, both firms’ payoffs are evaluated.

### 3.1 Centralized Management

We first consider a base case: The two firms are integrated or managed by one central entity. The outsourcer incurs the following costs in the system:

\[
c_w L(\mu) + c_g \lambda(1 - q).
\]

\(^1\)Under this linear relationship, the coefficient of correlation could be 1, 0, or −1. An alternative is to model \( b_n \) and \( b_q \) as two correlated random variables (e.g., they follow a bivariate normal distribution). However, a general two-dimensional screening model brings significant challenge for analysis. So far there have been limited results in the literature on optimal mechanism design involving multiple unobservables. See Armstrong and Rochet (1999) for a representative study.
The first term is the expected waiting cost, and the second term is the goodwill loss from dissatisfied customers. The supplier incurs the capacity and quality costs:

\[ bg(\mu) + (a + kb)h(q). \]

The total cost function for the integrated supply chain is (we use subscript \( c \) for centralized supply chain):

\[ C_c = c_wL(\mu) + c_g\lambda(1 - q) + bg(\mu) + (a + kb)h(q). \]  

**Proposition 1** \( C_c \) is jointly convex and the optimal capacity \( \mu^* \) and quality \( q^* \) for the centralized supply chain are given by

\[
\begin{align*}
    c_wL'(\mu^*) + bg'(\mu^*) &= 0, \\
    (a + kb)h'(q^*) - c_g\lambda &= 0.
\end{align*}
\]

It can be shown that in the integrated supply chain, the optimal service rate \( \mu^* \) is increasing in the arrival rate \( \lambda \) and waiting cost \( c_w \), but decreasing in capacity cost \( b \). The optimal quality level \( q^* \) is increasing in capacity cost \( b \) when \( k < 0 \), and is decreasing in \( b \) when \( k > 0 \). \( q^* \) is also increasing in the arrival rate \( \lambda \) and the loss of goodwill \( c_g \). Throughout the paper, we use increasing/decreasing in the weak sense, e.g., increasing means non-decreasing.

For later use, let

\[ C_c(b) = c_wL(\mu^*(b)) + c_g\lambda(1 - q^*(b)) + bg(\mu^*(b)) + (a + kb)h(q^*(b)) \]

be the supply chain’s optimal total cost when the supplier’s cost is \( b \).

## 4 Optimal Contingent Contracts

In this section, we consider the case of decentralized management. That is, the outsourcer and the service providers are managed separately, each trying to optimize its own expected payoff. In this situation, the outsourcer does not know the exact value of the service provider’s cost \( b \). We focus on deriving the outsourcer’s optimal contingent contract that uses the cost \( b \) as an input in contract terms. Although contingent contracts require a more complex format, they are clearly superior to contracts that are independent of the supplier’s cost realization (i.e., the non-contingent contracts form a subset of the contingent contracts).
The optimal contingent contract minimizes the outsourcer’s expected total cost. It is also called the optimal mechanism (OM). A mechanism is essentially a mapping from the supplier’s cost space to the space of a transfer payment and an action schedule. According to the Revelation Principle (Myerson, 1981 and 1983), there is a *direct* (i.e., the supplier only reports her private cost) and *truthful* (i.e., the supplier truthfully reports her cost) mechanism that is optimal. Specifically, we only need to search among mechanisms that satisfy two constraints: the incentive compatibility (IC) constraint (i.e., the supplier announces the true cost) and the individual rationality (IR) constraint (i.e., it is in the supplier’s best interest to take the contract).

Consider a menu of contracts \{\mu(x), q(x), T(x)\}: The supplier chooses from this menu by announcing her cost to be \(x\); then she needs to provide a service rate \(\mu(x)\) and a quality level \(q(x)\); and finally, the outsourcer pays her \(T(x)\) per customer served.\(^2\) Under this menu of contracts, both the transfer payment and action schedule are functions of the supplier’s announced cost \(x\). Let \(\pi_s\) denote the supplier’s expected profit (the subscript \(s\) refers to the supplier). The IC constraint implies

\[
\begin{align*}
  b = \arg \max_x \pi_s(x, b) &= T(x)\lambda - [bg(\mu(x)) + (a + kb)h(q(x))]
\end{align*}
\]

where \(b\) is the supplier’s true cost and \(x\) is the announced cost. The IR constraint is given by

\[
\begin{align*}
  \pi_s(b) \geq 0, \text{ for all } b \in [\underline{b}, \bar{b}].
\end{align*}
\]

That is, the supplier participates only if her profit is non-negative. An underlying assumption here is that the supplier’s opportunity cost (or the best profit the supplier can get from outside options) is zero. Introducing a strictly positive opportunity cost for the supplier does not change the analysis and qualitative insight. It is worth pointing out that according to (6), even the least efficient (or the highest cost) supplier accepts the contract, which implies that the outsourcer is willing to transact with any supplier types.\(^3\)

\(^2\)Note that \(T(\cdot)\) is the unit price for each customer served by the supplier. Since the demand rate \(\lambda\) is fixed in our model setting, specifying the unit price is equivalent to specifying the total payment. Another alternative is to pay for each customer satisfied by the supplier. The difference between these two payment schemes is a factor of \(q\) (the probability a customer is satisfied). It is straightforward to show that the optimal menu of contracts with pay-per-customer-satisfied is equivalent to the optimal menu of contracts with pay-per-customer-served and yields the same expected total cost for the outsourcer. The analysis is similar to the proof of Proposition 2 and therefore omitted.

\(^3\)Such a condition may not hold if it is not profitable for the outsourcer to deal with certain types of suppliers: For example, the revenue generated from customers may not be sufficient to cover the supplier’s cost. The outsourcer may choose to exclude some supplier types from the transaction in this case. See Ha (2001) and Corbett, Zhou and Tang (2004) for discussion of such contracts.
The outsourcer’s total cost consists of the transfer payment $T\lambda$ and the operational costs:

$$C_o(\mu, q) = T\lambda + c_wL(\mu) + c_g\lambda(1 - q).$$

The subscript $o$ stands for the outsourcer. The optimal menu of contracts is given by solving the following dynamic optimization problem:

$$\min_{\{\mu(\cdot), q(\cdot), T(\cdot)\}} \int_0^{\bar{b}} \left[ T(x)\lambda + c_wL(\mu(x)) + c_g\lambda(1 - q(x)) \right] f(x) dx$$

s.t. (5), (6)

We introduce some notations before presenting the solution to (7). Let $\eta_1(x) = F(x)$ and $\eta_2(x) = F(x) - 1$. Define $\{\mu_i(\cdot), q_i(\cdot)\}$ to be the solutions to the following simultaneous equations ($i = 1, 2$):

$$c_wL'(\mu) + [x + \eta_i(x)/f(x)]g'(\mu) = 0,$$

$$c_g\lambda - [(a + kx) + k\eta_i(x)/f(x)]h'(q) = 0.$$ (8), (9)

The following proposition presents the solution of the optimal mechanism.

**Proposition 2** In the optimal menu of contracts, there exist $b_1$ and $b_2$ ($b \leq b_1 \leq b_2 \leq \bar{b}$) such that the optimal capacity and quality functions $\{\mu^o(x), q^o(x)\}$ are given by:

$$\begin{cases} 
\mu^o(x) = \mu_1(x), & q^o(x) = q_1(x), & x \leq b_1, \\
\mu^o(x) = \mu_1(b_1), & q^o(x) = q_1(b_1), & b_1 \leq x \leq b_2, \\
\mu^o(x) = \mu_2(x), & q^o(x) = q_2(x), & x \geq b_2.
\end{cases}\)$$

The payment function $T^o(x)$ can be solved using (5) and (6).

The above optimal menu of contracts is quite complex and deserves some discussion. First we explain the two critical values $b_1$ and $b_2$, which determine the structure of the supplier’s profit (or information rent). In a traditional, single-dimensional screening problem, we know that the supplier’s profit decreases as her cost realization increases, and the least efficient supplier type receives a zero profit. However, this is not always true in our model setting. There are three possible cases in the above proposition depending on the values of $b_1$ and $b_2$ (these values are characterized in the proof of Proposition 2 in Appendix A):

**Case (1)** ($b_1 = b_2 = \bar{b}$): In this case, the supplier’s profit decreases in the capacity cost $x$, just as in a single-dimensional screening problem.

**Case (2)** ($b_1 = b_2 = b$): In this case, the supplier’s profit increases in the capacity cost $x$, which is the opposite of Case (1).
Case (3) \((\bar{b} < b_1 < \bar{b} \text{ or } \bar{b} < b_2 < \bar{b})\): In this case, the supplier’s profit decreases in the capacity cost for \(x \leq b_1\) (and reaches zero at \(x = b_1\)), remains zero on the interval \([b_1, b_2]\), and then increases for \(x \geq b_2\). In other words, there exists an interval \([b_1, b_2]\) where the supplier’s information rent is squeezed to zero (the interval collapses to a single point if \(b_1 = b_2\)).

The above observation is interesting because the supplier’s profit in the optimal menu of contracts can be either decreasing, increasing, or constant depending on the problem parameters. Now a natural question arises: When do the above three different cases happen? Intuitively, Case (1) would happen if the capacity cost is the dominant cost factor, while Case (2) might happen when the quality cost has a stronger effect. The following proposition confirms this intuition.

**Proposition 3** There exist \(k_2 < k_1 < 0\) such that the above Cases (1), (2), and (3) will happen if and only if \(k \geq k_1\), \(k \leq k_2\), and \(k_2 < k < k_1\), respectively.

This proposition further highlights how the optimal menu of contracts depends on the strength of correlation. Case (1) is when \(k\) is above some threshold value \(k_1\): As capacity cost \(x\) increases, the supplier’s quality cost either increases or decreases relatively slowly, so the supplier’s overall efficiency level decreases in \(x\). In this case, the supplier’s profit monotonically decreases in the realization of cost \(x\), just as in a single-dimensional screening problem. Case (2) is the opposite in the sense that the supplier’s profit increases in cost \(x\). This is mainly because of the strong negative correlation between the capacity and quality costs: As capacity cost \(x\) increases, her quality cost decreases at a relatively fast rate, so the supplier’s overall efficiency level increases in \(x\). In this case, the quality cost dominates, and a higher capacity cost implies a lower quality cost and thus higher total profit.

When the correlation strength \(k\) is in the intermediate range, as represented in Case (3), the supplier’s profit structure is not monotone any more, and it depends on her private cost \(x\). For low realizations of \(x\), the supplier’s profit is largely driven by her capacity cost, and its behavior is similar to that in Case (1). Her profit decreases in \(x\) in this interval. For high realizations of \(x\), her quality cost dominates, and her behavior is similar to that in Case (2), where her profit increases in cost \(x\). Interestingly, when her cost \(x\) is in its intermediate range, there may be an interval where the profit is constant zero. It means that under certain cost realizations, the outsourcer is able to extract all the information rent due to the fact that countervailing incentives are acting against each other to balance out the supplier’s information advantage. As a result, the supplier might be better off with an extreme value of the capacity cost. Figure 1 below illustrates the supplier’s profit function under the optimal contingent contracts in a numerical example with \(k < 0\). Note
that the U-shaped profit curve in Figure 1 is reduced to L-shaped if \( b_2 = \bar{b} \) (or reversed-L-shaped if \( b_1 = \bar{b} \)).

Figure 1. Supplier profit as a function of cost realizations in the optimal menu of contracts: an example.

Next we examine the optimal capacity function \( \mu^o(x) \) and the optimal quality function \( q^o(x) \) that the outsourcer should offer to the supplier. The next proposition shows that the shape of these two functions depends on the correlation strength \( k \):

**Proposition 4** *In the optimal menu of contracts, if \( k \geq 0 \), then both the optimal capacity \( \mu^o(x) \) and the optimal quality \( q^o(x) \) are decreasing in \( x \). On the other hand, if \( k < 0 \), then \( \mu^o(x) \) is decreasing, while \( q^o(x) \) is increasing, in \( x \).*

For illustration, Figure 2 plots the optimal \( \mu^o(x) \) and \( q^o(x) \) functions in the same example used by Figure 1. For comparison, the supply chain optimal capacity \( \mu^*(x) \) and quality \( q^*(x) \) are also shown.

Figure 2. Capacity \( \mu^o(x) \) and quality \( q^o(x) \) in the optimal menu of contracts: an example.
We can see that the two functions $\mu^o(x)$ and $q^o(x)$ could be flat for a subinterval, as we have noted above in Proposition 2.

Finally, by comparing the optimal menu of contracts to the supply chain’s optimal solution, we can check whether the supply chain is coordinated. We find that, in general, the outsourcer distorts the supply chain optimal solution in order to maximize his own objective due to the presence of asymmetric information. However, supply chain coordination might be achieved by the optimal menu of contracts for certain cost realization when Case (3) above occurs. The following proposition formalizes this result.

**Proposition 5** The following table summarizes the comparison between the supply chain optimal capacity and quality levels and those under the optimal menu of contracts (the values of $k_1$, $k_2$, and $k_3$ are given in Proposition 3):

<table>
<thead>
<tr>
<th>Case</th>
<th>Capacity</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Case (1)</strong></td>
<td>$\mu^o(x) &lt; \mu^*(x)$</td>
<td>$q^o(x) &lt; q^*(x)$ if $k &gt; 0$</td>
</tr>
<tr>
<td>$(k \geq k_1)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Case (2)</strong></td>
<td>$\mu^o(x) &gt; \mu^*(x)$</td>
<td>$q^o(x) &gt; q^*(x)$ if $k &lt; 0$</td>
</tr>
<tr>
<td>$(k \leq k_2)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Case (3)</strong></td>
<td>$\mu^o(x) &lt; \mu^<em>(x)$ for $x &lt; \hat{b}$; $\mu^o(x) = \mu^</em>(x)$ for $x = \hat{b}$; $\mu^o(x) &gt; \mu^*(x)$ for $x &gt; \hat{b}$.</td>
<td>$q^o(x) &gt; q^<em>(x)$ for $x &lt; \hat{b}$; $q^o(x) = q^</em>(x)$ for $x = \hat{b}$; $q^o(x) &lt; q^*(x)$ for $x &gt; \hat{b}$.</td>
</tr>
</tbody>
</table>

There exists a $\hat{b} \in (b_1, b_2)$: There exists a $\hat{b} \in (b_1, b_2)$:

The example in Figure 2 represents Case (3) in Proposition 5, where $\hat{b}$ represents the intersection of $\mu^o(x)$ and $\mu^*(x)$, as well as $q^o(x)$ and $q^*(x)$ (indeed we prove that both intersect at the same points). The traditional wisdom in supply chain contracting (with single-dimensional screening) is that the principal’s optimal menu of contracts never coordinates the supply chain except for the least efficient agent type. However, as we have shown in Proposition 5, this is no longer true when there are countervailing incentives: For the supplier with cost $\hat{b}$, the principal’s optimal menu of contracts could yield the supply chain optimal solution. Moreover, for suppliers with costs that are in the neighborhood of $\hat{b}$, the optimal menu of contracts generates capacity and quality levels that are close to system-optimal.
5 Non-contingent Linear Contracts

The optimal menu of contracts derived in the previous section require the supplier’s private cost information as an input. In particular, the outsourcer induces the supplier to truthfully report her cost value. Although theoretically appealing, truth telling could be difficult to achieve in reality: Firms are unlikely to reveal their cost structure for all sorts of reasons that are not present in our model (e.g., competition, long-term consideration, etc.) The optimal mechanism is also complex because it involves several nonlinear functions. These shortcomings may severely compromise the practical value of the optimal mechanism. The purpose of this section is to explore the effectiveness of commonly observed simple contracts in service outsourcing, particularly those involving linear functions that are not contingent upon the supplier’s cost structure.

We have surveyed service outsourcing contracts implemented in practice as well as in related literature and identified two broad categories of practical contracts. The first category is based on what we call a direct contracting approach. Contracts in this category articulate and mandate the outsourcer’s desired capacity and quality level. It is a natural way to simplify the optimal menu of contracts by specifying the capacity and quality levels the supplier should provide regardless of her cost realization. Indeed, many companies adopt this form of contracts in outsourcing services (Lacity and Willcocks 1998). Such contracts are straightforward to specify, easy to communicate, and relatively simple to implement because they make clear the outsourcer’s expectation and the service provider’s responsibility in both capacity and quality dimensions. This is probably the most widely used form of contracts in service outsourcing, despite the fact that it may have many variants in real-world applications. For example, in call center outsourcing, various service-level agreement (SLA) terms such as “80% calls have waiting time less than 120 seconds” are commonly used (Hasija, Pinker and Shumsky 2006, and Milner and Olsen 2008). They are in effect directly contracting on service providers’ capacity level. This is because in accepting such a service-level agreement, a service provider essentially agrees to commit to a certain level of capacity to achieve the corresponding service level. In this section, we first study such a category of contracts. We call them Specified Capacity and Quality (SCQ) contracts.

The second category is based on what we call an incentive contracting approach. For example, in call center outsourcing, it is not uncommon for the outsourcer to pay the service provider by each call satisfactorily resolved, as well as to impose a waiting penalty if customers encounter excessive delay while waiting for service. Such an incentive approach allows the supplier to choose her own actions. This may be desirable because it is the supplier who possesses private information. Indeed,
many companies are starting to use this incentive approach. For example, Bank of America charges its service providers penalties for under-performing in key metrics such as customer satisfaction rate (Ellram and Tate 2006). Hence we study this class of contracts with the following terms: There is a quality penalty for each unsatisfied customer, plus a waiting penalty per unit of time for customer waiting. Call this a Quality and Waiting Penalty (QWP) contract. Clearly, the first term is to induce a high level of quality, while the second term is to incentivize the service provider to obtain enough capacity.

The two contracts we study here (i.e., the SCQ and QWP contracts) are stylized versions of what we observe in practice. However, these two contracts capture the two fundamentally different approaches in service outsourcing contracts. Most of the real-world contracts fall into either of the two categories, or represent a mixture of the two. Note that they both are independent of the supplier’s cost realization, i.e., the outsourcer can design the outsourcing contracts without any input from the supplier’s private cost information. In addition, they only involve linear functions. Therefore, they are much easier to implement in practice compared to the optimal menu of contracts identified in the previous section. Hence, we call them “simple” contracts. We first explore how the outsourcer should utilize these two types of contracts in achieving a high capacity and quality level in outsourcing. Then we compare the performances of these two simple contracts to the optimal menu of contracts in Section 6.

5.1 Direct Contracts: Specified Capacity and Quality (SCQ)

Consider the following direct contract: The outsourcer states the minimum capacity $\mu^s$ and quality $q^s$ for the supplier (we use superscript $s$ to denote the SCQ contract):

$$\mu(x) \geq \mu^s, \quad q(x) \geq q^s,$$

for all $x$,

and as compensation, the outsourcer pays the supplier a unit price $T$ to ensure supplier participation. It is clear that the supplier will choose $\mu = \mu^s$ and $q = q^s$. Suppose $b^s \in [b, \bar{b}]$ maximizes the supplier’s effort cost to achieve a given pair of $(\mu^s, q^s)$, i.e.,

$$b^s = \arg \max_b [bg(\mu^s) + (a + kb)h(q^s)].$$

Then there must be

$$T\lambda = b^s g(\mu^s) + (a + kb^s)h(q^s),$$
which guarantees that all supplier types will participate and make a non-negative profit. In this SCQ contract, the outsourcer’s expected total cost will be

\[ C_o(\mu^s, q^s) = c_w L(\mu^s) + c_g \lambda (1 - q^s) + b^s g(\mu^s) + (a + kb^s) h(q^s). \]  

(10)

The outsourcer’s problem is to identify the optimal pair \((\mu^s, q^s)\) that minimizes \(C_o\). Note that because of the non-monotonicity of the supplier’s profit function, the outsourcer’s optimization problem consists of two stages: We first need to search for the least efficient \(b^s\) for a given \((\mu^s, q^s)\), and then examine all possible pairs of \((\mu^s, q^s)\) that minimizes the outsourcer’s cost, i.e., we have a min-max problem. The following result shows that the outsourcer’s optimal \((\mu^s, q^s)\) in the SCQ contract can be found by simply analyzing an equivalent max-min problem from the centralized supply chain. Let \(b^m\) to be the cost that maximizes the supply chain’s optimal total cost, i.e., \(b^m = \arg \max_b C_c(b)\) (see equation 4 for the definition of \(C_c(b)\)). Let \(\mu^m = \mu^*(b^m)\) and \(q^m = q^*(b^m)\), and recall from Proposition 1 that \((\mu^*(b^m), q^*(b^m))\) are the cost-minimizers for the supply chain under \(b^m\). Under this definition, \(C_c(b^m)\) is the supply chain’s highest, \textit{ex post} optimal total cost. Note \(b^m\) may not be the highest cost \(b\) in general because \(k\) may be negative.

**Proposition 6** The outsourcer’s optimal cost in the SCQ contract is equal to \(C_c(b^m)\). In addition, if \((\mu^m, q^m)\) is unique, then \(\mu^s = \mu^m\) and \(q^s = q^m\) are optimal for the outsourcer in the SCQ contract.

The insight from the above proposition is that the outsourcer’s optimal cost in the SCQ contract is equal to the supply chain’s optimal total cost under \(b^m = \arg \max_b C_c(b)\). Since the centralized supply chain is relatively easy to analyze, we can also derive the outsourcer’s optimal SCQ contract easily. In implementing the SCQ contract, the outsourcer needs to be able to verify the actual capacity and quality levels provided by the service provider. Although modern information technology has made this readily achievable, in some business situations such information may still not be perfectly observable. This is especially true when the service process is not deterministic. For example, a well-staffed service provider may provide slow service due to the randomness in service requirements from the customers. Therefore, companies need to take this into consideration when designing outsourcing contracts. As an alternative, the outsourcer may use incentives (and penalties) and let the supplier choose her preferred capacity and quality levels under the set of incentives without having to monitor her decisions. We study this class of contracts next.
5.2 Incentive Contracts: Quality and Waiting Penalty (QWP)

Consider the following incentive contract: The outsourcer charges the supplier a quality penalty $K$ for each unsatisfied job and a waiting penalty $w$ per unit of time a job spent in the system. In addition, a price $T$ is paid to the supplier for each incoming job. We call this a Quality and Waiting Penalty (QWP) contract. The supplier’s problem under the contract $\{w, K, T\}$ is

$$\max_{\mu, q} \pi_s = T\lambda - [bg(\mu) + (a + kb)h(q)] - K(1 - q)\lambda - wL(\mu).$$

The supplier’s optimal choices $(\mu^p, q^p)$ are given by (we use superscript $p$ to denote the QWP contract):

$$bg'(\mu^p) + wL'(\mu^p) = 0,$$

$$K\lambda - (a + kb)h'(q^p) = 0.$$

Let $b^p = \arg\max_b [bg(\mu^p(b)) + (a + kb)h(q^p(b))]$, i.e., the supplier’s total cost reaches the highest at cost $b^p$. Since the outsourcer wants to minimize his total cost, $T$ is determined such that the least efficient supplier will just break even in the QWP contract:

$$T\lambda - [b^pg(\mu^p(b^p)) + (a + kb^p)h(q^p(b^p))] - K(1 - q^p(b^p))\lambda - wL(\mu^p(b^p)) = 0$$

or

$$T\lambda = [b^pg(\mu^p(b^p)) + (a + kb^p)h(q^p(b^p))] + K(1 - q^p(b^p))\lambda + wL(\mu^p(b^p)).$$

Thus, the outsourcer’s problem in this contract is as follows:

$$\min_{w,T} C_o(w, T) = E_b [(c_w - w)L(\mu^p) + cg(1 - q^p) + K\lambda (1 - q^p) - T\lambda],$$

where $\mu^p$, $q^p$ and $T$ are three functions dependent on $b$, $w$ and $K$.

There is no closed-form solution for $w$ and $K$ without assuming any specific structures on $L(\cdot)$, $g(\cdot)$, and $h(\cdot)$. These optimal contract parameters can be derived using a numerical search, although it is less straightforward than the analysis of the SCQ contract (where we only need to analyze the centralized supply chain). However, we can prove the following proposition, which shows that the QWP contract always performs better than the SCQ contract from the outsourcer’s standpoint.

**Proposition 7** The outsourcer’s expected total cost in the QWP contract is lower than that in the SCQ contract.
This result implies that the outsourcer is always better off by offering a carefully designed incentive contract than a straightforward SCQ contract. Of course, as we see from above, there is some trade-off in choosing the QWP contract: It requires (slightly) more computational efforts in calculating the optimal penalty terms. The SCQ contract, on the other hand, just requires the outsourcer to directly specify his desired capacity and quality levels, which are relatively straightforward to implement. However, as we have pointed out, there are business situations where the actual capacity and quality levels cannot be perfectly observable or verifiable (Anand and Aron 2008), especially due to the randomness in service systems. In such situations, the incentive approach may be more suitable for the outsourcer.

There may be variations of the penalty terms in the above QWP contract. For example, some call center outsourcing contracts impose a fixed penalty if the time a customer spends in the system exceeds a target level, while some other contracts charge waiting penalty only for additional time that a customer waits beyond a threshold value. It can be shown that although these terms appear to be different, they are essentially the same from the outsourcer’s point of view: They all provide incentives for the supplier to choose the right actions.

6 Comparison of Contracts

This section compares the contracts described in the previous sections. We are interested in the performances of the widely used simple contracts against that of the optimal menu of contracts. Section 6.1 studies M/M/1 queueing systems and Section 6.2 considers GI/G/m queueing systems. Explanations and intuition have also been provided to help understand the findings.

6.1 M/M/1 System

For tractability, we first use a stylized case of M/M/1 queue and solve in closed-form the various contracts we have studied so far. In an M/M/1 queuing system, we have \( L(\mu) = \frac{\lambda}{\mu - \lambda} \). Also consider the following specific function forms: \( g(\mu) = \alpha \mu \) (\( \alpha > 0 \)) and \( h(q) = \frac{1}{\gamma} \ln \left( \frac{\beta}{1 - q} \right) \) (\( 0 < \beta \leq 1, \gamma > 0 \)). The inverse of this \( h(q) \) function is: \( q(h) = 1 - \frac{\beta}{e^{\gamma h}} \). It is meant to capture the following properties:

1. \( q(h) \) is bounded by a non-negative constant \( q(0) = 1 - \beta \) from below, and by 1 from above,
2. \( q(h) \) is increasing in \( h \) and (3) there are diminishing returns to effort (i.e., \( q(h) \) is concave).

The specific function form we use possesses all of the above properties and is tractable. We may interpret \( 1 - \beta \) as the lower bound of the supplier’s initial quality level (i.e., \( q \geq 1 - \beta \)), and \( \gamma \) as a measure of the effectiveness of effort. Appendix B considers another set of plausible effort
functions and obtains similar results.

First, for the system-optimal solution, we apply the results from Proposition 1 and we have the following:

**Corollary 1** With an M/M/1 system and the above functional forms, the optimal capacity and quality for the centralized supply chain are given by

\[
\mu^*(x) = \lambda + \sqrt{\frac{\lambda c_w}{\alpha x}} \\
q^*(x) = \max\left(1 - \beta, 1 - \frac{a + kx}{\lambda c_g \gamma}\right).
\]

Given that the outsourcer does not have perfect information of \(b\), he can offer to the supplier the optimal menu of contracts, which we identified in Section 4.

**Corollary 2** With an M/M/1 system and the above functional forms, the outsourcer’s optimal menu of contracts is characterized by (see also Proposition 2):

\[
\mu_i(x) = \lambda + \sqrt{\frac{\lambda c_w}{\alpha (x + \eta_i(x)/f(x))}} \\
q_i(x) = \max\left(1 - \beta, 1 - \frac{a + kx + k\eta_i(x)/f(x)}{\lambda c_g \gamma}\right), \quad i = 1, 2
\]

where \(\eta_1(x) = F(x)\) and \(\eta_2(x) = F(x) - 1\).

Comparing the above solutions with those of the system-optimal, we can see the effects of information asymmetry on the supplier’s actions: They appear as an extra term of \(\eta_i(x)/f(x)\) for capacity, and \(k\eta_i(x)/f(x)\) for quality. Note also that the strength of correlation \(k\) has no direct impact on the supplier’s capacity choice, but plays a significant role in her quality choice.

Next, we can also solve the SCQ contract in closed-form. Applying the results from Proposition 6, we immediately have the following (recall \(b^m\) is the cost realization that maximizes the optimal total cost of the supply chain as defined in equation 4):

**Corollary 3** With an M/M/1 system and the above functional forms, the capacity and quality observed under the SCQ contract is:

\[
\mu^* = \lambda + \sqrt{\frac{\lambda c_w}{\alpha b^m}}, \quad q^* = \max\left(1 - \beta, 1 - \frac{b^m}{\lambda c_g \gamma}\right).
\]

There is no closed-form solution for the optimal QWP contract, although from Proposition 7 we know that QWP is more efficient than SCQ from the outsourcer’s point of view. Later we evaluate the performance of this contract numerically.
Recall that the outsourcer’s objective is to minimize his expected total cost by designing an outsourcing mechanism. Next we compare the simple contracts (SCQ and QWP) with the optimal menu of contracts (OM) for the outsourcer using a numerical study. The design of the numerical study is as follows. For the correlated costs, we adopt a function \( v(b) = \bar{b} + kb \) and test the \( k \) values taken from \(-0.9\) to \(1\) with a step of \(0.1\) (there are 20 values for \(k\) in total). The coefficient \(\alpha\) is normalized to be 1. The cost parameters \(c_w\) and \(c_g\) take values from \(\{0.1, 1, 10\}\) and \(\{0.1, 1, 10, 100\}\), respectively. Thus, the ratio \(c_w/c_g\) covers a wide range of values. Demand rate \(\lambda\) ∈ \(\{0.1, 10, 100\}\).

We assume \(b\) follows a uniform distribution with \(\bar{b} = \theta(1-\delta)\) and \(\bar{b} = \theta(1+\delta)\), where \(\theta\) ∈ \(\{0.1, 5, 30\}\) and \(\delta\) ∈ \(\{0.4\}\). That is, \(\theta\) is the mean of the distribution and \(\delta\) measures the variation in the cost. Note that \(\delta = 0.4\) represents a significant uncertainty in the supplier’s cost: \(\frac{\bar{b}}{\hat{b}} = 2.33\), i.e., a 133\% difference between the most and the least efficient supplier types. It is clear that the performance of the simple contracts improves as \(\delta\) decreases, so we do not consider \(\delta\) values less than 0.4. Finally, \(\beta\) is chosen from \(\{0.1, 0.5, 0.9\}\) and \(\gamma\) from \(\{0.01, 0.1, 1, 10\}\). There are 1296 scenarios (or parameter combinations) for each \(k\) value in this numerical study.

In each scenario we evaluate the performance of the QWP and SCQ contracts, measured by the percentage cost increase relative to the optimal menu of contracts (OM). Figure 3 summarizes the performance of each contract by presenting the maximum and average percentage cost increases. We highlight the following two observations based on the graph: First, both the QWP and SCQ contracts are close to optimal when \(k\) is not very negative (e.g., for \(k = -0.4\), the maximum percentage cost increase is 2.25\% and 2.52\% for the QWP and SCQ contract, respectively); however,
the performance of the two simple contracts deteriorates as $k$ decreases (e.g., for $k = -0.8$, the maximum percentage cost increase is 16.10% and 16.12% for QWP and SCQ, respectively). Thus the simple contracts tend to be more effective when the capacity and quality costs are not strongly negatively correlated. Second, the performance of the SCQ contract is very close, but not as good as that of the QWP contract, which is consistent with Proposition 7. In this numerical study, the resulting supplier’s average utilization (i.e., $\rho = \lambda/\mu$) in the OM ranges from 0.0246 to 0.9949, the quality level ranges from 0.1 to 0.999999, and the ratio of the outsourcer’s quality-related cost to the total cost ranges from 0.01% to 99%. So the numerical study is quite comprehensive, and the findings apply to a wide range of situations.

Why is the performance of the simple contracts much worse when $k$ is strongly negative? To further understand the behavior of the simple contracts, we compare the supplier’s profit ex post under different cost realizations, $b$. Note that the supplier’s profit serves as a good indicator of the performance of the simple contracts, i.e., a large amount of information rent means a higher cost for the outsourcer. Figure 4 plots the supplier’s profit in the OM and the SCQ contract as a function of the $b$ for the a specific example: $\theta = 0.1, \delta = 0.4, \alpha = 1, \gamma = 0.01, c_w = 0.1, c_g = 0.1, \lambda = 0.1$, and $\beta = 0.1$. The supplier’s profit in the QWP contract is very close to the profit in the SCQ contract; therefore, we omit the QWP curve in the graph.

![Figure 4](image_url)  
Figure 4. Supplier’s profit under different supplier cost realizations ($k = 0.8$ vs. $k = -0.8$).

In this example, $b$ has a support on $[0.06, 0.14]$, i.e., there is a remarkable variation in cost $b$. Two immediate observations can be made from Figure 4(a) with $k = 0.8$. First, the supplier’s profit is decreasing in the cost realization in both the OM and the SCQ contract, and the two curves are quite close to each other. Second, the simple contracts may perform better than the
OM contract *ex post* because it leaves less profit for the supplier for certain cost realizations. These observations together may explain why the OM contract does not perform much better than the simple contracts in expectation when countervailing incentives are not present. In contrast, with \( k = -0.8 \), we can see that the curve for the OM contract is decreasing first and then flat in the cost realization. Furthermore, for a subinterval of cost realizations of \( b \), the supplier’s profit is suppressed to zero due to the existence of countervailing incentives. The SCQ contract, on the other hand, is not as effective since it always leaves some positive information rent to the supplier and does not squeeze the supplier’s profit to zero. This provides an intuitive explanation why the simple contracts are less effective when \( k \) is very negative: The OM contract can eliminate the supplier’s information rent under the presence of countervailing incentives, while the SCQ contract is not able to do so.

Finally, to test the robustness of the above observations, we have conducted a numerical study using a normal distribution for \( F \). The design of the numerical study is the same as before except that now cost \( b \) follows a normal distribution with mean \( \mu = \theta \) and standard deviation \( \sigma = \frac{\delta}{\sqrt{2}} \). Figure 5 summarizes the results from this numerical study. As we can see, again, the simple contracts generally perform well for \( k > 0 \), but the performance deteriorates significantly as \( k \) enters the negative domain.

![Figure 5. Outsourcer’s percentage cost increase relative to OM (normal cost distribution).](image-url)
6.2 GI/G/m System

In the previous subsection, we focus on queuing systems with a single server. However, this restriction omits an important set of practical situations, where more than a single server is required to achieve the desirable capacity. Examples include production systems with multiple parallel machines/workers, large-scale call centers, and any other service facilities involving multiple servers. In this subsection, we test the robustness of the previous results using queuing systems with multiple identical servers. In particular, we consider GI/G/m queueing systems: The inter-arrival time follows a general distribution (GI), the service time follows a general distribution (G), and there are \( m \) servers.

Since the accurate congestion measures for a GI/G/m system are quite involved and difficult to analyze, we employ an approximation as follows. Let \( c_a \) be the coefficient of variation of the inter-arrival time and \( c_p \) be the coefficient of variation of the service time. Recall \( \lambda \) is the arrival rate to the system, and let \( \mu_0 \) denote the service rate for an individual server. Then the average number of customers in the system can be approximated by

\[ L = \left( \frac{c_a^2 + c_p^2}{2} \right) \left( \rho \frac{\sqrt{2m+1} - 1}{m(1 - \rho)} \right) \left( \frac{\lambda}{\mu_0} \right) + \frac{\lambda}{\mu_0}, \]

where \( \rho = \frac{\lambda}{m\mu_0} \) is the utilization for the queuing system. This closed-form approximation has been first proposed by Sakasegawa (1977). Whitt (1993) offers a detailed discussion of its performance along with several other approximations. It has been shown that the approximation performs remarkably well except in very light traffic. Recently, it has been adopted as standard textbook formula (see Hopp and Spearman 1996 and Cachon and Terwiesch 2006).

Assume that the service rate for each individual server, \( \mu_0 \), is fixed. Then with GI/G/m queues, the supplier’s capacity decision becomes \( m \), the number of servers to use, and \( q \), the quality of service. Furthermore, assume that \( m \) is large so that we can take it as a continuous decision variable. These assumptions are appropriate, for example, in the case of large-scale call centers, where \( \mu_0 \) is relatively stable across individual employees and \( m \) is the staffing level the call center chooses to use (the total capacity is \( \mu = m\mu_0 \)). Finally, assume that \( g(m) = \alpha m \) is the effort function to achieve a staffing level \( m \) and \( h(q) \) takes the same function form as in Section 6.1. Note that \( h(q) \) is independent of \( m \), i.e., the quality-related effort function is independent of the number of servers in the system. This assumption is to maintain tractability and has also been used by Ren and Zhou (2008). Under these assumptions, the total cost function for the supply
chain is given by

\[
C(m, q) = c_w L + c_g \lambda (1 - q) + bg(m) + (a + kb) h(q)
\]

\[
= c_w \left( \frac{c_a^2}{2} \right) \left( \frac{\rho \sqrt{2(m+1)-1}}{m(1 - \rho)} \right) \left( \frac{\lambda}{\mu_0} \right) + \frac{\lambda}{\mu_0} + c_g \lambda (1 - q) + b \alpha m + \frac{a + kb}{\gamma} \ln \left( \frac{\beta}{1 - q} \right)
\]

\[
= \left( \frac{c_a^2 + c_p^2}{2} \right) \left( \frac{c_w}{s \sqrt{2(\phi s + 1) - 1}} \right) \left( \frac{\lambda}{\mu_0} \right) + c_w \phi + c_g \lambda (1 - q) + b \alpha \phi s + \frac{a + kb}{\gamma} \ln \left( \frac{\beta}{1 - q} \right),
\]

where \( \phi = \frac{\lambda}{\mu_0} \) and \( s = \frac{m}{\phi} > 1 \). Since \( \phi = \frac{\lambda}{\mu_0} \) is fixed, we may view \( s \) as the decision variable on capacity.

It can be shown that \( C(m, q) \) in the above equation is jointly convex in \( (m, q) \) (or equivalently, jointly convex in \( (s, q) \)), and the centralized supply chain has a unique optimal solution. Further, all the results in Sections 4 and 5 apply here. Although there is no a closed-form solution, we may compare the outsourcing contracts presented in the previous sections numerically. Assume \( c_a = c_p = 1 \) in the numerical study (i.e., an M/M/m queue). The qualitative results would not change with general values of \( c_a \) and \( c_p \). The parameters are the same as in Section 6.1, except \( \phi \in \{10, 100\} \), or equivalently, \( \mu_0 \in \{\frac{\lambda}{100}, \frac{\lambda}{10}\} \). The results are summarized in Figure 6 for uniform and normal cost distributions, respectively. It can be seen that the qualitative observations with multiple servers are similar to those with a single server.

Figure 6. Outsourcer’s percentage cost increase relative to OM with multiple servers:

uniform (left) and normal (right) cost distributions.
7 Conclusion

Service outsourcing is an ongoing trend in today’s global economy. However, many firms are surprised to find that their outsourcing endeavors do not result in better efficiency and lowered cost. Instead, many experience a deterioration in service quality and profitability. This paper aims to help firms improve their outsourcing operations. We attribute the less-than-desired outcome to two root causes. First, many firms lack full knowledge about their outsourcing partners, mostly because of geographical distances and unfamiliarity with local environments. Second, they may have failed to consider the impact of the trade-off between various costs for providing fast, quality services. In particular, many firms just focus on the savings from capacity cost by outsourcing, but ignore the associated increase in quality-related cost.

We illustrate the complexity in managing service outsourcing with a principal-agent model with possible countervailing incentives. In our model, the service provider’s capacity and quality costs are private information, and they are correlated with each other. We identify the optimal outsourcing contract that minimizes the total costs for the outsourcer. Such an optimal contract requires the supplier to reveal her cost information by choosing from a non-linear, complex menu of actions. It has been shown that the form of the optimal contract depends on the correlation structure between the service provider’s capacity cost and quality cost. One noteworthy finding is that under negatively correlated costs, the optimal contract may squeeze the supplier’s profit to zero for a non-degenerate interval of cost realizations. In addition, the service supply chain could be coordinated under the optimal contract. These observations are in contrast with those from the literature that involve information asymmetry on a single cost dimension.

Then we study the effectiveness of relatively simple contracts that do not require the cost information as an input. There are two major categories of simple contracts being observed in practice: those that contract directly on capacity and quality levels and those that use financial incentives. We show that financial-incentive contracts are always better for the outsourcer than direct contracts. We also find that both contracts generally perform very well compared to the optimal menu of contracts, as long as the capacity and quality costs are not negatively correlated. However, their performance could be much worse when the supplier’s costs are negatively correlated. Therefore, the relationship between the supplier’s capacity and quality costs plays an important role in the performance of the simple outsourcing contracts.

Our results suggest that outsourcing companies should first invest in understanding the local outsourcing environments of their potential service providers. Particular attention should be paid
to the form and extent of correlated costs of service providers, not just the potential savings on labor costs. Are the agents of service providers easy to train to achieve a high service quality? Is their capacity cost positively or negatively correlated with quality cost? If there is a negative correlation and the resulting countervailing incentives are relatively strong, then firms should adopt more sophisticated contracts, such as those that are contingent on service provider’s cost information, to offset part of the inefficiency from information asymmetry and the trade-off between capacity and quality costs.

On the other hand, if there are no countervailing incentives in local outsourcing environments, firms can use relatively simple and robust contracts, such as those that specify desired capacity and quality levels, to assure a high level of capacity and service quality even when they do not have perfect information about the service provider. This bodes well for companies that are seeking outsourcing, but at the same time are concerned about their lack of deep knowledge about their partners and their potential loss of control over the quality of the work being outsourced.

Even though we study quality and information asymmetry in the context of service outsourcing, we would like to mention that our model embraces the case of make-to-order manufacturing outsourcing. In this environment, the make-to-order manufacturing process is a queue, and the defective rate is the measure of quality. All the results and insights obtained in this paper would apply to the make-to-order setting.

This research can be extended in several directions. First, our model assumes that the demand rate $\lambda$ is common knowledge. In reality, the outsourcer may have better demand forecast information. It would be interesting to study a signalling game where the outsourcer wants to credibly convey the demand information to the supplier. Second, we consider a static setting in this paper. When there are repetitive interactions between the outsourcer and the supplier, reputation would play a major role in making contracting decisions. Also the outsourcer may learn about the supplier’s private cost information over time. Finally, we only consider perfect (either positive or negative) cost correlation in this research. Whether the results will continue to hold with a general correlation structure remains an open question. Though analytically challenging, outsourcing contract design under multi-dimensional information asymmetry is an interesting direction for future research.
Acknowledgments

The authors would like to thank Philipp Afeche, Mor Armony, Ravi Aron, Gerard Cachon, Albert Ha, Victor Martinez de Albeniz, Serguei Netessine, Robert Shumsky, Hongtao Zhang, Yong-Pin Zhou, the participants at the MSOM and INFORMS conferences (2007, 2008), and seminar participants at Hong Kong University of Science and Technology, Chinese University of Hong Kong, and University of Southern California for their helpful comments.

References


Appendix A: Proofs

Proof of Proposition 1 The proof is by noticing that $L(\cdot), g(\cdot)$, and $h(\cdot)$ are all convex functions.

Proof of Proposition 2 We focus on the proof for $k < 0$ because it is more involved. The proof for $k \geq 0$ can be derived similarly.

If a supplier has cost realization $b$ but announces $x$, then her profit function is given by

$$\pi_s(x) = T(x)\lambda - [bg(\mu(x) + (a + kb)h(q(x))].$$

The first-order condition for truth telling is (i.e., $x = b$ maximizes the supplier's profit):

$$T'(x)\lambda - [xg'(\mu)\mu'(x) + (a + kx)h'(q)q'(x)] = 0 \text{ for all } x. \quad (12)$$

The second-order condition for truth telling is

$$\lambda T''(x) - x[g''(\mu)\mu'^2(x) + g'(\mu)\mu''(x)] - (a + kx)h''(q)q'^2(x) + h'(q)q''(x)] \leq 0. \quad (13)$$

Differentiating (12) and plugging into (13) gives

$$g'(\mu)\mu'(x) + kh'(q)q'(x) \leq 0 \text{ for all } x. \quad (14)$$

(12) and (14) are the local incentive constraints, which ensure that the supplier will truthfully announce her cost locally. For global optimality, we need to show that

$$\pi_s(b) \geq \pi_s(x) \text{ for all } x \neq b.$$

We have

\[
\begin{align*}
\pi_s(b) - \pi_s(x) &= T(b)\lambda - [bg(\mu(b)) + (a + kb)h(q(b))] - T(x)\lambda + [bg(\mu(x)) + (a + kb)h(q(x))] \\
&= [T(y)\lambda - yg(\mu(y)) - (a + ky)h(q(y))]|_x^b + (b - x)g(\mu(x)) + [(a + kb) - (a + kx)]h(q(x)) \\
&= \int_x^b [T(y)\lambda - yg(\mu(y)) - (a + ky)h(q(y))] dy + \int_x^b g(\mu(x))dy + \int_x^b kh(q(x))dy \\
&= \int_x^b \{-g(\mu(y)) - kh(q(y)) + g(\mu(x)) + kh(q(x))\} dy,
\end{align*}
\]
where the last equality is by applying (12). Assume $\mu'(x) \leq 0$ and $q'(x) \geq 0$ in the optimal mechanism. This is intuitive because it implies that the capacity (quality) level decreases (increases) in the supplier’s cost realization, $x$, which also satisfies (14). Then it follows that

$$\pi_s(b) - \pi_s(x) = \int_x^b \{[g(\mu(x)) - g(\mu(y))] + k[h(q(x)) - h(q(y)))] \, dy \geq 0. \quad (15)$$

Since $k < 0$, it is clear that (14) and (15) hold if $\mu' \leq 0$ and $q' \geq 0$. Truth telling implies that

$$\pi_s(x) = T(x)\lambda - [xg(\mu(x)) + (a + kx)h(q(x))]$$

for all $x$.

Taking derivative with respect to $x$ and by (12), we have

$$\pi'_s(x) = -[g(\mu(x)) + k h(q(x))]. \quad (16)$$

Since $T(x) = \pi_s(x) + [xg(\mu) + (a + kx)h(q)]$, the outsourcer’s problem in (7) can be rewritten as

$$\min_{\{\mu(\cdot), q(\cdot), \pi_s(\cdot)\}} \int_b^x \left[ c_w L(\mu) + c_y \lambda (1 - q) + \pi_s(x) + xg(\mu) + (a + kx)h(q) \right] f(x) \, dx$$

s.t. (16), and (6),

where the first constraint is derived from the necessary condition for truth telling and the second is the IR constraint. This is an optimal control problem: $\pi_s$ is the state variable and $\mu$ and $q$ are the control variables. Define

$$H(\mu, q, \pi_s, \eta, x) = -[c_w L(\mu) + c_y \lambda (1 - q) + \pi_s(x) + xg(\mu) + (a + kx)h(q)] f(x) - \eta(x) [g(\mu) + k h(q)]$$

as the Hamiltonian, where $\eta$ is the co-state variable. Optimizing $H$ with respect to $\mu$ and $q$ yields the necessary conditions for $\mu^o$ and $q^o$:

$$c_w L'(\mu^o) + [x + \eta(x)/f(x)] g'(\mu^o) = 0, \quad (17)$$

$$c_y \lambda - [(a + kx) + k \eta(x)/f(x)] h'(q^o) = 0. \quad (18)$$

Since $L$, $g$, and $h$ are all convex functions, we know that $H$ is concave in $\mu$ and $q$, which implies that (17) and (18) are also sufficient conditions for optimality. Note that $\mu' \leq 0$, $q' \geq 0$ and (16) together are sufficient for truth telling. Thus we may solve the outsourcer’s problem by identifying a pair of functions that satisfy the following conditions: $\mu' \leq 0$, $q' \geq 0$, (16), (17) and (18).

Start with the optimality conditions in (17) and (18). We need to determine the co-state variable $\eta(x)$. By the Pontryagin principle, we have

$$\eta'(x) = -\partial H/\partial \pi_s = f(x)$$
at all differentiable points. The co-state variable $\eta(x)$ can be solved using the transversality conditions, which depend on the service provider’s profit function $\pi_s(x)$ at the boundary points. Recall that $g$ and $h$ are convexly increasing functions. Under the conditions $\mu' \leq 0$ and $q' \geq 0$, we know that $\pi'_s(x)$ in (16) is increasing in $x$. So there are three possibilities in the optimal menu of contracts: (1) $\pi'_s(x) \leq 0$ for all $x$; (2) $\pi'_s(x) \geq 0$ for all $x$; (3) $\pi'_s(x)$ starts at a negative value, and then increases to a positive value ($\pi'_s(x)$ could be zero in a non-degenerate subinterval). In each case, the transversality conditions for determining $\eta(x)$ will be different. Next we consider each of the three possible cases.

**1** $\pi'_s(x) \leq 0 \text{ for all } x$  In this case, the supplier’s profit decreases in cost $x$. It must be true that the supplier’s profit in the optimal mechanism is $0$ at $x = \bar{b}$ (otherwise the outsourcer can reduce the supplier’s profit without affecting the truth-telling incentives). This means that the state variable $\pi_s(x)$ is fixed at $x = \bar{b}$ while it is free at $x = \underline{b}$. Thus the transversality condition is $\eta(\bar{b}) = 0$ and we know $\eta(x) = F(x) = \eta_1(x)$. As a result, the solution for $\mu^o$ and $q^o$ is given by $\mu^o = \mu_1(x)$ and $q^o = q_1(x)$ for all $x$. Note that we need $\pi'_s(\bar{b}) \leq 0$, or equivalently, $-g(\mu_1(\bar{b})) - kh(q_1(\bar{b})) \leq 0$ to ensure $\pi'_s(x) \leq 0$ for all $x$. Therefore, if $g(\mu_1(\bar{b})) + kh(q_1(\bar{b})) \geq 0$, then the outsourcer’s optimal menu of contracts is given by $\mu^o = \mu_1(x)$ and $q^o = q_1(x)$. Since $F$ is log-concave (i.e., $F(x)/f(x)$ is increasing), it is straightforward to show that $\mu'^o \leq 0$ and $q'^o \geq 0$. We can set $b_1 = b_2 = \bar{b}$ in this case.

**2** $\pi'_s(x) \geq 0 \text{ for all } x$  This is the opposite of case (1). In particular, the supplier’s profit increases in cost $x$. It must be true that the supplier’s profit in the optimal mechanism is $0$ at $x = \underline{b}$ (otherwise the outsourcer can reduce the supplier’s profit without affecting the truth-telling incentives). This means that the state variable $\pi_s(x)$ is fixed at $x = \underline{b}$ while it is free at $x = \bar{b}$. Thus the transversality condition is $\eta(\underline{b}) = 0$ and we know $\eta(x) = F(x) - 1 = \eta_2(x)$. As a result, the solution for $\mu^o$ and $q^o$ is given by $\mu^o = \mu_2(x)$ and $q^o = q_2(x)$ for all $x$. Note that we need $\pi'_s(\underline{b}) \geq 0$, or equivalently, $-g(\mu_2(\underline{b})) - kh(q_2(\underline{b})) \geq 0$ to ensure $\pi'_s(x) \leq 0$ for all $x$. Therefore, if $-g(\mu_2(\underline{b})) - kh(q_2(\underline{b})) \geq 0$, then the outsourcer’s optimal menu of contracts is given by $\mu^o = \mu_2(x)$ and $q^o = q_2(x)$. Given that $f(x)/(1 - F(x))$ is increasing, we can show that $\mu'^o \leq 0$ and $q'^o \leq 0$ hold in this case. Here we have $b_1 = b_2 = \underline{b}$.

**3** $\pi'_s(x)$ increases from the negative domain to the positive domain  In this case, $\pi'_s(x) = -[g(\mu(x)) + kh(q(x))] = 0$ may hold for a non-degenerate subinterval in $[\underline{b}, \bar{b}]$. Denote this
subinterval by \([b_1, b_2]\) (if \(b_1 = b_2\), then the subinterval collapses to a single point). Clearly, there must be \(\pi_s(x) = 0\) for \(x \in [b_1, b_2]\), because otherwise the outsourcer can reduce the supplier’s profit without affecting the truth-telling incentives. Since \(\pi_s(b) > 0\) and is not fixed, the transversality condition is \(\eta(b) = 0\) and we know \(\eta(x) = F(x) = \eta_1(x)\) for \(x < b_1\). Similarly, we know \(\eta(x) = F(x) - 1 = \eta_2(x)\) for \(x > b_2\). Thus, \(b_1\) and \(b_2\) must satisfy the following two equations

\[
\begin{cases}
\pi'_s(b_1) = g(\mu_1(b_1)) + kh(q_1(b_1)) = 0 \\
\pi'_s(b_2) = g(\mu_2(b_2)) + kh(q_2(b_2)) = 0
\end{cases},
\]

respectively (if a solution does not exist to one of the equations, then set \(b_1 = \bar{b}\) or \(b_2 = \bar{b}\)). Since \(\mu'_i \leq 0\), \(q'_i \geq 0\), \(k < 0\), and \(\eta_1(x) > \eta_2(x)\), we know there must be \(b_1 \leq b_2\). Further, since \(\pi'_s(x)\) remains a constant in a subinterval, both \(\mu\) and \(q\) must also be constants. (This implies \(x + \eta(x)/f(x) = b_1\), from which we can solve the co-state variable \(\eta(x)\) for \(x \in [b_1, b_2]\).) In addition, there must be \(\mu^o(x) = \mu_1(b_1) = \mu_2(b_2)\) and \(q^o(x) = q_1(b_1) = q_2(b_2)\) for \(x \in [b_1, b_2]\). Therefore, the optimal menu of contracts are characterized by

\[
\begin{cases}
\mu^o(x) = \mu_1(x), q^o(x) = q_1(x) \text{ for } x \leq b_1, \\
\mu^o(x) = \mu_1(b_1), q^o(x) = q_1(b_1) \text{ for } b_1 < x \leq b_2, \\
\mu^o(x) = \mu_2(x), q^o(x) = q_2(x) \text{ for } x > b_2.
\end{cases}
\]

Given log-concavity and increasing failure rate for \(F\), we know that \(\mu^o \leq 0\) and \(q^o \geq 0\) hold in this case as well.

In all of the above three cases, once \(\mu^o\) and \(q^o\) are determined, we can solve \(T^o\) using the IC and IR constraints in (16), (6).

Finally, we can derive the optimal menu for \(k \geq 0\) similarly. With \(k \geq 0\), the structure of the optimal menu is the same as in Case (1) above where there is \(\pi'_s(x) \leq 0\) for all \(x\). In other words, we have

\[
\mu^o(x) = \mu_1(x), q^o(x) = q_1(x) \text{ for all } x.
\]

The details are omitted. \(\Box\)

**Proof of Proposition 3** For easy reference, the truth-telling condition is given by

\[
\pi'_s(x) = -[g(\mu(x)) + kh(q(x))],
\]

and the optimality conditions are

\[
\begin{align*}
c_wL'(\mu^o) + [x + \eta(x)/f(x)]g'(\mu^o) & = 0, \\
c_2\lambda - [(a + kx) + k\eta(x)/f(x)]h'(q^o) & = 0.
\end{align*}
\]
We first show that there exists a $k_1 < 0$ such that Case (1) happens if and only if $k \geq k_1$. If $k \geq 0$, then Case (1) will happen for sure. So we only need to focus on $k < 0$. Consider $\pi'_s(x) = -[g(\mu_1(x)) + kh(q_1(x))]$, where $\mu_1$ and $q_1$ are the solutions for $\eta(x) = F(x)$ as defined in Proposition 2. From (21) and $\eta(x) = F(x)$, it is easy to show that $q_1(x)$ decreases in $k$ (note $k < 0$), which implies $kh(q_1(x))$ increases in $k$. Thus, for any given $x$, $\pi'_s(x) = -[g(\mu_1(x)) + kh(q_1(x))]$ decreases in $k$. Now, consider $\pi'_s(\bar{b}) = -[g(\mu_1(\bar{b})) + kh(q_1(\bar{b}))]$. For $k$ close enough to zero, there must be $\pi'_s(\bar{b}) \leq 0$; in addition, $\pi'_s(\bar{b})$ increases as $k$ decreases (i.e., become more negative). Thus, there exists a $k_1$ such that $\pi'_s(\bar{b}) = 0$ at $k = k_1$. Next we argue that scenario (1) happens if and only if $k \geq k_1$. Since $\pi'_s(\bar{b})$ is monotone decreasing in $k$, we know that there is $\pi'_s(\bar{b}) \leq 0$ for any $k \geq k_1$. Note that $\mu_1(x)$ decreases in $x$ and $q_1(x)$ increases in $x$, so for any fixed $k$, $\pi'_s(x)$ increases in $x$. Hence there is $\pi'_s(x) \leq 0$ for all $x$ for any $k \geq k_1$. This means that scenario (1) happens if $k \geq k_1$. Note that for any $k < k_1$, scenario (1) cannot be optimal since $\pi'_s(\bar{b}) > 0$. Thus $k \geq k_1$ is necessary for scenario (1) to happen. This completes the first part of the proof.

Second, we show that there exists a $k_2$ such that scenario (2) happens if and only if $k < k_2$. Consider $\pi'_s(x) = -[g(\mu_2(x)) + kh(q_2(x))]$, where $\mu_2$ and $q_2$ are the solutions for $\eta(x) = F(x) - 1$ as defined in Proposition 2. From (21) and $\eta(x) = F(x) - 1$, we know that $q_2(x)$ decreases in $k$ (note that $k < 0$ and $x + \eta(x)/f(x) > 0$ in (20)), which implies $kh(q_1(x))$ increases in $k$. Thus, for any given $x$, $\pi'_s(x) = -[g(\mu_1(x)) + kh(q_1(x))]$ decreases in $k$. Now, consider $\pi'_s(\bar{b}) = -[g(\mu_2(\bar{b})) + kh(q_2(\bar{b}))]$. Define $k_2$ to be such that $\pi'_s(\bar{b}) = 0$ at $k = k_2$. Next we argue that scenario (2) happens if and only if $k < k_2$. Since $\pi'_s(\bar{b})$ is monotone decreasing in $k$, we know that there is $\pi'_s(\bar{b}) \geq 0$ for any $k < k_2$. Note that $\mu_2(x)$ decreases in $x$ and $q_2(x)$ increases in $x$, so for any fixed $k$, $\pi'_s(x)$ increases in $x$. Hence there is $\pi'_s(x) \geq 0$ for all $x$ for any $k < k_2$. This means that scenario (2) happens if $k < k_2$. Note that for any $k > k_1$, scenario (2) cannot be optimal since $\pi'_s(\bar{b}) < 0$. This completes the second part of the proof.

Finally, it follows immediately that scenario (3) happens if and only if $k_2 \leq k_2 < k_1$. □

**Proof of Proposition 4** The proof is by checking the optimal capacity function $\mu^o(x)$ and the optimal quality function $q^o(x)$ in Proposition 2 and its proof. □

**Proof of Proposition 5** Parts (1) and (2) are straightforward by comparing the optimal menu in Proposition 2 to the supply chain optimal solution characterized in Proposition 1.

For (3), note that $\mu^o(x)$ decreases in $x$ in general and is flat on $[b_1, b_2]$, while $\mu^*(x)$ is (strictly) decreasing in $x$ for all $x$. In addition, we know (from Parts 1 and 2) that $\mu^o(x) > \mu^*(x)$ for
x < b_1 and \( \mu^*(x) < \mu^*(x) \) for \( x > b_2 \). This implies that there exists a unique point \( \hat{b} \) such that \( \mu^*(\hat{b}) = \mu^*(\hat{b}) = \mu^*(b_1) \). In particular, from (2) and (8), \( \mu^*(\hat{b}) = \mu^*(b_1) \) implies

\[
b_1 + \eta_1(b_1)/f(b_1) = \hat{b}.
\]

Together with (3), (9) and (22), similar arguments lead to \( q^*(\hat{b}) = q^*(\hat{b}) = q^*(b_1) \). Thus the supply chain is coordinated at \( x = \hat{b} \). \( \square \)

**Proof of Proposition 6** In the SCQ contract, the outsourcer’s problem can be written as

\[
\min_{(\mu^*, q^*)} C_o(\mu^*, q^*) = [c_wL(\mu^*) + c_g\lambda(1 - q^*) + b^*g(\mu^*) + v(b^*)h(q^*)].
\]

Since \( b^* = \arg \max_b [bg(\mu^*) + v(b)h(q^*)] \), we have

\[
\min_{(\mu^*, q^*)} C_o(\mu^*, q^*) = \min_{(\mu^*, q^*)} \left\{ c_wL(\mu^*) + c_g\lambda(1 - q^*) + \max_b [bg(\mu^*) + v(b)h(q^*)] \right\}
\]

\[
= \min_{(\mu^*, q^*)} \max_b \left\{ c_wL(\mu^*) + c_g\lambda(1 - q^*) + [bg(\mu^*) + v(b)h(q^*)] \right\}
\]

\[
= \min_{(\mu, q)} \max_b C_c(\mu, q, b).
\]

Recall \( C_c \) is the supply chain’s total cost. It is easy to verify that \( C_c(b, (\mu, q)) \) is jointly convex in \( (\mu, q) \) for every \( b \) and it is concave in \( b \) for every pair \( (\mu, q) \). According to the min-max theorem (Ponstein, 1965), we know that

\[
\min_{(\mu, q)} \max_b C_c((\mu, q), b) = \max_{b} \\min_{(\mu, q)} C_c((\mu, q), b)
\]

\[
= \max_{b} \min_{(\mu, q)} [c_wL(\mu) + c_g\lambda(1 - q) + bg(\mu) + v(b)h(q)].
\]

\[
= C_c(b^m).
\]

Thus we have shown that the outsourcer’s optimal cost in the SCQ contract is equal to the supply chain’s maximal cost under \( (b^m, (\mu^m, q^m)) \). In addition, if \( (\mu^m, q^m) \) is unique, then we know that it is both a max-min point and a min-max point. That is, \( \mu^* = \mu^m \) and \( q^* = q^m \). \( \square \)

**Proof of Proposition 7** Consider the following parameters in the QWP contract: the outsourcer sets \( w = c_w \) and \( K = c_g \). Recall that \( c_w \) and \( c_g \) are the waiting cost and the goodwill cost incurred by the outsourcer. Under such a contract, the outsourcer essentially transfers all the costs in the service supply chain to the supplier and the supplier will choose the centralized optimal solution given in Proposition 1. Further, the supplier’s total cost will be the supply chain’s optimal cost \( C_c(\mu^*, q^*) \). To ensure participation, the outsourcer needs to pay the supplier a lump-sum payment
\( T\lambda = C_c(b^m) \), i.e., the supply chain’s optimal cost with \( b^m \). From Proposition 6, we know that in the SCQ contract, the outsourcer’s total cost is equal to \( C_c(b^m) \) too. Since \( w = c_w \) and \( K = c_g \) is a special case of the QWP contract, the claimed result then follows. \( \square \)

Appendix B: Extensions and Additional Numerical Studies

In Appendix B we discuss some extensions of the basic model and present more numerical results. The first extension deals with a situation where dissatisfied customers may come back, and the second considers impatient customers, who may leave the queue in a random fashion. Finally, we test the robustness of the results in the paper by using an alternate set of capacity and quality effort functions.

Retrial

Retrial is a common phenomenon in service operations. A customer may come back again after experiencing a failed service (e.g., call center). In a manufacturing setting, rework may happen if a defective product is returned to the manufacturer. This extension introduces retrial into the basic model. Consider an M/M/1 queueing system and each customer is satisfied with probability \( q \). Assume among the dissatisfied customers, a portion of \( 1 - p \) will return to the system for service again until they are satisfied. One may imagine these dissatisfied customers are put in the front of queue immediately after they return. This indicates that the dissatisfied customers have higher priority to be served and also simplifies the analysis. Actually, where to put the returned customers does not matter if one is only interested in the average time a customer spends in the system, which is our case. It can be readily shown that the impact of retrial is that the service rate becomes \([1 - (1 - q)p]\mu\) instead of the original \( \mu \) (i.e., retrials essentially reduce the server’s speed by a certain factor). Therefore, given a capacity level \( \mu \) and a quality level \( q \), the supply chain’s total cost becomes

\[
C(\mu, q) = \frac{c_w\lambda}{[1 - (1 - q)p]\mu - \lambda} + c_g\lambda(1 - q) + [bg(\mu) + v(b)h(q)].
\]

The analysis of contracts (especially the OM) becomes more complicated because there exists a product term of the two decision variables \( \mu \) and \( q \). Nevertheless, we can still evaluate the contracts numerically. In our numerical study, all the parameters are the same as in the first numerical study in Section 6.1, except now we have an additional parameter \( p = \{0.1, 0.5, 0.9\} \). The purpose here is to test whether the simple contracts still perform well with positive \( k \) values (the numerical studies in the main paper have already demonstrated that the simple contracts may perform poorly with negative \( k \) values), so we allow \( k \) to take values in \( \{0, 0.2, 0.4, 0.6, 0.8, 1.0\} \). We compare the SCQ
and the OM and find that the percentage cost increase of the SCQ over the OM is at most 1.453% and the average is 0.369% among all scenarios. We know that the QWP is even better than the SCQ by applying the same argument in Proposition 3. This shows that the simple contracts continue to perform well for the outsourcer even taking customer retrial into consideration.

**Impatient Customers** In this extension, we consider a queueing system with abandonment. For example, a customer who is put on hold may drop a call if the waiting time exceeds a certain threshold value. An M/M/N+M abandonment model is used to study the impact of customer impatience. That is, customers' patience is exponentially distributed and there are N identical servers in the system. Again $\mu_0$ stands for the individual server’s service rate. Let $\eta$ denote the abandonment rate for customers and for each abandoned call, there is a good will loss $c_d$ incurred by the outsourcer. Define $A(N)$ to be the equilibrium abandonment probability given a capacity level $N$. For tractability, we apply the diffusion approximation for $A(N)$ characterized by the following equations:

$$A(N) = \left(1 - \frac{z(\xi \sqrt{\mu_0 / \eta})}{z(\xi \sqrt{\mu_0 / \eta} + \sqrt{\eta/(N\mu_0)})}\right)\psi\left(-\xi, \sqrt{\frac{\mu_0}{\eta}}\right),$$

where

- $\xi = N\sqrt{\frac{\mu_0}{\lambda}} - 1$,
- $z(x) = \frac{\phi(x)}{1 - \Phi(x)}$,
- $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, $\Phi(x) = \int_{-\infty}^{x} \phi(y)dy$,
- $\psi(x, y) = \left(1 + \frac{z(-xy)}{yz(x)}\right)^{-1}$.

The diffusion approximation is appropriate for moderate to large-scale call centers that are operated under high utilization (see Garnett, Mandelbaum and Reiman, 2002). Therefore, we also assume that $N$ is continuous under this approximation. It can be shown that the average queue length in this M/M/N+M model is $\lambda A(N)/\eta$. Thus the service supply chain’s total cost, given a capacity level $N$ and a quality level $q$, can be expressed as

$$C(N, q) = c_w\lambda A(N)/\eta + c_g\lambda(1-q)(1-A(N)) + c_d\lambda A(N) + [bg(N) + v(b)h(q)].$$

Notice that $A(N)$ is complex and also there exists a product of the two decision variables in the cost function (i.e., $c_g\lambda(1-q)(1-A(N))$), which complicates the analysis of the contracts. Although we can manage to evaluate the contracts numerically, we have to restrict to the effort
function \( h(q) = \frac{1}{\gamma} \ln \left( \frac{\beta}{1-q} \right) \) and the scenarios in which an interior solution exists for the quality decision, \( q \). The following parameters have been used: \( v(b) = \bar{b} + kb \) and \( k \) takes values from \( \{0, 0.2, 0.4, 0.6, 0.8, 1.0\} \), \( \lambda = 100 \) and \( \mu_0 \) is normalized to be 1, \( g(N) = \alpha N \) and \( \alpha \) is normalized to be 1, \( \eta = \{0.1, 1, 10\} \), \( \theta = \{0.1, 1, 10\} \), \( \delta = \{0.4\} \), \( c_w = \{0.1, 1, 1\} \), \( c_g = c_d = \{1, 10, 30\} \), \( \beta = \{0.1, 0.5, 0.9\} \) and \( \gamma = \{1, 20\} \). There are 486 scenarios in total for each \( k \) value. We find that the maximum percentage cost increase of the SCQ over the OM is 0.919\% and the average is only 0.161\%. Again, the performance of the QWP is even better according to the argument in Proposition 7. Therefore, the simple contracts continue to perform well for the outsourcer after counting for impatient customer behavior.

**Alternate Capacity and Quality Effort Functions** Recall that \( g(\mu) \) is the effort function for achieving a target capacity level \( \mu \) and \( h(q) \) is the effort function for achieving a target quality level \( q \). In order to compare the contracts, Section 6.1 uses specific functional forms: \( g(\mu) \) and \( h(q) \) take linear and logarithmic functions, respectively. Here we study an alternate set of effort functions. Consider \( g(\mu) = \alpha \mu^t \) (\( \alpha > 0 \) and \( t \geq 1 \)) and

\[
    h(q) = \left[ \frac{q - 1 + \beta}{\gamma(1-q)} \right]^{\frac{1}{\gamma}},
\]

or

\[
    q(h) = (1 - \beta) + \beta \left( \frac{\gamma h^\tau}{1 + \gamma h^\tau} \right), \tag{23}
\]

where \( 0 \leq \beta \leq 1 \), \( \gamma > 0 \), and \( 0 < \tau < 1 \). Note that \( q(h) \) in (23) is concave, increasing, and bounded by \( [0, 1] \). This \( q(h) \) function has been first proposed by Little (1970) to study the sales response to marketing effort. Although it is complex, such a functional form has been commonly used in the management science literature due to its versatility in constructing different function curves by adjusting parameters. For instance, Li and Rajagopalan (1998) applies it to describe the relationship between product quality and process improvement effort.

We use the following parameters to compare the performances of contracts: \( v(b) = \bar{b} + kb \) and \( k \) takes values from \( \{0, 0.2, 0.4, 0.6, 0.8, 1.0\} \), \( \lambda = \{10, 100\} \), \( \theta = \{5, 30\} \), \( \delta = \{0.4\} \), \( c_w = \{1, 10\} \), \( c_g = \{1, 10, 100\} \), \( \beta = \{0.5, 0.9\} \), \( \gamma = \{0.1, 1\} \), \( t = \{1.5, 2\} \), and \( \tau = \{0.6, 0.99\} \). We find that the maximum percentage cost increase of the SCQ over the OM is 0.502\% and the average is only 0.195\%. Together with Proposition 7 (which shows that the QWP yields a lower cost than the SCQ), we know that the simple contracts again perform well for the outsourcer with the alternate capacity and quality effort functions.
References

