Asymmetric Information and Economies of Scale in Service Contracting

Mustafa Akan
Tepper School of Business, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, akan@cmu.edu

Barış Ata, Martin A. Lariviere
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208
[b-ata@kellogg.northwestern.edu, m-lariviere@kellogg.northwestern.edu]

We consider outsourcing in two important service settings: call center and order fulfillment operations. An important factor in both is the inherent economies of scale. Therefore, we advance a unifying model covering both applications and study the associated contracting problem under information asymmetry. At the time of contracting, the outsourcing firm, “the originator,” faces uncertainty regarding the demand volume but has private information about its probability distribution. The true demand is quickly observed once the service commences. The service provider invests in capacity before the start of the operation and offers a menu of contracts to screen different types of the originator. Adopting a mechanism design approach, we prove that a menu of two-part tariffs achieves the full-information solution. Hence, it is optimal among all possible contracts (in both settings) because of economies of scale and contractibility of realized demand.

Key words: service outsourcing; call centers; order fulfillment operations; economies of scale; information asymmetry; screening

History: Received: September 5, 2007; accepted: April 10, 2010. Published online in Articles in Advance October 13, 2010.

1. Introduction

The last 20 years have seen an increasing use of outsourcing. Beginning in manufacturing, firms have turned to others to do work for them. Over time, this practice has escalated from purchasing simple commodity parts to having suppliers provide complex parts and subassemblies. For its new 787 Dreamliner, Boeing is counting on outside firms to deliver doors, landing gear, and even entire wings (Niezen and Weller 2006). In the auto industry, it is not uncommon for purchased materials to account for more than half the cost of a car.

Outsourcing, however, is no longer utilized only for manufacturing. Service outsourcing began with simple ancillary activities such as janitorial services, but has grown to include outsourcing of entire business processes (such as order taking) and departments (such as information technology). TPI, a consulting firm that specializes in sourcing, estimates that the value of new business process outsourcing (BPO) contracts exceeded $90 billion in 2008. TPI’s estimates do not include government contracts, deals under $50 million, or contracts renewed without the help of outside consultants. Looking at broader measures, Cohen and Young (2006) forecasted that the BPO market exceeded three-quarters of a trillion dollars in 2008.

Call center outsourcing is a significant slice of the overall BPO market. Furness (2005) estimated that the U.S. call center outsourcing market would reach £23.9 billion by 2008, whereas Beasty (2006) reported that there are over 150 outsourcing vendors for call centers in North America. Call center contracts take a variety of forms. One form of contracts includes hourly charges based on the number of agents available to take calls for a client. A second alternative is a fee for each minute that an agent is engaged with a caller. The third possibility is a charge per call handled. Centris Information Systems preannounces contracts pairing a per-minute charge with a minimum utilization level. Accolade Support offers packages in which buyers purchase a minimum number of minutes, with additional minutes provided at higher prices. The total payment of the originator consists of a fixed amount paid in the beginning, plus a usage fee increasing linearly with volume. This scheme resembles a two-part tariff where the customer pays an initial fixed fee plus a constant price for each unit. The targeted service level, whether measured by waiting time targets or abandonment rates, affects charges, where higher service lev-

1 Much of the information on typical call center contracts comes from Centris Information Systems (2007) and conversations with a Centris executive.

els increase costs. Costs increase in service levels, but the rate of increase depends on the call volume. Therefore, certain call centers naturally exhibit economies of scale, so it is cheaper on a per-call basis to provide good service when the call volume is high. This is often reflected in pricing, as in the case of Accolade Support, which is also incorporated in our analysis.

Outsourcing order fulfillment operations is another growing segment of the BPO market. E-commerce firms find outsourcing order fulfillment, warehousing, and inventory operations especially attractive. By doing so, firms exploit the benefits of the Internet while avoiding the investment and ongoing management expenses required to effectively stock and distribute their products. For instance, Amazon.com’s Fulfillment by Amazon (FBA) service enables merchants to fulfill orders from sales channels other than Amazon.com using the inventory that they store in the Amazon fulfillment centers. Participants in the program pay Amazon an inventory storage fee based on daily average volume (measured in cubic feet), and Amazon ultimately ships items to customers when they are ordered online (and charges the seller a variable fee that also depends on the weight of the item).

Another example is ProLog Logistics, which operates distribution facilities in San Diego, California, and Lexington, Kentucky. In addition to its order fulfillment and warehousing services, ProLog Logistics provides inventory management and replenishment services. ProLog works with its clients to develop reorder and safety-stock parameters. These are used when ProLog executes purchase orders on behalf of the client, who is relieved of closely monitoring the replenishment of their inventory held at the ProLog distribution centers. ProLog’s pricing is largely based on a per-order and per-piece basis, although volume discounts are also typical. For warehouse storage, clients pay according to the space they use. Similar to call centers, achieving service-level targets increases costs, yet certain inventory systems also exhibit economies of scale. This makes it less costly on a per-volume basis to achieve a given service level for larger systems.

We study service contracting in both call center and order fulfillment settings. Despite their disparities, certain call centers and order fulfillment systems both exhibit economies of scale. More precisely, queuing systems with multiple servers and inventory systems with constant leadtimes exhibit economies of scale, i.e., larger systems operate more efficiently. We use the term economies of scale if the average costs are strictly decreasing with scale, whereas diseconomies of scale refers to the reverse effect. The remaining case is constant returns to scale or linear costs. The economies of scale common to certain queuing and inventory systems enabled a unifying formulation. The results for the two applications follow as special cases.

There is information asymmetry at the time of contracting because the outsourcing firm (which we call the originator) and the service provider do not have access to the same information when they sign a contract. The originator faces uncertainty about its demand but has private information on its probability distribution. The originator can be one of two types: high or low. A high-type originator anticipates a larger demand volume than a low type. At the time of contracting, the originator knows his type with certainty but the service provider only has a prior over the two possibilities. The information asymmetry can arise when the originator is able to affect uncertain demand through exerting promotion effort or increasing quality. Both of these activities improve the originator’s assessment of customer interest for the product. One motivating example is outsourcing the handling of sales calls (or fulfillment of customer orders) that result from one-time promotional deals, for which the time horizon of required service is also short. For instance, commercial banks often mail credit card offers but have a contractor process the applications who cannot foresee the rate of applications linked to campaigns. Another example is online retail stores’ short-term coupons that increase consumer demand for products but cannot be fully known by the service provider. In a related setting, Soyer and Tarimcilar (2008) analyzes call center data based on the sales of a consumer electronics producer and show that the media expenses and advertisements have significant effects on the call volume.

Capturing the demand information privately held by the originator is significant for the service provider to plan ahead. Namely, the service provider needs to train personnel in the call center setting and lease warehouse space in the order fulfillment setting for the specific service requested by the originator. We focus on the case where the service provider uses the same technology to serve both types of the originator, e.g., she uses similar switching systems, automated call distribution (ACD), computer telephony integration (CTI), and interactive voice response services in the call center setting, and has similar presorting, packaging, RFID, and tracking capabilities, as well as the material-handling equipment, in the order fulfillment setting. It is well known that different technologies could be used at different scales, in which case identifying the type of the originator clearly decreases the service provider’s cost. Instead, we study how the service provider can manage variation in scale within a range that would lead to a single class of technology and still show that capturing private demand information is significant to reduce costs. Satisfying larger demand requires training more personnel or leasing more space. At the same time, the demand is uncertain before the start of service, although the realized demand can quickly be inferred after the service commences.
The service provider faces a newsvendor-like problem in choosing the size of the personnel to train or the amount of space to lease prior to service, where both decisions depend on the anticipated demand. Planning ahead saves costs. It is cheaper to train personnel or lease space before the start of the operation. Identification of proper talent and its training take time. Alternatively, the organization can hire through a temporary service agency that offers talent at short notice, but demands a higher price than the “usual” hiring process (cf. Milner and Pinker 2001). Training too many personnel or leasing too much space in advance is risky as well. The training and leasing costs, once incurred, cannot be reversed. Learning the true demand distribution from the originator enables the service provider to lower her expected training or leasing costs by planning accordingly.

To lower her expected training and leasing costs, the service provider prefers to contract with the originator before the start of the operation and separate the two types of the originator. In other words, the service provider offers a menu of contracts, and identifies, or screens, the originator’s type from his choice of contract from the menu.

The originator and the service provider contract on a service level and an associated payment arrangement. We measure service quality in the call center setting as the fraction of calls answered, or equivalently, the abandonment rate. Once the call center is operational, its manager can observe the arrival rate of the calls and adjust the staffing levels to satisfy the service level specified in the contract. The fill rate is the measure of service quality in the context of outsourcing order fulfillment operations. Likewise, the inventory manager can quickly learn the demand rate for the product and adjust the inventory level to fulfill her contractual obligations. Note that the staffing level (the storage room for inventory) is potentially far below the total number of trained personnel (total leased space). Because the demand volume can be quickly verified ex post, the service level and payment may depend on its realized value.

By taking a mechanism design approach, we look for the profit-maximizing direct-revelation mechanisms for the service provider when she only knows the distribution of the originator’s type. The canonical problem in mechanism design considers a seller offering multiple levels of quality to customer segments that differ in their willingness to pay for quality. This leads to the standard results that the full-information solution cannot be implemented in an incentive-compatible mechanism; and in the optimal mechanism the most favorable type receives an efficient level of quality and captures an information rent whereas the least favorable type’s allocation is distorted (see Salanié 1997). In contrast, we find that a menu of two-part tariffs implements the full-information solution and hence is optimal among all possible contracts.

Economies of scale coupled with the observability of the realized demand enables a menu of two-part tariffs that implement the full-information solution. The originator’s outside option is modeled as his expected profit if he were to conduct the operations in house. Consequently, the high-type originator’s outside option (per demand volume) is higher because he enjoys the benefits of economies of scale. On the one hand, a low-type originator is deterred by the large fixed payment in the contract designed for the high-volume originator. On the other hand, the high-type originator is content with paying a large lump-sum payment as opposed to being subject to the payment schedule of the low type, where the per-volume charge exceeds the high-type’s lump-sum payment divided by his expected demand volume.

Two-part tariffs are widely adopted mainly because they are simple and easy to implement (cf. Wilson 1993). We show that they are also effective in revealing the originator’s demand information. Our analysis validates the use of two-part tariffs and proposes that any new such contracts between the call center and inventory managers should take these forms; they can achieve perfect price discrimination by simply offering the originator the choice to select his payment from a menu that consists of bundles of fixed payments and linear payment schedules. The two-part tariffs also arise as the optimal pricing scheme in access services when the capacity is limited and customers have different usage rates (cf. Essegaier et al. 2002).

The rest of the paper is structured as follows: Section 2 provides a literature review. Section 3 presents a unifying model of service outsourcing and the mechanism design problem. Section 4 solves the mechanism design problem. Sections 5 and 6 apply the results to outsourcing of call center and order fulfillment operations, respectively. Section 7 discusses the results and concludes the paper.

2. Literature Review
The theory of contracts analyzes the strategic interactions among agents when informational asymmetries are present; see Salanié (1997) for an introduction to agency models. The literature on contracting for service operations is thin. Milner and Olsen (2008) consider a call center with both contract and non-contract customers that prioritizes contract customers only in the off-peak hours. The authors show that this behavior is rational under contracts on the percentage of delay, but also undesirable for the contract customers; therefore, they also propose novel contracts that eliminate such behavior. Lee et al. (2009) consider the outsourcing of a two-level service process.
where contracts occur under full information. Aksin et al. (2008) consider a service provider who faces an uncertain volume and can outsource all or part of the calls. The authors look at the impact of contract terms on how capacity is planned and what service is outsourced. In a newsvendor framework, they question whether the firm should outsource its base load of calls or its peak calls. Unlike Aksin et al. (2008), we consider complete outsourcing and information asymmetry regarding the distribution of demand. Aksin and Masini (2008) study insourcing and show that its effectiveness depends on the complementarities between the needs of the business environment and the capabilities developed to address these needs. In contrast, we study outsourcing between two independent parties. See Aksin et al. (2007) for a survey of the recent literature on call center operations management.

Milner and Pinker (2001) consider a firm that outsources the job of identifying capable workers to an external labor supply agency. The agency offers flexibility in an environment where both the demand and the supply of contingent labor are uncertain. As in our setting, postdemand hiring is more costly. Milner and Pinker (2001) study the effect of labor supply uncertainty on the labor contracts signed between the firm and the labor agency. The authors find that when the uncertainty is about labor productivity, a coordinating contract exists, whereas if it is about the availability of labor, then such contracts are not possible. In contrast, we study the complete outsourcing of service operations, which includes the service provider’s investment for hiring and training necessary agents.

Recently, contract design under asymmetric information in queuing settings has received more attention. Shumsky and Pinker (2003) consider the incentive issues that arise between the firm and the gatekeeper of a service, who may attempt to solve customers’ problems or may refer jobs to specialists. Hasija et al. (2008) study a variety of contracts for call center outsourcing when there is information asymmetry about worker productivity, which can be high or low. The authors show that under full information, coupling pay-per-call or pay-per-time contracts with service-level agreements can induce the call center to choose the optimal staffing level for the whole service chain. On the other hand, when the call center is privately informed about the service rate, a menu of pay-per-call contracts cannot both extract optimal staffing levels from different types and be incentive compatible. Neither can a menu of pay-per-time contracts achieve full separation. However, offering a menu consisting of a pay-per-call contract, intended for the high-productivity call center, and a pay-per-time contract, intended for the low-productivity call center, can achieve the full-information solution.

Ren and Zhou (2008) study contracting in a service chain where the outsourcing party’s revenue is from successfully resolved calls. To increase the fraction of successfully answered calls, the call center must exert effort. However, a moral hazard problem arises because exerting effort is costly and not necessarily observed by the outsourcing party. Still, the throughput and call resolution frequency can be observed and contracted upon. If the call center accepts the outsourcing firm’s contract, he chooses the staffing and effort levels according to contract terms. Ren and Zhou (2008) show that if the effort is observable and contractible, paying the call center based on the number of resolved calls and sharing the call center’s costs results in the selection of systemwide optimal staffing and effort levels. If the effort is unobservable, a partnership contract that sells the revenues to the call center, shares the call center’s costs, and collects payments based on call volume can overcome the moral hazard problem.

Ren and Zhang (2009) examine service outsourcing contracts when the service provider’s capacity cost—i.e., the cost for providing faster service—and quality cost—i.e., the cost of increasing the fraction of answered calls resolved successfully—are private information. The capacity and quality costs may be correlated. The service provider earns zero profit if he turns down the outsourcing party’s contract offer. The parties contract on capacity, call resolution frequency, and the outsourcing party’s payment. The authors solve for the optimal menu of contracts under asymmetric information and show that it distorts the systemwide optimal capacity and effort levels where the distortions depend on the correlation between the cost terms. Moreover, the outsourcing party is not able to extract the entire expected surplus from the service provider.

Cachon and Zhang (2006) also study contract design in a queuing framework where a buyer sources a good or service from a supplier whose capacity cost is private information. The buyer offers a menu of contracts to the supplier aiming to minimize the procurement plus holding and backlogging costs. The menu of contracts specifies the capacity level as well as the per-unit payment by the buyer as a function of the supplier’s reported capacity cost. The optimal menu of contracts is nonlinear in the reported capacity cost and the supplier builds less capacity than the service chain optimal levels. Hence, the full-information solution is cannot be implemented. One of the important differences between our setting and Cachon and Zhang (2006) is that in ours the service provider can offer contracts that depend on the realized demand volume because it is immediately observable after the operation starts. This and economies of scale enable a full separation of types via two-part tariffs. Because the parties also enjoy economies of scale in Cachon and
of the parties, i.e., adverse selection; see Chen (2003). The asymmetry stems from access to different information by one of the parties, i.e., moral hazard problem, is one source of information asymmetry in a supply chain and has been studied by Baiman et al. (2000) and Lim (2001) in the context of quality provision with costly effort. However, information asymmetry in the majority of articles on supply chain contracting stems from access to different information by one of the parties, i.e., adverse selection; see Chen (2003) and references therein for an overview. To cite several examples, Corbett and de Groot (2000), Corbett (2001), Ha (2001), and Lutze and Ozer (2008) focus on information asymmetry in cost, whereas Cachon and Lariviere (2001) are concerned with information asymmetry in market demand and forecasts. Cachon and Lariviere (2001) also formulate a capacity provision problem where the customer with private information on demand offers a contract to the supplier to signal her type.

Typically, the efforts to generate demand and the operational planning to satisfy it are studied separately in the literature. In a related study, Soyer and Tarimcilar (2008) advance a modulated Poisson process model to describe and analyze the arrival data to a call center of a consumer electronics producer who sells most of its products through the call center. The company promotes its products via different ads in print media and also offers various promotions such as free shipping and free installment. Soyer and Tarimcilar (2008) assess the effectiveness of different advertising strategies on increasing the call volume to the call center and find that the amount spent on the advertisements, media venue, and the specific promotion strategies all have significant effects on the call volume.

Outsourcing of inventory operations under full information is studied in several contexts in the operations management literature. See Kouvelis et al. (2006) for an overview of vendor-managed inventory (VMI) systems. Nagarajan and Rajagopalan (2008) study the performance of holding-cost subsidy contracts in improving channel performance in a VMI system. Fry et al. (2001) consider a type of VMI system called a (z, Z) contract and compare its performance with the performance of retail-managed inventory. Perhaps the most relevant paper to ours is Cachon and Harker (2002), which traces the competition between two firms under full information to economies of scale. In particular, Cachon and Harker (2002) show that scale economies intensify competition. The authors also consider the outsourcing decision of the two competing firms in a two-stage game when the supplier neither has a cost advantage over either of the outsourcing firms nor can pool demand across both firms. Scale economies lead to the surprising result that even then, there exists a set of contracts that all three firms agree to sign because outsourcing mitigates downstream competition.

3. A Unifying Model of Service Outsourcing

An originator and a service provider contract for a short-term outsourcing project. The originator captures a unit margin of \( m > 0 \) for each customer served. Before the operation starts, the demand volume \( \lambda \) is
unknown to both parties, although the originator has a better idea of it. Once service starts, the demand can quickly be observed. The contract occurs prior to service because the service provider needs to plan for training and facilities in advance. She trains personnel in the call center setting and leases warehouse space in the order fulfillment setting. Hereafter, we use the term investment to refer to the expenditure incurred for training and leasing. The level of up-front investment depends on the anticipated demand, because larger demand requires more trained personnel or more space. The investment carried out before service is not reversible. Although the service provider can train more personnel or lease more space if necessary after the demand is realized, this will be more costly than investing a priori (cf. Milner and Pinker 2001). To mitigate her share of risk from uncertain demand, the service provider offers contracts that depend on the realized demand λ.

The originator can be of two types, depending on the probability distribution λ is drawn from: high or low, indexed by H and L, respectively. The demand volume of type i takes values on the positive real line with density function function \( g_i(\cdot) \) and distribution function \( G_i(\cdot) \) for \( i = L, H \). We assume that the originator knows his market better than the service provider, and, in particular, he knows his type at the time of contract signing. The service provider believes the originator to be type i with probability \( q_i \in (0, 1) \) for \( i = L, H \). A high-type originator anticipates a larger volume. More specifically, we assume the following:

**Assumption 1.** \( \lambda_H \) is equal in distribution to \( a\lambda_L \) for some constant \( a > 1 \), i.e., \( G_H(\lambda) = G_L(\lambda/a) \) for \( \lambda \geq 0 \).

Assumption 1 implies that the volume \( \lambda_H \) of the high-type first order stochastically dominates volume \( \lambda_L \) of the low type, that is, we have \( G_i(\lambda) \geq G_{H}(\lambda) \) for \( \lambda \geq 0 \). Although Assumption 1 is sufficient for the results, it is not necessary. An alternative sufficient condition will be discussed in the next section.

The information asymmetry is significant for the service provider’s investment decision because it depends on the anticipated demand. Let \( N(\lambda) \) denote the level of investment required to satisfy a demand volume of \( \lambda \). We assume that \( N(\lambda) \) is strictly increasing in \( \lambda \), where \( N(0) = 0 \).

Note that due to shift-based staffing and the outsourcing project’s extension to multiple shifts, the personnel trained \( N(\lambda) \) is in general substantially larger than the number of agents utilized in any given shift. Likewise, in the order fulfillment setting, the space leased \( N(\lambda) \) is much larger than the storage room for the physical items, whereas the rest of the space is allocated to equipment and offices. Moreover, as \( \lambda \) increases, the relative effect of uncertainty diminishes. This corresponds to the diminishing effect of unforeseen personnel turnover, illnesses, shift management difficulties in the call center setting, and the reduced need for auxiliary space in the order fulfillment setting. Thus, we assume that \( N(\lambda) \) is concave in \( \lambda \).

A natural form for the level of investment \( N(\lambda) \) needed to satisfy \( \lambda \) is \( N(\lambda) = \phi_1(\lambda + \phi_2(\lambda)) \), where \( \phi_2(\lambda) \) is increasing concave and satisfies \( \phi_2(b\lambda) \leq \tau(b)\phi_2(\lambda) \) with \( \tau(b) \leq b \) and \( b > 1 \). This functional form lends itself to an intuitive interpretation: \( \phi_1(\lambda) \) is the nominal level of investment to manage \( \lambda \), whereas \( \phi_2(\lambda) \) is the investment needed in excess of nominal requirements due to uncertainty. If \( \lambda \) increases by a factor of \( b > 1 \), the level of investment in excess of nominal requirements increases by \( \tau(b) \leq b \) because the relative effect of uncertainty diminishes with scale.

The service provider faces a newsvendor-like problem in choosing the level of investment prior to operation. Personnel can be trained (or space can be leased) at a lower unit cost before the operation starts. Because the service provider learns demand once the operation starts, it is feasible but more costly to train additional personnel on short notice because this may require new hirings, outside help, and schedule changes. Let \( \ell \) denote the cost per unit investment before service is initiated, whereas \( \ell' > \ell \) is the cost per unit investment after the service is initiated. Although investment is a lump-sum payment, it can easily be converted to a cost rate over the outsourcing project horizon. Without loss of generality, we adopt this interpretation. Investment is sunk if it exceeds the level necessitated by the realized demand. Therefore, it is also risky to train too many personnel (or lease too much space) before the demand is realized.

The service provider’s eventual investment depends on the realized demand and the investment carried out before the start of operation. To describe the service provider’s planning problem before the service commences, let \( \psi(\lambda, \bar{\lambda}) \) denote the investment she incurs if she anticipates a demand \( \bar{\lambda} \) and invests accordingly before the service commences, whereas the demand turns out to be \( \lambda \), i.e.,

\[
\psi(\bar{\lambda}, \lambda) = I[N(\bar{\lambda}) + \ell(N(\lambda) - N(\bar{\lambda}))^+],
\]

where \( (x)^+ = \max\{x, 0\} \). If \( \lambda \leq \bar{\lambda} \), then the service provider can handle it without additional investment. However, if the service provider plans for a small demand volume \( \lambda < \bar{\lambda} \), then she needs to train (lease) \( N(\lambda) - N(\bar{\lambda}) \) additional personnel (space) at the higher unit cost \( \ell' \).

Before the operation starts, the service provider chooses an investment level to minimize her expected investment. From her perspective, choosing an investment level of \( N(\lambda) \) is equivalent to choosing an anticipated demand volume \( \lambda \). If she expects the originator to be of type \( i \), conditional on the originator’s type,
the service provider chooses to plan for a demand volume $\lambda$ to

$$\text{minimize } E[\psi(\lambda, \bar{\lambda})]. \quad (2)$$

The following proposition describes the optimal investment decision. It also compares the minimum expected investments to serve the two types of the originator. Proposition 1 will be instrumental in solving the mechanism design problem and its proof is provided in the appendix.

**Proposition 1.** The optimal level of investment for serving a type $i$ originator is $N(\lambda^*_i)$, where

$$\lambda^*_i := G^{-1}_i(1 - \frac{1}{t}) \quad \text{for } i = L, H. \quad (3)$$

Moreover, under Assumption 1, the minimum expected investment for serving a type $i$ originator defined as $\Psi_i := E[\psi(\lambda^*_i, \lambda_i)]$ for $i = L, H$ satisfies

$$\Psi_L \leq \Psi_H \quad \text{and} \quad \Psi_H / E[\lambda_H] \leq \Psi_L / E[\lambda_L].$$

The demand volume $\lambda^*_i$ that can be satisfied with $N(\lambda^*_i)$ corresponds to the critical fractile formula in a newsvendor solution: the underage cost is the extra per-unit investment $t - \bar{t}$ to pay for further investment needed on short notice after the demand is realized. An underage cost equal to $\bar{t}$ is incurred for the excessive investment that cannot be reversed. Proposition 1 proves that a larger expected investment is required to serve the high type, whereas the expected investment per volume is less for the high type, hence exhibiting economies of scale. Finally, the optimal investment depends on the anticipated $\bar{\lambda}$. As a result, knowing the true distribution of demand reduces the service provider’s expected investment, motivating the service provider to distinguish, or screen, the originator types.

Once the demand volume $\lambda$ is realized, the service provider incurs an operating cost rate $c(\alpha, \lambda)$ to deliver a service level $\alpha$ given $\lambda$. We assume that

$$c(\alpha, \lambda) = c_1 \lambda + c_2 \sqrt{\lambda} \beta(\alpha, \lambda), \quad (4)$$

where $c_1, c_2 \geq 0$ are constants. This cost structure arises naturally in both settings we consider. In the call center setting, $\alpha$ denotes the fraction of calls answered (equivalently, $1 - \alpha$ is the fraction of calls abandoned), and $c(\alpha, \lambda)$ is the staffing cost. In the order fulfillment example, $\alpha$ denotes the fill rate and $c(\alpha, \lambda)$ is the inventory holding cost. Note that the number of agents (item storage room) utilized at any given time is potentially much lower than the total number of trained personnel (total leased space). The linear term $c_1 \lambda$ captures staffing (or holding) cost for the nominal demand ignoring variability, whereas $c_2 \sqrt{\lambda} \beta(\alpha, \lambda)$ is the staffing cost of the safety capacity (or the holding cost for the safety stock). We interpret $\beta(\alpha, \lambda)$ as the standardized excess capacity and assume it is convex and strictly increasing in $\alpha$. Notice that if $\beta(\alpha, \lambda)$ does not depend on $\lambda$, the average cost $c(\alpha, \lambda) / \lambda$ is decreasing in $\lambda$, i.e., it exhibits economies of scale. As the readers will see, this is indeed the case in the full information solution, cf. Proposition 2.

The originator also incurs a cost rate $\kappa(\alpha, \lambda)$ of customer dissatisfaction if the service level is $\alpha$ and demand is $\lambda$. In the call center setting, $\kappa(\alpha, \lambda)$ corresponds to the lost revenue from abandoning consumers. In an order fulfillment setting, it corresponds to the cost of backlogging orders. We assume that

$$\kappa(\alpha, \lambda) = \kappa_1 \lambda + \kappa_2 \sqrt{\alpha} f(\beta(\alpha, \lambda)), \quad (5)$$

where $\kappa_1, \kappa_2 \geq 0$ are constants, $f(\cdot)$ is a decreasing convex function, and $\beta(\alpha, \lambda)$ is as in (4). To simplify the analysis, we assume $f(\cdot)$ is a function only of $\beta(\alpha, \lambda)$ and this turns out to be correct in the two applications we consider. As with $c(\alpha, \lambda)$, the form of $\kappa(\alpha, \lambda)$ can be justified for both applications; further, $\kappa$ too exhibits economies of scale.

**The Originator’s Outside Option.** The originator’s outside option is given by his expected profit from conducting operations in house. The originator incurs a fixed cost $C$ of starting up the operations, which we convert to a cost rate over the project horizon. We assume that both types of the originator use the same technology. The originator is subject to the same cost $t$ per unit of investment before the operation starts and cost $\bar{t}$ after the operation starts as the service provider. Our results continue to hold if the originator incurs a higher investment compared to the service provider. Hence, a type $i$ originator incurs an expected investment equal to $\Psi_i$, cf. Proposition 1. We assume that the originator is less efficient in his operations. He incurs an operating cost equal to

$$\bar{c}(\alpha, \lambda) = \bar{c}_1 \lambda + \bar{c}_2 \sqrt{\alpha} \beta(\alpha, \lambda), \quad (6)$$

where $\bar{c}_1 > c_1$ and $\bar{c}_2 > c_2$ implying $\bar{c}(\alpha, \lambda) > c(\alpha, \lambda)$ for each $\alpha, \lambda$. Once the service center is established and the demand is learned, the originator chooses the service level $\bar{\alpha}$ to maximize his revenue minus the operating cost. That is, given $\lambda$, a type $i$ originator chooses $\alpha$ to maximize

$$m\lambda - \kappa(\alpha, \lambda) - \bar{c}(\alpha, \lambda), \quad (7)$$

because the fixed cost and investment are already sunk. Using (5)–(6), the service level that maximizes (7) is given by the first-order condition

$$-\kappa_2 f'(\beta(\alpha, \lambda)) \frac{\partial \beta(\alpha, \lambda)}{\partial \alpha} = \bar{c}_2 \frac{\partial \beta(\alpha, \lambda)}{\partial \alpha}, \quad (8)$$

which is not only necessary but also sufficient because $f(\beta(\alpha, \lambda))$ and $\beta(\alpha, \lambda)$ are both convex in $\alpha$. Because $\partial \beta(\alpha, \lambda)/\partial \alpha > 0$ for all $\alpha, \lambda$, we can simplify (8) to write
\[ f'(\beta(\alpha, \lambda)) = -\tilde{c}_2/\kappa_2. \] Thus, for each \( \lambda \), both originator types set \( \tilde{\alpha}(\lambda) \) such that
\[
\tilde{B} := \beta(\tilde{\alpha}(\lambda), \lambda) = (f')^{-1}(\frac{-\tilde{c}_2}{\kappa_2}),
\] (9)
where \((f')^{-1}\) is the inverse of \( f' \). The service level \( \tilde{\alpha}(\lambda) \) is efficient: it compensates the marginal cost \( \delta c(\alpha, \lambda)/\delta \tilde{\alpha} \) of increasing the service level with the reduction \( \delta \alpha(\alpha, \lambda)/\delta \alpha \) in the cost of customer dissatisfaction. Given \( \tilde{\alpha}(\lambda) \), the originator’s profit for a specific \( \lambda \) is equal to
\[
(m - c - \kappa_1)\lambda - (c\tilde{B} + \kappa_2 f(\tilde{B}))\sqrt{\lambda} - \psi(\lambda^*, \lambda) - C. \tag{10}
\]
Therefore, the outside option of a type \( i \) originator is given by
\[
K_i := (m - c - \kappa_1)E[\lambda_i] - (c\tilde{B} + \kappa_2 f(\tilde{B}))E[\sqrt{\lambda_i}] - \Psi_i - C
\]
for \( i = L, H \), (11)
because \( \Psi_i = E[\psi(\lambda^*, \lambda)] \) by Proposition 1. Note that the outside options of the two types are different.

**Mechanism Design Problem.** The service provider, the uninformed party, offers a menu of contracts to the originator, the informed party, to distinguish, or screen, an originator who anticipates a high volume from one who anticipates a low volume. In what follows, we formulate the problem of the service provider as a mechanism design problem and focus on contracts of the form \((\alpha, p)\), where \( \alpha \) is the agreed service level and \( p \) is the fee paid by the originator. After the contract is signed and the operation starts, the service provider can quickly deduce the true demand volume. The service provider can adjust staffing and inventory levels to fulfill her contractual obligation. Accordingly, the service provider can offer a contract where \( \alpha \) and \( p \) depend on the type reported by the originator as well as the realized demand \( \lambda \). That is, the service provider chooses a menu \((\alpha_i(\lambda), p_i(\lambda))\) to
\[
\text{maximize} \quad \sum_{i=L, H} q_i(E[p_i(\lambda_i) - c_i \lambda_i]
\]
subject to
\[
E[(m - c - \kappa_1)\lambda_i - \kappa_2 \sqrt{\lambda_i} f(\beta(\alpha(\lambda), \lambda)) - p_i(\lambda)] \geq K_i
\]
for \( i = L, H \), (IR)
\[
E[(m - c - \kappa_1)\lambda_i - \kappa_2 \sqrt{\lambda_i} f(\beta(\alpha(\lambda), \lambda)) - p_i(\lambda)]
\geq E[(m - c - \kappa_1)\lambda_i - \kappa_2 \sqrt{\lambda_i} f(\beta(\alpha(\lambda), \lambda)) - p_i(\lambda)]
\]
for \( i \neq j \), (IC)
where the minimum expected investment \( \Psi_i \) to serve type \( i \) is given by Proposition 1, and the outside option \( K_i \) of type \( i \) is given by (11), for \( i = L, H \).

The service provider’s profit from serving type \( i \) under demand \( \lambda \) is given by the payment \( p_i(\lambda) \) she receives minus the sum of the operating cost to support \( \alpha_i(\lambda) \) and the investment to meet \( \lambda \). Therefore, the objective function of the service provider maximizes expected profits, where the expectation is taken with respect to the service provider’s prior belief about originator’s type and subsequently with respect to the demand distribution given the originator’s type. The first constraint imposes individual rationality (IR), or participation. It ensures that each type of originator prefers the contract to his outside option. The second constraint imposes incentive compatibility (IC). The constraint (IC) ensures that the originator of type \( i \) \((i = L, H)\) prefers the contract devised for him over the contract devised for the other type. When both IR and IC constraints are satisfied, each type of originator self-selects the contract intended for him. The next section solves the mechanism design problem.

### 4. Solution to the Mechanism Design Problem
We first consider the full-information case where the service provider can observe the type of the originator. Under full information, the IC constraints can be ignored. The following proposition provides the full-information solution, which is an upper bound on the profit that can be obtained under asymmetric information. The proof of Proposition 2 is provided in the appendix.

**Proposition 2.** The optimal standardized excess capacity under full information is given by \( \beta^* := (f')^{-1}(\frac{-\tilde{c}_2}{\kappa_2}) \). Moreover, for each \( \lambda \), the optimal service level \( \alpha^*(\lambda) \) under full information is given by the unique solution of
\[
\beta^* = \beta(\alpha^*(\lambda), \lambda), \tag{12}
\]
and the optimal payments \( p^*_i(\lambda), p^*_j(\lambda) \) for \( i \neq j \) must satisfy
\[
E[p^*_i(\lambda)] = (m - \kappa_1)E[\lambda_i] - \kappa_2 f(\beta^*)E[\sqrt{\lambda_i}] - K_i
\]
for \( i = L, H \). (13)

The full-information solution is efficient in the sense that the marginal benefit of increasing the service level is equal to its marginal cost. Therefore, under full information, both types receive the same service level for each \( \lambda \). Also, \( \beta^* \) exceeds the optimal standardized capacity \( \tilde{\beta} \) chosen by the originator if he did the work in-house, cf. (9). This follows from \( \tilde{c}_2 > c_2 \), i.e., the service provider’s cost advantage over the originator. Along the same lines, \( \alpha^*(\lambda) \) is higher than the optimal service level \( \tilde{\alpha}(\lambda) \) of the originator were he to establish his own operations for each \( \lambda \) because \( \beta^* > \tilde{\beta} \) and \( \beta(\alpha, \lambda) \) is increasing in \( \alpha \).
If a menu of contracts generates the same expected profit as the full-information solution and satisfies the IR and IC constraints, it is indeed optimal. In what follows, we guess the optimal solution to the mechanism design problem as a menu of two-part tariffs and then verify its optimality. Specifically, we show that it is feasible and achieves the expected profit of the full-information solution. A two-part tariff is a payment plan consisting of a fixed fee and a per-volume fee. In the menu of two-part tariffs we propose, the low type pays only a per-volume fee, whereas the high-type’s payment consists only of a fixed fee.

**Theorem 1.** The menu of two-part tariffs that consists of service levels \(\alpha^*(\lambda)\) for \(\lambda \geq 0\) and payment schedule

\[
p_i(\lambda) = \lambda \frac{(m - \kappa_i)E[\lambda_i] - \kappa_2 f(\beta^*)\sqrt{\lambda_i} - K_i}{E[\lambda_i]} \tag{14}
\]

and

\[
p_H(\lambda) = (m - \kappa_i)E[\lambda_H] - \kappa_2 f(\beta^*)\sqrt{\lambda_H} - K_H \tag{15}
\]

is optimal for the mechanism design problem.

**Proof of Theorem 1.** We first show that the menu of two-part tariffs enables the service provider to extract the entire expected surplus from the originator and yields the same expected profit as the full-information solution. Second, we verify that the menu of two-part tariffs satisfies IR and IC constraints, and hence it is optimal over all possible mechanisms.

By Proposition 2, under full information, the service level is set at \(\alpha^*(\lambda)\) for \(\lambda \geq 0\) for each type. Also by Proposition 2, the IR constraints bind, and the optimal payments \(p'_i(\lambda), p_H(\lambda)\) satisfy (12). This implies that the expected revenue of the service provider under full information is equal to

\[
\sum_{i=L, H} q_i \left[ (m - \kappa_i)E[\lambda_i] - \kappa_2 f(\beta^*)\sqrt{\lambda_i} - K_i - \Psi - C \right]. \tag{16}
\]

Under (14)-(15), the expected payment of type \(i\) is equal to \((m - \kappa_i)E[\lambda_i] - \kappa_2 f(\beta^*)\sqrt{\lambda_i} - K_i\) for \(i = L, H\), which implies that the expected surplus of the originator is extracted.

Moreover, given \(\lambda\), the service provider inverts an investment \(\Psi(\lambda, \lambda)\) and an operating cost \(c(\alpha^*(\lambda), \lambda) = c_1 \lambda + c_2 \beta^* \sqrt{\lambda}\) to provide the originator with \(\alpha^*(\lambda)\). Because type \(i\) happens with probability \(q_i\), the expected profit from the menu of two-part tariffs is equal to the expected profit under full information, cf. (16). That is, the menu of two-part tariffs achieves the upper bound on the objective function value of the mechanism design problem.

It suffices to check that the menu of two-part tariffs is feasible. Recall that type \(i\) originator’s expected payment is \((m - \kappa_i)E[\lambda_i] - \kappa_2 f(\beta^*)\sqrt{\lambda_i} - K_i\), implying that (IR) holds with equality for \(i = L, H\). We need to check that (IC) is also satisfied for \(i = L, H\). Because \(\alpha_L(\lambda) = \alpha_H(\lambda) = \alpha^*(\lambda)\) for \(\lambda \geq 0\), we can rewrite the IC constraints as follows:

\[
E[p_i(\lambda_i)] \leq E[p_j(\lambda_j)] \quad \text{for} \quad i, j \in \{L, H\} \quad \text{and} \quad i \neq j.
\]

That is, the IC constraints are satisfied if and only if the expected payment of type \(i\) under the contract designed for him is less than the expected amount he would pay if he pretended to be the other type. The IC constraint may be simplified using (14)-(15):

\[
(m - \kappa_i)E[\lambda_i] - \kappa_2 f(\beta^*)\sqrt{\lambda_i} - K_i - \Psi - C \leq (m - \kappa_i)E[\lambda_H] - \kappa_2 f(\beta^*)\sqrt{\lambda_H} - K_H,
\]

and

\[
(m - \kappa_i)E[\lambda_H] - \kappa_2 f(\beta^*)\sqrt{\lambda_H} - K_H \leq (m - \kappa_i)E[\lambda_i] - \kappa_2 f(\beta^*)\sqrt{\lambda_i} - K_i,
\]

Recall that the outside option \(K_i\) of type \(i\) is given by

\[
K_i = (m - c_1 - \kappa_i)E[\lambda_i] - (c_2 \beta + \kappa_2 f(\beta^*))\sqrt{\lambda_i} - \Psi - C.
\]

cf. (11), where \(\tilde{\beta} = (f')^{-1}(\tilde{c_2} / \kappa_2)\). Moreover, \(\beta^* \geq \tilde{\beta}\), and hence, \(f(\beta^*) \leq f(\beta)\) because \(f(\cdot)\) is decreasing. Substituting for \(K_i\) and \(K_H\) in (IC) and (IC\(j\)) and arranging terms, we obtain the following expressions for the IC constraints:

\[
\begin{align*}
c_1 (E[\lambda_i] - E[\lambda_H]) + \Psi - \Psi_H & \leq [\kappa_2 f(\beta) - f(\beta^*)] + c_2 \sqrt{\lambda_i} (E[\sqrt{\lambda_i}] - E[\sqrt{\lambda_H}]), \tag{IC\(i\)}
\end{align*}
\]

and

\[
\frac{C + \Psi_H}{E[\lambda_H]} + \frac{C + \Psi}{E[\lambda_i]} \leq [\kappa_2 f(\beta) - f(\beta^*)] + c_2 \beta \left( \frac{E[\sqrt{\lambda_i}]}{E[\lambda_i]} - \frac{E[\sqrt{\lambda_H}]}{E[\lambda_H]} \right). \tag{IC\(H\)}
\]

Using integration by parts, we can write \(E[\lambda_i] = \int_0^{\infty} (1 - G_i(\lambda)) \, d\lambda\). Therefore,

\[
E[\lambda_H] = \int_0^{\infty} (1 - G_H(\lambda)) \, d\lambda
\]

\[
= a \int_0^{\infty} (1 - G_L(\lambda)) \, d\lambda = a E[\lambda_L],
\]

where the second equality is true because \(G_H(\lambda) = G_L(\lambda / a)\) from Assumption 1. Thus, \(E[\lambda_i] \leq E[\lambda_H]\). Similarly, \(E[\sqrt{\lambda_H}] = \sqrt{a} E[\sqrt{\lambda_L}]\), and hence, \(E[\sqrt{\lambda_i}] \leq E[\sqrt{\lambda_H}]\). The right-hand side of (IC\(j\)) is positive because
f(\beta) \geq f(\beta^*)$. The left-hand side is negative because $\Psi_l \leq \Psi_H$ from Proposition 1. Thus, (IC$_s$) is satisfied by (14)–(15).

Because $\Psi_H/E[\lambda_{HI}] \leq \Psi_I/E[\lambda_I]$, cf. Proposition 1, and $E[\lambda_H] = aE[\lambda_I]$, the left-hand side of (IC$_s$) is negative. Further, $E[\sqrt{\lambda_{HI}}] = \sqrt{a}E[\sqrt{\lambda_I}]$ implies $E[\sqrt{\lambda_{HI}}]/E[\lambda_{HI}] \leq E[\sqrt{\lambda_I}]/E[\lambda_I]$, and the right-hand side is positive because $f(\beta) \geq f(\beta^*)$. This proves that (IC$_s$) is satisfied as well.

The menu of two-part tariffs achieves the full-information solution because of two features of the contracting environment. First, the demand can be observed and contracted upon; second, the service operations we consider exhibit economies of scale. To elaborate on the role of contingent contracts, recall that in Cachon and Zhang (2006), the parties enjoy economies of scale but the capacity cost of the supplier is privately known but never revealed, leading to a postcontract inefficiency. Therefore, their work suggests that contracting on demand should be necessary for the menu of two-part tariffs to achieve the full-information solution in our setting. The role of economies of scale is to render the menu of two-part tariffs incentive compatible. Economies of scale imply that not only is the outside option, but also the outside option per demand volume, is greater for a high type than that for the low type. Consequently, the high type is willing to pay less on a per-volume basis but more in aggregate. The high-type’s deviation is deterred by the linear payment schedule designed for the low type, whose per-volume fee is greater than the low-type’s lump-sum payment divided by $E[\lambda_{HI}]$. The low-type’s deviation is prevented by the large fixed payment charged to the high type. Notice that if the high type were willing to pay more both in aggregate and on a per-volume basis, a menu of two-part tariffs would not be incentive compatible because the high type would pretend to be the low type, thus enjoying a lower per-volume fee.

Finally, Assumption 1 is sufficient, but not necessary, for the results. An alternative set of sufficient conditions is as follows: $\lambda_I$ is less than $\lambda_{HI}$ in the hazard-rate order; $\lambda_L$ is greater than $\lambda_{HI}$ in the star order; and

$$G^{-1}_{HI} \left(1 - \frac{1}{\beta}\right)E[\lambda_I] \leq G^{-1}_{L} \left(1 - \frac{1}{\beta}\right)E[\lambda_{HI}].$$

We refer the reader to Shaked and Shanthikumar (1994) for properties of the hazard-rate and star orderings. Notice that Assumption 1 neither implies nor is implied by these latter conditions. However, $\lambda_I$ and $\lambda_{HI}$ satisfy these latter conditions if they satisfy Assumption 1, and in addition, $\lambda_L$ has an increasing generalized failure rate.

In the next two sections, we demonstrate that call center and order fulfillment applications can both be analyzed within the unifying model framework.

5. Call Center Outsourcing

In this section, an originator outsources his call center operations to a call center manager (service provider). The calls generate a revenue of which the originator captures a margin $m > 0$. As stated in §3, the arrival rate $\lambda$ is uncertain at the time of contracting. We maintain the assumptions of the unifying model, cf. §3, on the information asymmetry. As discussed above, $\lambda$ may be uncertain at the time of contracting because consumer response to marketing effort is unknown. Information asymmetry could then arise if the service provider is unable to monitor all of the originator’s marketing activities. The service provider’s training needs and the ensuing expenses are given by the investment formulation of the unifying model, cf. §3. For brevity, we do not repeat those here. Instead, we next provide a model of call center operations.

A Queuing Model of the Call Center Operations.

The call center is modeled as a multiserver queue with a Poisson arrival process and exponentially distributed service times. The arrival rate is $\lambda$, which for the moment is assumed to be commonly known. Without loss of generality, the mean service time is one. Each pending customer may abandon. Time-to-abandon is exponentially distributed with rate $\eta$. In other words, using the standard terminology, we model the call center as an $M/M/N + M$ queue.

Assuming $\lambda$ is large, i.e., $\lambda \gg 1$, the approximation results of Garnett et al. (2002) for large call centers accurately capture the queuing dynamics in our setting. Besides developing approximations of various performance metrics for large call centers, Garnett et al. (2002) also argue that basic economics requires managing such large call centers in a quality-and-efficiency-driven regime. The economic optimality of this operating regime is justified by Maglaras and Zeevi (2003). Also, see Randhawa and Kumar (2010) for an analysis of the optimal capacity sizing under nonlinear delay costs. Although deriving different staffing rules, the optimal safety capacity in Randhawa and Kumar (2010) is still sublinear in $\lambda$, and hence, it exhibits (statistical) economies of scale, which is essentially what is needed for our results.

The quality-and-efficiency-driven regime leads to staffing decisions based on a square-root rule, which dictates the number of call center agents $S$ to be equal to

$$S = \lambda + \beta \sqrt{\lambda}. \quad (17)$$

Given Poisson arrivals, $\sqrt{\lambda}$ is the standard deviation of demand arriving per unit of time. Then, $\beta$ refers to the system’s excess capacity measured in units of the standard deviation of demand per unit of time, i.e., the standardized excess capacity.

We borrow the following approximation from Garnett et al. (2002) as a tractable model of call center
operations. We assume that the abandonment probability is given by
\[ P(\text{Abandonment}) = \frac{\Delta(\beta)}{\sqrt{\lambda}}, \] (18)
which provides a good approximation for moderate-to-large values of \( \lambda \). To be more specific, Garnett et al. (2002) proposes the approximation
\[ P(\text{Abandonment}) = \Delta(\beta)/\sqrt{\lambda}, \]
which is equivalent to the approximation (18) asymptotically, and both are justified through the same limiting argument.

Let \( \Phi \) and \( \Phi \) denote the standard normal probability density and cumulative distribution functions, respectively. Garnett et al. (2002) provides the following explicit formula for \( \Delta(\cdot) \) in terms of the hazard-rate function \( h(x) := \phi(x)/\Phi(-x) \) of a standard normal random variable,
\[ \Delta(\beta) = \left[ \sqrt{\eta} h(\beta/\sqrt{\eta}) - \beta \right] \left[ 1 + \frac{\sqrt{\eta} h(\beta/\sqrt{\eta})}{h(-\beta)} \right]^{-1}. \]
The relevant feature of \( \Delta(\cdot) \) for our analysis is that \( \Delta(\beta) \) is convex decreasing for \( \beta \geq 0 \), which is precisely the regime of interest for revenue-generating call centers.

The approximation (18) captures economies of scale in large call centers. To elaborate, define the standardized excess capacity \( \beta(\alpha, \lambda) \) needed to answer a fraction \( \alpha \) of the calls for a given arrival rate \( \lambda \) as follows:
\[ \beta(\alpha, \lambda) = \Delta^{-1}((1 - \alpha)\sqrt{\lambda}), \] (19)
where \( \Delta^{-1}(\cdot) \) is the inverse of \( \Delta(\cdot) \). Then, given \( \alpha \), the corresponding standardized excess capacity \( \beta(\alpha, \lambda) \) is decreasing in the arrival rate \( \lambda \). Moreover, the square-root staffing rule, cf. (17), implies that the number of agents staffed per arrival rate \( \lambda \) to answer a fraction \( \alpha \) of calls is also decreasing provided that we fix \( \beta(\alpha, \lambda) \) for each \( \lambda \). In other words, a larger call center is more efficient and provides better service for a given level of standardized excess capacity, i.e., economies of scale.

**Contracting for Call Center Outsourcing.** Next, we describe the results for call center outsourcing, which simply follows from specializing the results of §4 to the contracting problem in this context. Learning the true distribution of \( \lambda \) lowers the expected training cost for the call center manager. The call center manager offers a menu of contracts of the form \((\alpha, p)\), where \( \alpha \) is the fraction of answered calls and \( p \) is the fee paid by the originator. Both \( \alpha \) and \( p \) depend on the reported type and realized \( \lambda \). The parties do not explicitly contract on the number of agents, but on \( \alpha \), which is a common feature of trade (cf. Hasija et al. 2008). Once service is established, the call center manager can rapidly infer \( \lambda \) and adjust his staffing level to answer \( \alpha \) of calls. The outside option of the originator is his expected profit from establishing his own call center by incurring a higher staffing cost rate \( \bar{k} \) than the call center manager’s staffing cost rate \( k \).

The call center manager’s problem is a special case of the formulation in §3. In particular,
\[ c(\alpha, \lambda) = k(\lambda + \sqrt{\lambda}\beta(\alpha, \lambda)) \]
and
\[ \kappa(\alpha, \lambda) = m(1 - \alpha)\lambda. \] (20)

Namely, the call center manager incurs a staffing cost rate of \( c(\alpha, \lambda) = k(\lambda + \sqrt{\lambda}\beta(\alpha, \lambda)) \) to answer \( \alpha \) fraction of calls given \( \lambda \). The originator incurs the cost rate of \( \kappa(\alpha, \lambda) = m\lambda(1 - \alpha) \) because of revenue loss from abandoning customers. Thus, the mechanism design formulation of the call center manager is a special case of that advanced in §3 with \( c_1 = c_2 = \bar{k}, c_1 = \bar{k}, \kappa_1 = 0, \kappa_2 = m, \) and, finally, \( f(\cdot) = \Delta(\cdot) \).

Note that \( c \) and \( \kappa \) in (20) satisfy the assumptions of the unifying model. Customer abandonment cost \( \kappa(\alpha, \lambda) \) is linearly decreasing in \( \alpha \). Keeping \( \lambda \) fixed, staffing cost \( c(\alpha, \lambda) \) is increasing convex in \( \alpha \) because \( \beta(\alpha, \lambda) = \Delta^{-1}((1 - \alpha)\sqrt{\lambda}) \) and \( \Delta \) is decreasing convex. Then, applying Theorem 1 to the call center outsourcing problem establishes the optimality of a menu of two-part tariffs.

Under Theorem 1, the call center manager chooses the same standardized excess capacity \( \beta^* \) for both types. The originator’s callers thus have a higher service level for high arrival rates. Moreover, \( \beta^* > \bar{\beta} \), where \( \bar{\beta} = (\Delta)^{-1}(-\bar{k}/m) \) is the standardized excess capacity that would be chosen by the originator on his own. As a consequence, the corresponding service levels satisfy \( \alpha^*(\lambda) \geq \bar{\alpha}(\lambda) \) for all \( \lambda \geq 0 \), i.e., the originator achieves a higher service level under the optimal contract than would be attained if he were to establish his own call center. Finally, note that for a fixed \( \alpha \), both \( c(\alpha, \lambda)/\lambda \) and \( \kappa(\alpha, \lambda)/\lambda \) are decreasing in \( \lambda \), exhibiting economies of scale.

In summary, the results of §4 yield that a menu of two-part tariffs is the optimal contract thanks to economies of scale inherent in certain call center operations and the ability of the service provider to contract on the realized demand.

### 6. Outsourcing of Inventory Operations

In this section, an originator outsources his order fulfillment operations to an inventory manager (service provider). The originator captures a margin \( m > 0 \) from the sale of a single good. As before, the demand arrival rate \( \lambda \) is uncertain when the contract is signed. We also maintain the previous assumptions on the information asymmetry. To run the originator’s order fulfillment operation, the inventory manager needs to lease warehouse space, which is formulated by the investment
modeling of §3. For brevity, those will not be repeated here. Instead, we next provide a model of order fulfillment operations.

An Inventory Model of Order Fulfillment Operations. The inventory manager controls a single item, and she faces a constant leadtime of \( L \) when she places an order with her supplier. The demand for the good is a Poisson process with rate \( \lambda \). The inventory manager follows a base-stock policy with base-stock level \( S \). The unmet demand is backlogged. The long-run fraction of demand satisfied from stock (i.e., not backlogged) is denoted by \( \alpha \). In other words, \( \alpha \) is the fill rate, and \( 1 - \alpha \) is the fraction of backordered demand. The other performance measures of interest are the average backorders \( \bar{B} \) and the average inventory \( \bar{I} \). Given a service-level target, i.e., a lower bound on the fill rate, the service provider sets the base-stock level \( S \) large enough to meet the service-level target, but otherwise as small as possible.

Assuming \( \lambda \) is large, i.e., \( \lambda \gg 1 \), the normal approximation of §6 of Zipkin (2000) accurately captures the characteristics of supply and demand in this setting. As a tractable model of inventory operations, we borrow the following approximation of Zipkin (2000), which is based on the approximation of a Poisson distribution with a large mean by a normal distribution. The standard normal loss function is given by

\[
\mathcal{L}(z) = \int_{-\infty}^{\infty} (x - z) \phi(x) \, dx = \Phi(z) - z(1 - \Phi(z)),
\]

where \( \phi \) and \( \Phi \) denote the standard normal probability density and cumulative distribution functions, respectively. Also note that \( L(-x) = L(x) + x \) for all \( x \).

Then, the fill rate \( \alpha \), average inventory \( \bar{I} \), and average backorders \( \bar{B} \) are given by

\[
\begin{align*}
\alpha &= 1 - \Phi(z), \\
\bar{I} &= \mathcal{L}(-z) \sqrt{\lambda L}, \quad \text{and} \\
\bar{B} &= \mathcal{L}(z) \sqrt{\lambda L}, \quad \text{(21)}
\end{align*}
\]

where \( z \) denotes the standardized value of the base-stock level \( S \), i.e., \( z = (S - \lambda L) / \sqrt{\lambda L} \). It is common to interpret \( z \) as the safety factor, which has a one-to-one correspondence with fill rate \( \alpha \) through the identity \( \alpha = 1 - \Phi(z) \).

The approximation (21) leads to stocking decisions based on a square-root rule. More specifically, given a service-level requirement that the fill rate exceeds \( \alpha \), the inventory manager sets the base-stock level equal to

\[
S(\alpha, \lambda) = \lambda L + z(\alpha) \sqrt{\lambda L},
\]

where the safety factor is given by \( z(\alpha) = \Phi^{-1}(1 - \alpha) \). Given Poisson arrivals, \( \sqrt{\lambda L} \) is the standard deviation of the demand during lead time. Then, \( z(\alpha) \) is the system’s excess inventory, which is measured in units of \( \sqrt{\lambda L} \). Also note that the approximation (21) highlights the following: the average number of backlogs \( \bar{B} \) is decreasing and convex in \( S \), whereas the average inventory level \( \bar{I} \) is increasing and convex in \( S \). A larger inventory system is more efficient and can provide better service for a given level of safety factor: \( S(\alpha, \lambda) / \lambda \) is decreasing in \( \lambda \).

Contracting for Order Fulfillment Outsourcing. Next, we describe the results for outsourcing order fulfillment operations, which follow directly from specializing the results of §4 to the contracting problem in this context. The inventory manager offers a menu of contracts of the form \( (\alpha, p) \), where \( \alpha \) is the agreed-upon fill rate and \( p \) is the fee paid by the originator. As before, \( \alpha \) and \( p \) depend on realized \( \lambda \) and the reported type. After the arrival rate is observed, the inventory manager can adjust her inventory level to provide the agreed-upon fill rate. She incurs a physical holding cost \( h \) to hold one unit in the warehouse per unit of time. Financial holding cost is incurred by the originator, who owns the inventory. Each backlogged item results in a penalty of \( b \) per time unit, which the originator incurs due to loss of customer goodwill toward the outsourcing party. The originator’s outside option is his expected profit from conducting his own operations by incurring a higher per-unit holding cost \( h > h \).

The inventory manager’s problem is a special case of the formulation in §3, where

\[
c(\alpha, \lambda) = h \mathcal{L}(-z(\alpha)) \sqrt{\lambda L}, \quad \kappa(\alpha, \lambda) = b \mathcal{L}(z(\alpha)) \sqrt{\lambda L}. \quad \text{(22)}
\]

By approximation (21), the inventory manager incurs a holding cost of \( c(\alpha, \lambda) = h L(-z(\alpha)) / \sqrt{\lambda L} \), whereas the originator incurs a backlogging cost of \( \kappa(\alpha, \lambda) = b L(z(\alpha)) / \sqrt{\lambda L} \). Namely, the inventory manager’s mechanism design problem is a special case of the one advanced in §3, with \( c_1 = c_4 = \kappa_1 = 0 \), \( c_2 = h \sqrt{\lambda} \), \( \bar{c}_2 = \bar{h} \sqrt{\lambda} \), \( \kappa_2 = h \sqrt{\lambda} \), \( \beta(\alpha, \lambda) = L(-z(\alpha)) \), and finally, \( f(x) = x + L^{-1}(x) \).

Note that \( c \) and \( \kappa \) satisfy the assumptions made in §3 for fill rates of 50% or higher, which is what we focus on. In particular, \( \kappa(\alpha, \lambda) \) is convex decreasing in \( \alpha \) for a given \( \lambda \), because the average number of backlogs \( L(z(\alpha)) \) is convex decreasing in \( z(\alpha) \), which itself is increasing convex in \( \alpha \). Similarly, for a given \( \lambda \), the holding cost \( c(\alpha, \lambda) \) is convex increasing in \( \alpha \). Hence, the optimality of two-part tariffs by Theorem 1. Finally, note that given \( \alpha \), both \( c(\alpha, \lambda) / \lambda \) and \( \kappa(\alpha, \lambda) / \lambda \) are decreasing in \( \lambda \), exhibiting economies of scale.

In summary, the results of §4 yield that a menu of two-part tariffs is the optimal contract, thanks to the economies of scale inherent in certain order fulfillment operations and the ability of the service provider to contract on the realized demand.
7. Concluding Remarks

We examine how asymmetric demand information affects contracting between two parties. The demand volume is uncertain at the time of contracting, but can be quickly observed and verified by the service provider once the service commences. The service provider invests in training personnel and leasing space, which is cheaper if done before the initiation of service. The service provider offers a menu of contracts to screen different types of originators. The outside option of the originator is modeled as his profit were he to perform the operations in-house. We study two relevant outsourcing settings that can be studied in this framework, outsourcing call center and order fulfillment services. Both are increasingly important parts of the business process outsourcing market. Studying call centers demands a queuing framework, whereas order fulfillment operations are best analyzed via an inventory model. Economies of scale are common to multiserver queuing systems and constant leadtime inventory systems, meaning larger systems are more efficient for any given service level. In both environments, the expected revenue (sales) is proportional to the demand volume \( \lambda \), but the staffing cost for safety capacity or the holding cost for safety inventory are proportional to the square root of \( \lambda \).

The economies of scale and contractibility of demand volume simplify contracts significantly. In particular, although monopoly screening problems cannot generally implement the full-information solution, here we can implement the full-information solution by offering efficient service levels and a menu of two-part tariffs. Thus, a menu of two-part tariffs is optimal among all possible mechanisms.

The optimal menu of two-part tariffs provides the efficient standardized excess capacity \( \beta^* \) to both types at each arrival rate. Similarly, under the optimal menu of two-part tariffs, the originator obtains the full-information fill rate \( \alpha^* \) regardless of his type and the realized demand volume. In both call center and order fulfillment settings, if the originator anticipates high demand volume, he chooses the two-part tariff that charges him a high payment up front but no per-demand fee. On the contrary, if the originator expects a low arrival rate, he prefers the two-part tariff that involves no payment up front but charges him a high linear rate for the demand volume.

Full separation by a menu of two-part tariffs owes to contracting on the realized demand and the economies of scale inherent in certain queuing and inventory systems. Contractibility of demand increases the service provider’s contractual power by enabling contingent contracts. Because the outside option of the originator is defined as the profit the originator would earn if he were to conduct the operations in-house, the high-type originator would benefit from economies of scale for high demand volume when he performs the operations himself. Therefore, because of economies of scale, the high type is willing to pay more for the service in aggregate, but less on a per-volume basis, which renders the menu of two-part tariffs as in Theorem 1 incentive compatible.

Two-part tariffs are commonly used as screening devices in environments where consumers face uncertainty about their demand. An implication of Theorem 1 is that call center and inventory managers can restrict themselves to menus of two-part tariffs as possible forms of contracts. Also, the optimal menu of two-part tariffs in Theorem 1 does not require the service provider to have a correct estimate of \( q_i \), for the relative frequency of different originator types in order to implement the full-information solution.

Although we restrict our model to two classes, the analysis applies to the setting with an arbitrary number of types of originators whose demand volume distributions satisfy the following: \( \lambda_{i,1} \) is equal in distribution to \( a_i \lambda_i \), for \( a_i > 1 \) and \( 1 \leq i < N \). The optimal menu of two-part tariffs then involves a payment schedule \( p_i(\lambda) = \phi_i + v_i \lambda \), for type \( i = 1, \ldots, N \), where \( \phi_i \) is the fixed fee and \( v_i \) is the per-volume charge. The lowest type pays no fixed fee, whereas the highest type’s payment schedule involves no per-volume charge, i.e., \( \phi_i = v_i = 0 \). Moreover, as the expected demand increases, the fixed fee increases, whereas the per-volume payment decreases, i.e., \( \phi_i \leq \phi_j \), and \( v_i \geq v_j \), for \( i < j \). The fixed fee \( \phi_i \) and per-volume charge \( v_i \) are chosen to satisfy the individual rationality constraints with equality for each type. Incentive compatibility of such a simple menu of two-part tariffs similarly relies on economies of scale. Oftentimes, contracts offered in outsourcing markets involve multiple options for up-front payments and per-volume fees. As predicted, typically the subscription fees increase for larger outsourcers, whereas the per-volume fees decrease.

The cost advantage of early investment induces the service provider to sign the contract before demand is realized and distinguish the two types of the originator. As the cost difference between investing before and after service initiation decreases, so does the benefit of obtaining the demand information by distinguishing types. In the limiting case, there would exist a single contract that would be acceptable to both types and achieve the full-information solution. Moreover, offering the contract before demand realization is the service provider’s equilibrium behavior (as opposed to contracting after \( \lambda \) is realized). This is because if the contract is not signed before \( \lambda \) is learned, type \( i \) can still guarantee himself a payoff exceeding \( K_i \) in equilibrium, whereas the total surplus that can be extracted by the service provider decreases because of cost disadvantage of late investment.
We took the abandonment probability, or equivalently, the fraction of calls answered, as our measure of service quality for call center outsourcing, whereas we restricted attention to fill rate as the service-level criterion in the order fulfillment setting. This is largely done for purposes of transparency. Alternatively, one could have other service-level criteria for call centers, such as an upper bound on the probability of waiting more than a certain amount. The average number of backorders and average customer waiting times are also relevant measures of service for outsourced operations. For such criteria, one can use approximations similar to ours to show that our structural results continue to hold.

The margin $m$ earned per-demand volume is held constant throughout the analysis, and our results carry through if $m$ is a random variable that is independent of the demand volume and realizes after the service commences. If the high type also expects a higher margin, the slope of the low type’s payment scheme might not be steep enough to deter deviations from the high type. In that case, the low type’s per volume fee can be increased to a level that deters high type’s deviation by offering the low type a signing bonus (a negative fixed fee).

Obviously, the problem of service outsourcing is quite rich, and our model is incomplete. One important dimension not considered here is the possibility that the service provider can contract with multiple originators. If so, she could possibly build capacity that is jointly used by several originators, and therefore create additional economies of scale. Although we do not investigate such an environment, we anticipate that full separation of types would still rely on the high-type originator’s outside option being higher both on aggregate and per volume basis. The additional economies of scale that benefit only the service provider because she is able to offer higher service levels and fees, extract all surplus from the originators, whereas the originators earn no rents but still enjoy higher service levels.

Acknowledgments

Although this study was initially motivated by call center outsourcing, the authors thank an anonymous associate editor for suggesting the outsourcing of inventory operations as another application area. They thank an anonymous referee for pointing out the emergence of a single optimal contract when the cost advantage of early investment vanishes. They also thank Gérard Cachon, the associate editor, and the three anonymous referees for numerous helpful and constructive comments on earlier versions of the manuscript.

Appendix. Proofs

Proof of Proposition 1. Using (1), we can write

$$E[\phi(\tilde{\lambda}, \lambda_i)] = \int N(\tilde{\lambda}) + \int_{x_i}^{\infty} [N(\lambda) - N(\tilde{\lambda})] g_i(\lambda) d\lambda,$$

and the first-order condition for the minimization problem (2) dictates that the optimal demand $\lambda_i^*$ to plan for is given by the solution of the following:

$$N'(\lambda_i^*) [1 - \bar{f}(1 - G_i(\lambda_i^*))] = 0.$$  \hfill (23)

We can simplify (23) to obtain $\lambda_i^* = G_i^{-1}(1 - \bar{f})$ because $N'(x) > 0$ for all $x \geq 0$.

Because $G_i(\lambda) = G_i(\lambda/a)$ we have $g_i(\lambda) = g_i(\lambda/a)/a$. Then, $\lambda_i^* = a \lambda_i^*$ because $G_i'(\lambda_i^*) = \lambda_i^*/a = 1 - \bar{f}$. Notice that because $G_i(\lambda_i^*) = 1 - \bar{f}$, it follows from (1) and Proposition 1 that

$$\Psi_i = \bar{f} \int_{\lambda_i^*}^{\infty} N(\lambda) g_i(\lambda) d\lambda \quad \text{for } i = L, H.$$  \hfill (24)

From (24),

$$\Psi_i = \bar{f} \int_{\lambda_i^*}^{\infty} N(\lambda) g_i(\lambda) d\lambda = \bar{f} \int_{\lambda_i^*}^{\infty} N(ax) g_i(x) dx,$$

where the second equality follows from $\lambda_i^* = a \lambda_i^*$ and $g_i(\lambda) = g_i(\lambda/a)/a$. Because $a > 1$ and $N(-)$ is increasing, (24)–(25) imply that $\Psi_i(\lambda) \leq \Psi_i(L)$. Moreover, because $N(\cdot)$ is concave and $N(0) = 0$, $N(ax) \leq a N(x)$ for $x \geq 0$. Therefore, (24)–(25) imply that $\Psi_i(x) \leq \Psi_i(L)$.

Proof of Proposition 2. If IC constraints are ignored, it is optimal to increase the payments such that IR constraints bind, which implies that

$$E[p_i(\lambda_i)] = (m - \kappa_i)E[\lambda_i] - \kappa_i E[\sqrt{\lambda_i} f(\beta_1(\alpha_i, \lambda_i))] - K_i$$

for $i = L, H$.  \hfill (26)

Substituting $E[p_i(\lambda_i)]$ and $E[p_i(\lambda_i)]$ in the objective function, the problem becomes one of choosing the service levels $(\alpha_i, \lambda_i)$ for each $\lambda$ so as to maximize

$$\sum_{i = L, H} q_i \left[ (m - \kappa_i - c_i) E[\lambda_i] - E[\sqrt{\lambda_i} \beta(\alpha_i, \lambda_i)] - c_2 \beta(\alpha_i, \lambda_i)] - K_i - \Psi_i \right].$$

In other words, the problem reduces to choosing the efficient service level for each $\lambda$ to maximize the systemwide profits. The objective can be maximized for each type separately. Then, for each type $i$, the problem is to choose the service level $\alpha$ for each $\lambda$ so as to

$$\min \kappa_2 \sqrt{\lambda} f(\beta(\alpha, \lambda)) + c_2 \sqrt{\lambda} \beta(\alpha, \lambda).$$  \hfill (27)

The first-order condition, which is necessary and sufficient because $\beta(\cdot, \lambda)$ and $f(\beta(\cdot, \lambda))$ are convex, gives rise to the first-order condition (12), where $\beta^* = (f')^{-1}(-c_2/\kappa_2).$ Once the service levels $\alpha(\lambda)$ are set, the overall profit of the system is fixed and the payment scheme needs to satisfy only the IR constraints in order to extract all the expected surplus from the originator. That is, the payment scheme must satisfy (13).

References


