Call Center Outsourcing: Coordinating Staffing Level and Service Quality

Z. Justin Ren* Yong-Pin Zhou†

July 7, 2006

Abstract

In this paper, we study the contracting issues in an outsourcing supply chain consisting of a user company and a call center that does outsourcing work for the user company. We model the call center as a $G/G/s$ queue with customer abandonment. Each call has a revenue potential, and we model the call center's service quality by the percentage of calls resolved (revenue realized). The call center makes two strategic decisions: how many agents to have and how much effort to exert to achieve service quality.

We are interested in the contracts the user company can use to induce the call center to both staff and exert effort at levels that are optimal for the outsourcing supply chain (i.e., chain coordination). Two commonly used contracts are analyzed first: piece-meal and pay-per-call-resolved contracts. We show that although they can coordinate the staffing level, the resulting service quality is below system optimum. Then, depending on the observability and contractibility of the call center's effort, we propose two contracts that can coordinate both staffing and effort. These contracts suggest that managers pay close attention to service quality and its contractibility in seeking call center outsourcing.

*Operations and Technology Management Department, Boston University School of Management.

†Department of Management Science, University of Washington Business School

1 Introduction

An increasing number of companies are moving their call center operations offshore. According to market researcher Datamonitor, the total value for the U.S. outsourcing market will be worth almost $24 billion by 2008, compared with the current $19 billion. According to Datamonitor, "By 2008, 1 in 15 agent positions (workstations) will be outsourced to a foreign market, from 1 in 24 currently. By year-end 2003, offshore outsourcers, climbing to 201,000 by 2008, will staff 121,000 agent positions."[1]

Despite lower labor cost, companies in practice have experienced mixed results from outsourcing their call centers. In fact, some companies' outsourcing strategies have backfired, causing them to re-evaluate or abort their outsourcing mission. In November 2003, DELL was forced to move its call center operations for OptiPlex desktops and Latitude laptops from India back to the U.S., after customers complained about language difficulties and delays in reaching senior technicians (Financial Times, November 26, 2003). During the same time period, Lehman Brothers, a leading financial services company, had to shift some call center operations from India back to the U.S. after its customers complained about the quality of service (Financial Times, December 17, 2003).

One important reason not all companies benefit from outsourcing is the lack of understanding of the economics of outsourcing, and how to coordinate the outsourcer to better serve the company that initiates the outsourcing (we call this the *user company*, or *user*). Indeed, it has been noted that, "Expectations in cost reduction are not always met because outsourcing contracts can be developed with a poor understanding of current costs ..." (United States Government Accountability Office, 2004)

This lack of understanding of the call center outsourcing contracts may be attributed to the fact that there has been little academic research on the call center outsourcing supply chain and its coordination. The call center outsourcing supply chain differs from the physical goods or inventory supply chain in that when a unit of physical goods is sold to the customer, the retailer, who

[1] `http://callcentermagazine.com/shared/article/showArticle.jhtml?articleId=17200246`

owns the inventory, reaps a revenue; however, in a call center outsourcing supply chain, when a service is provided, the call center usually does not gain revenue directly from the service. Instead, the revenue goes to the user company and the call center is compensated by the user company. Furthermore, call centers operate as queueing systems. Customers call in and are placed in a queue if no servers are immediately available. Waiting cost is incurred. Customers may also drop out of the queue, i.e., abandon, due to impatience. Such costs are unique to a queueing system and are normally absent in the research on inventory supply chains. Finally, call centers provide service through the phone line and are invisible to the end customers. As a result, customers do not distinguish the call center from the user company. Any costs incurred during the service provided by the call center (waiting cost, abandonment cost, loss of goodwill from unsatisfactory service) will be imposed upon the user company, rather than the call center itself.

Service quality is especially important in call center outsourcing. Poor service quality by a call center is directly reflected upon its user company. Improving the service quality of the call center agents is vital to the user company's profitability. In revenue-generating call centers (e.g., phone-order services), a well-trained and motivated sales agent can answer customers' inquiries to their satisfaction and successfully increase the likelihood of a sale. In non-revenue-generating call centers (e.g., technical support centers), a knowledgeable agent can solve technical problems in a timely fashion, while an incompetent agent can aggravate a customer's frustration, lead to customer complaints, reduce the likelihood of future sales, and hurt the user company's image (as illustrated by the DELL and Lehman Brothers examples above).

This paper aims to address these outsourcing challenges by answering the following questions:

1. *What is different about call center outsourcing, in terms of coordinating the whole chain?*

It is well known that in an inventory supply chain, a linear wholesale contract causes 'double marginalization', where the retailer stocks less than the supply chain optimal quantity. In the outsourcing supply chain, a different form of double marginalization exists because the call center, unlike the retailer, is paid by the user company rather than the customers. Where double marginalization occurs is in the call center's effort to achieve service quality. When the call center's

profit margin does not match that of the integrated outsourcing supply chain, the call center will rationally exert less effort, resulting in a service quality inferior to that in an integrated system.

2. *How effective are the call center outsourcing contracts commonly observed in practice?*

The piecemeal and pay-per-call-resolved contracts are commonly observed in call center outsourcing. Are they capable of inducing system-optimal staffing level and effort (to improve service quality) from the call center? What are their implications for social welfare? Our analysis indicates that these contracts induce very different effort levels, as well as the corresponding social welfare.

3. *How to achieve system coordination, i.e., system-optimal staffing and effort levels, with a contracting mechanism?*

Different forms of call center outsourcing have been observed in practice. Some companies take a 'hands-off' approach, while other companies form partnerships with their outsourcers, sharing set-up and operating costs. What form of outsourcing should a company choose? In this paper we show that although some of the contracts in practice fail to coordinate the outsourcing supply chain along the service quality dimension, there are two types of contracts that can. These two contracts differ in whether the call center's effort is observable and contractible. Our results shed light on how to choose the right form of outsourcing. In particular, we find that when service quality is important but hard to monitor, a close collaboration is needed between the user company and its outsourcer to achieve system coordination.

This paper makes two contributions. First, we are among the first to study supply chain coordination in the context of call center outsourcing. Much of the previous research has studied call center operations either from a queueing theoretical perspective with no explicit cost considerations, or as a stand-alone cost-minimizing or profit-maximizing company. Our paper builds on the existing research, but studies the outsourcing supply chain as a whole. Specifically, we study how to coordinate the different players in the chain to achieve system optimality.

Second, we study the issues related to both staffing and service quality in call center operations. We measure service quality by the percentage of calls that are served to customers' satisfaction (or, 'call resolution probability'). Outsourcing often brings immediate cost savings, but companies

should keep in mind that there can also be some 'hidden costs' in outsourcing (BusinessWeek, 2003), one of which is service quality cost. Because the call center is largely invisible to the end customer, when it under-performs, it is the user company that suffers the consequences. For this reason, the service quality of call centers must be taken into account and carefully managed by the user company. However, since each service encounter is unique (and sometimes unobservable), it may not be possible to contract directly on service quality. In this paper we provide contracts that can induce the call center to exert effort to achieve supply-chain-optimal service quality.

The results of this paper have important managerial implications. We find that commonly used contracts, such as the piecemeal and pay-per-call-resolved contracts, can coordinate the staffing level, but not the effort level. Moreover, we find that when the service quality effort can be observed, a pay-per-call-resolved plus cost-sharing contract can coordinate the outsourcing supply chain on both staffing and service quality. When the service quality effort is not observable, a 'partnership'-type contract can coordinate the outsourcing supply chain. Our results thus provide insights that can help companies to decide what form of relationship to pursue with their outsourcers.

The rest of the paper is organized as follows. Section 2 surveys the literature. After presenting the model setup in Section 3, we study a centrally managed outsourcing supply chain as a benchmark and define our research question in Section 4. Sections 5 and 6 study contracts in the decentralized setting. Section 5 shows that two contracts commonly used in practice fail to coordinate the outsourcing supply chain. Then in Section 6, we propose two contracts that can coordinate the system, where effort can be observable or non-observable, respectively. In Section 7 we use numerical analysis to generate more insights on the expected profit function and the effectiveness of contracts. Finally, we discuss the limitations of this research and conclude in Section 8.

2 Literature Review

Call center has become an increasingly productive research area in recent years. For a review of the state-of-the-art call center research, see Gans et al. (2003).

Most research in this area focuses on the queueing dynamics of call centers. The queueing model contained in this paper is a general $G/GI/s$ queue with customer abandonment, where impatient customers may leave after joining the queue but before being served. Empirical evidence suggests that in the context of call centers, customer impatience plays an important role in the behavior of queues (Zohar et al. 2002, Brown et al. 2005). Garnett et al. (2002) study the simplest multi-server model with abandonment: exponential arrival and service rate, unlimited waiting space, and exponential abandonment rate, denoted as the $M/M/s/\infty+M$ model. They point out that abandonment is very important in understanding the dynamics of real-world call centers. Other recent papers that incorporate customer abandonment in queueing models include Brandt and Brandt (1999), Whitt (1999), Mandelbaum and Shimkin (2000), Akşin and Harker (2001), Shimkin and Mandelbaum (2004), Zeltyn and Mandelbaum (2004).

Because multi-server queues with abandonment are difficult to analyze exactly, researchers have made significant progress on various approximation schemes. Garnett et al. (2002) provide a diffusion approximation for the $M/M/s/\infty+M$ model, and Whitt (2004a) provides fluid and diffusion approximations for the $M/M/s/r+M$ model in the overloaded regime. Whitt (2004b) shows that "fluid approximation yields a remarkably simple approximation for the performance of the $G/GI/s+GI$ queue, but one which is quite insightful." To capture the first-order performance description for multi-server queues with abandonment, Whitt (2006) develops deterministic fluid approximations for the general $G/GI/s+GI$ models. Based on these results, we adopt the fluid approximation in this paper as well. More details are given in Section 4.

Another important customer queueing behavior is retrial. Falin and Templeton (1997) is a good reference on this subject. Hoffman and Harris (1986) estimate the retrial rates by both blocked and abandoned callers at IRS's call center, and Aguir et al. (2004) incorporate these retrials into a call center queueing model. Because our fluid approximation applies to a stationary queue, the retrial behaviors do not affect our analysis, so we do not model them explicitly.

The objective of many recent call center models is to minimize total cost including staffing, waiting, abandonment, and telecommunication costs (or a subset thereof). Examples include Bas-

samboo et al. (2005), Gurvich et al. (2004), and Harrison and Zeevi (2005). Models that also consider revenue include Andrews and Parsons (1993), Helber (2004), and Koole and Pot (2004). In terms of the overall profit function, our model is closest to Whitt (2004b), but we differ in several aspects. First, our model includes a service quality component. Not all served customers generate revenue–only those served *and resolved* customers do. The resolution rate depends on the service quality. For those calls served but not resolved, there is an additional penalty. Second, we introduce an additional decision variable, the call center's effort, which influences its service quality. We allow effort to be unobservable and unverifiable. Third, while Whitt (2004b) studies the optimal staffing level for a stand-alone call center, we study the optimal staffing and effort levels from the perspective of the whole outsourcing supply chain.

While our paper assumes complete outsourcing, both Gans and Zhou (2005) and Akşin et al. (2004) allow the user company to outsource some, but not all, calls. Gans and Zhou (2005) focus on the queueing control and capacity planning aspects, while Akşin et al. (2004) suppress queueing details to focus on the higher-level contract design. Akşin et al. (2004) 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'. Chevalier et al. (1998) also study the service subcontracting issue, but their focus is on the 'make or buy' decision. None of these models consider service quality.

There exists a large body of literature on service quality. The SERVQUAL model (Parasuraman et al. 1990) lists ten dimensions of service quality, which is then narrowed down to five. Call centers have long used wait-related measures, such as average call waiting time and call waiting probability, as measures of service quality. But these measures have little to do with the actual service encounter and the customer's satisfaction from that encounter. Recently Gans (2002) uses the customer loyalty and defection probability to model the service quality, and de Véricourt and Zhou (2005) use call resolution probability to model quality in customer-service oriented call centers.

In this paper, we define service quality as the probability of a customer's inquiry call being successfully resolved. When the calls generate revenue, call resolution means the conversion of a

customer inquiry into sales. When the calls are for customer service, call resolution means that the customer is satisfied and will not call back for the same question, thus reducing the customer's future service cost and increasing the customer's future consumption. Shumsky and Pinker (2003) recognize the importance of call resolution in providing the agents with the appropriate economic incentive. They show 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. This is similar to the "pay-per-call-resolved" contract studied in our paper (Section 5.2).

The coordination of inventory supply chain is another related research area, where various types of contracts have been identified that can achieve system coordination. These include buy-back (Pasternack 1985), quantity-flexibility (Tsay 1999), sales-rebate (Taylor 2002), and revenue-sharing (Cachon and Lariviere 2005) contracts. Cachon (2003) has an extensive discussion about the simple linear wholesale contract, and Cohen et al. (2003) provide an industrial example of its effectiveness. In this paper, we study contracts that are specific to service outsourcing, such as piece-meal or pay-per-call-resolved, and thus are different from those used in inventory supply chains.

3 Model Setup

We consider an outsourcing supply chain consisting of two companies: a user company and a call center (outsourcer). The call center is typically large, and is modeled as a multi-server queueing system with customer abandonment. With arrival rate λ, the call center staffs s servers (i.e., agents) each with service rate μ. Customers enter a queue if not served upon arrival. They are impatient, and will abandon after a random amount of time, which has continuously differentiable PDF f and CDF F. The waiting cost rate is c_w, and each time a customer abandons, there is a cost of c_a.

Of the calls that are eventually served, a portion, p, are satisfactorily resolved. Each time a call is resolved, a revenue r is earned. For the rest $1-p$ portion of the calls (served but not resolved), there is a loss of goodwill c_g for each of them. In a revenue-generating call center (e.g., catalog-shopping order service), a well-trained sales agent can satisfactorily answer customers' questions

and increase the likelihood of a sale or introduce new products to customers resulting in increased revenue. In a non-revenue-generating call center (e.g., technical support), a knowledgeable service agent can solve technical problems successfully in a timely fashion. In both cases, an incompetent service agent can lead to customer complaints, reducing the likelihood of both immediate and future sales. Repeated customer phone calls for the same problem can also increase system load and cost.

We assume that the call resolution probability p is a non-negative continuous random variable with support $[0,1]$ and CDF $G(p)$. Moreover, we assume that p is influenced by the call center's effort e, which may be unobservable or unverifiable by the user. Such effort may include hiring human-resource consultants to improve the recruiting process, providing productivity-enhancing facilities (e.g., better work environment) and amenities, or purchasing equipment and training to improve servers' service quality. The expected call resolution probability for a given effort level e, denoted as $\overline{p}(e)$, is then

$$\overline{p}(e) = \int_0^1 [1 - G(p|e)]dp. \tag{1}$$

It is natural to assume that effort positively impacts the call resolution probability, and that the marginal impact is decreasing: $\frac{\partial G(p|e)}{\partial e} < 0$ and $\frac{\partial^2 G(p|e)}{\partial e^2} > 0$. It follows immediately that $\overline{p}'(e) > 0$ and $\overline{p}''(e) < 0$. For convenience, we also assume $\overline{p}'(e)|_{e=0} = \overline{p}'_0 > 0$, and $\lim_{e\to\infty} \overline{p}'(e) = 0$.

Effort is costly to the call center, at a rate of c_e. Hiring staff also costs the call center at a rate of c_s. To rule out uninteresting cases where the call center finds itself having a negative profit, we restrict $r\overline{p}(0) > c_s + (1 - \overline{p}(0))c_g$.[2]

Our notation is summarized below.

[2]In reality, it is quite plausible that when an outsourcer does not spend enough effort, the percentage of resolved calls are so low that it is no longer profitable to be in service. It is straightforward to incorporate this into our model. However in order to highlight the insights from our model, we rule this out for the sake of a cleaner presentation.

λ, μ	arrival and service rates
$p(e)$	random percentage of calls resolved as a function of effort
s	number of servers
$T(s)$	number of customers served per time unit in steady state
$L(s)$	abandonment in steady state, $\lambda = L(s) + T(s)$
$W(s)$	waiting time (before either service or abandonment occurs) in steady state
r	expected revenue from each served and resolved call
c_g	loss of goodwill from calls served but not resolved
c_s, c_e, c_a, c_w	unit staffing, effort, abandonment, and waiting costs

4 Benchmark: An Integrated Outsourcing Supply Chain

As a benchmark, we first look at the integrated outsourcing supply chain where the call center and the user are owned by the same company who makes the centralized decision on staffing and effort levels. By staffing s agents and exerting effort e, the system's total expected profit is:

$$\pi^I(s,e) = \underbrace{r\overline{p}(e)T(s)}_{\text{revenue}} \quad \underbrace{-c_s\mu s - c_e e}_{\text{staffing and effort costs}} \quad \underbrace{-c_a L(s) - c_w \lambda W(s) - c_g\left(1-\overline{p}(e)\right)T(s)}_{\text{abandonment, waiting, and loss of goodwill costs}}. \tag{2}$$

Note that our linear cost and revenue structure is similar to that in Whitt (2004b). Also note that we assume effort cost is independent of the staffing level. The integrated system solves the following profit maximization problem:

$$\max_{s,e} \pi^I(s,e). \tag{3}$$

Real-world call center operations often stipulate a certain performance requirement, such as 'ASA (average speed of answer, or, average waiting time in queue) $\leq$ 30 seconds'. The optimization program (3) can be augmented by some service level constraints, but we choose not to explicitly model these constraints because conceptually they can be 'dualized' into costs. For example, a constraint on ASA can be dualized into a waiting time cost (and add to the existing c_w). Moreover, as we will show later, with the fluid approximation, it is optimal for the call center to staff sufficiently so that no customer waits or abandons. In this case, the service level constraints would not apply. Of course, this holds only for large systems where the fluid approximation is appropriate.

The $G/GI/s{+}GI$ system in the optimization problem (3) is hard to analyze exactly. Garnett et al. (2002) show that even the analysis of a much simpler system, $M/M/s{+}M$, is hard. To simplify the analysis, they propose a diffusion approximation method. Whitt (2006) provides a fluid approximation for the general $G/GI/s{+}GI$ models, and shows that the approximation is remarkably accurate. Moreover, a fluid approximation allows for analytical tractability, from which one can gain important managerial insights. Therefore, our approach follows Whitt's fluid approximation. In the fluid approximation, throughput rate is the minimum of the arrival rate and the maximum service rate: $T(s) = \min(\lambda, \mu s)$. Any arrival in excess of the maximum service rate is abandoned:

$$L(s) = (\lambda - \mu s)^+ = \lambda - T(s), \tag{4}$$

where $(x)^+ = \max\{0, x\}$.

Furthermore, by using a Taylor series approximation of the CDF F around $t = 0$, and with a mild assumption that $f(0) \neq 0$,[3] one can obtain a relatively simple relationship on steady state waiting time (for details see Whitt 2004b):

$$W(s) = \frac{L(s)}{f(0)\lambda}. \tag{5}$$

This suggests that in the fluid approximation, waiting time is proportional to abandonment. Of course, as Whitt (2004b) points out, this relationship does not apply to the original stochastic model, but it captures the first-order effects of the queueing system. Substituting (1), (4), and (5) in (2), we obtain

$$\begin{aligned}
\pi^I(s,e) &= r\overline{p}(e)(\lambda - L(s)) - c_s\mu s - c_e e - c_a L(s) - \frac{c_w}{f(0)}L(s) - c_g(1-\overline{p}(e))(\lambda - L(s)) \\
&= [r\overline{p}(e) - c_g(1-\overline{p}(e))]\lambda - \left[r\overline{p}(e) - c_g(1-\overline{p}(e)) + c_a + \frac{c_w}{f(0)}\right] L(s) - c_s\mu s - c_e e. \qquad (6)
\end{aligned}$$

[3] This assumption is certainly satisfied by the commonly used exponential patience distribution. Equation (5) is essentially based on the 1st-order Taylor expansion when $f(0) \neq 0$. When $f(0) = 0$, let n be the smallest integer such that $F^{(n)}(0) \neq 0$. The n^{th}-order Taylor approximation will then yield $W(s) = \sqrt[n]{\frac{n!L(s)}{f^{(n-1)}(0)\lambda}}$. The fact that $W(s)$ is concave in this case will complicate analysis.

Recall that we assume $r\overline{p}(0) > c_s + (1 - \overline{p}(0))c_g$ to guarantee that the integrated system can obtain a positive profit by operating its call center. Since $L(s)$ is piece-wise linear and convex in s, the expected profit is also nicely behaved with respect to e. By solving the first-order conditions of (6), we obtain the following proposition. All the proofs of this paper are in the on-line appendix.

Proposition 1 *The profit-maximizing staffing level s^I and effort e^I for the integrated system satisfy*

$$s^I = \lambda/\mu, \tag{7}$$

$$\overline{p}'\left(e^I\right) = \frac{c_e}{(r + c_g)\,\lambda}, \ \textit{if} \ \frac{c_e}{(r + c_g)\,\lambda} < \overline{p}_0'; \qquad e^I = 0 \ \textit{o.w.} \tag{8}$$

The solution in (7) indicates that the integrated system would staff just enough to meet demand. Therefore in steady state, there is no customer abandonment or waiting. This result is not new. It is the same result obtained in Whitt (2004*b*) for the linear revenue and cost case. On the other hand, the solution in (8) indicates that the system optimally balances the cost and benefit of effort. For example, when the revenue rate (r), goodwill cost (c_g), and/or customer arrival rate (λ) increase, the optimal effort also increases because the expected benefit from more resolved calls becomes larger. When the cost of effort c_e increases, however, the call center will optimally exert less effort. When the maximum marginal increase in call resolution probability, $\overline{p}_0'$, is so small that $\overline{p}_0' < \frac{c_e}{(r+c_g)\lambda}$, no effort would be expended.

At optimality, the total expected outsourcing supply chain profit is

$$\pi^I(s^I, e^I) = \left[r\overline{p}(e^I) - c_g(1 - \overline{p}(e^I)) - c_s\right]\lambda - c_e e^I. \tag{9}$$

With this integrated model as a benchmark, we now investigate the case of a decentralized outsourcing supply chain.

5 Decentralized System: Outsourcing

In the decentralized outsourcing supply chain, the user and the call center are two independent entities. The user company offers the call center a contract to take all of its calls. If the call center

accepts the contract, it chooses its staffing and effort levels to serve incoming calls from the user's customers. Depending on the service outcomes, the call center is paid by the user according to the contract. The user, on the other hand, receives revenue from its own customers only if they are served and resolved by the call center. Because the call center is invisible to the customers, it is the user that bears the negative consequences of customer abandonment, customer dissatisfaction from waiting, and calls not being resolved. The amount the user pays to the call center is denoted by Ψ, which is specified in the contract and could be a function of a number of variables, such as T, W, or L. The task of contract design is to determine what factors determine Ψ, and how.

In a decentralized setting, the total expected outsourcing supply chain profit in (2) is decomposed into two parts: π^c for the call center and π^u for the user:

$$\begin{aligned} \pi^c(s,e) &= \mathsf{E}(\Psi) - c_s\mu s - c_e e, \\ \pi^u(s,e) &= r\overline{p}(e)T(s) - c_a L(s) - c_w\lambda W(s) - c_g\left(1-\overline{p}(e)\right)T(s) - \mathsf{E}(\Psi). \end{aligned}$$

Our goal is to identify contracts that can achieve two objectives in the decentralized setting: (1) achieve the system-optimal staffing and effort level, and (2) achieve an arbitrary split of system profit between the user and the call center. The reason is that by coordination, the outsourcing supply chain can achieve the maximum possible profit. Then with a proper contract, the total profit can be split between the user and the call center such that both parties are better off than when the outsourcing supply chain is not coordinated. In other words, with a coordinating contract both parties can first make a bigger pie, and then share it in such a way that each gets a bigger piece than before.

The key here is to find out the call center's optimal staffing and effort decisions in the decentralized setting. It obviously depends on the specific Ψ. For example, the simplest form of Ψ is a fixed payment, i.e., $\Psi = \sigma$ where σ is a constant. It is easy to see that a fixed payment will not induce the call center to staff or exert effort at the system-optimal level. In the next sections we investigate four specific contract forms: a piecemeal contract, a pay-per-call-resolved contract, a pay-per-call-resolved plus cost sharing contract, and a partnership contract.

5.1 Piecemeal (PM) Contract

Piecemeal contract, also called wholesale contract or linear contract, is commonly observed in the industry. The user pays the call center a unit rate b for each call served:

$$\Psi = bT(s). \tag{10}$$

In this contract, the user only needs to decide b and the payment depends only on $T(s)$. In order for the contract to be acceptable to both parties, we must have $c_s < b < r$. Under this contract, the profit for both outsourcing supply chain parties are:

$$\begin{aligned} \pi^c_{PM}(s,e) &= bT(s) - c_s\mu s - c_e e \\ \pi^u_{PM}(s,e) &= r\overline{p}(e)T(s) - c_a L(s) - c_w\lambda W(s) - c_g\left(1-\overline{p}(e)\right)T(s) - bT(s). \end{aligned} \tag{11}$$

Proposition 2 *Under the piecemeal contract, the call center will achieve the system-optimal staffing level, i.e., $s_{PM} = s^I$. However, it will exert no effort, i.e., $e_{PM} = 0$.*

Because the piecemeal contract links the call center's income to the number of calls served, the call center has an incentive to staff adequately. Unfortunately, the piecemeal contract does not link the call center's income to the resolution of calls. As a rational response, the call center does not exert any effort. Consequently, the outsourcing supply chain reaches optimality only along one of the two important dimensions; it is not coordinated.

Proposition 2 indicates that if we examine the call center on a grand scale, where the deterministic aspects of a fluid approximation dominates the stochastic variations at the detailed queue level, then a piecemeal contract can work well in terms of coordinating the staffing level. Besides, it has a simple form and is easy to implement. These help to explain the popularity of this contract in practice. However, such a contract does not consider service quality. Therefore when service quality is uncertain and depends on the call center's private actions (i.e., effort), the user needs more than just a piecemeal contract to achieve the service quality it desires.

5.2 Pay-per-call-resolved (PPCR) Contract

In order to account for the call center's effort, a natural extension of the piecemeal contract is to link the payment to the call center to its call resolution:

$$\Psi = bp(e)T(s).$$

Unlike in the PM contract, here b is the unit rate the user pays the call center for each call served and resolved. The payment to the call center now depends on $p(e)T(s)$, which is determined by both s and e. Let e_{PPCR} be the call center's effort level under a PPCR contract. b needs to be large enough to ensure that the call center's profit is non-negative. In this case, $b > \frac{c_e e_{PPCR}/\lambda + c_s}{\overline{p}(e_{PPCR})}$.

As an example, when Britain's Virgin Train outsourced its sales call center function to Cap Gemini, its contract compensated the call center based not on the number of inquiry calls, but on the ticket sales. The call center not only needed to meet call volumes, but also had a strong incentive to convert customer inquiries into actual sales. The result was a 400% increase in revenue (K'djah Worldwide, 2003).

Under this contract, the profits for both outsourcing supply chain parties are:

$$\begin{aligned} \pi^c_{PPCR}(s,e) &= b\bar{p}(e)T(s) - c_s\mu s - c_e e, \\ \pi^u_{PPCR}(s,e) &= r\overline{p}(e)T(s) - c_a L(s) - c_w\lambda W(s) - c_g\left(1-\overline{p}(e)\right)T(s) - b\bar{p}(e)T(s). \end{aligned} \tag{12}$$

We find that the pay-per-call-resolved contract can not only coordinate staffing level, but also motivate the call center to exert effort to improve service quality.

Proposition 3 *Under the pay-per-call-resolved (PPCR) contract, the call center will set staffing level at the system-optimal level: $s_{PPCR} = s^I$. It will also exert effort to improve service quality:*

$$\overline{p}'\left(e_{PPCR}\right) = \frac{c_e}{b\lambda},\ \textit{if}\ \frac{c_e}{b\lambda} < \overline{p}'_0; \qquad e_{PPCR} = 0\ \textit{o.w.}$$

Similar to the PM contract, the PPCR contract induces the call center to staff enough people so that no calls would be lost, because its revenue is directly tied to the volume of calls served. Furthermore, because the call center is compensated only when a call is served and resolved, it has

an incentive to exert more effort to increase the service quality and the volume of calls resolved. However, because $\frac{c_e}{b\lambda} > \frac{c_e}{(r+c_g)\lambda}$ and $\overline{p}'(e)$ is decreasing, under the PPCR contract the call center's effort is less than the outsourcing supply chain optimum:

Corollary 1 $e_{PPCR} \leq e^I$.

Underlying this result is the difference in the profit margin of the integrated system and that of the call center. In the integrated system, each additional unit of effort (at the marginal cost of c_e) results in an increase of system revenue from the additional resolved calls and a reduction in the loss-of-goodwill cost, $(r + c_g)\,\lambda$. For the call center under the PPCR contract, however, each additional unit of effort would only increase the call center's revenue by $b\lambda < (r + c_g)\,\lambda$. Due to this difference in the profit margin, the call center will under-invest in the effort. This under-investment is the result of a 'double marginalization' problem analogous to that in an inventory supply chain.

Is there a contract that can fix the double marginalization problem in the outsourcing supply chain? The answer is yes, but the form of the 'solution' contract depends on whether effort is observable and contractible. We study both cases next.

6 Chain Coordination in Call Center Outsourcing

6.1 Observable and Contractible Effort

Some efforts by the call center, such as leasing new buildings and facilities, and purchasing software for technical support or training purposes, can be measured and verified. Moreover, in practice many companies already use a 'cost-plus' type of contract, which calls for the sharing of cost information. In these cases, the user can propose to share a proportion of the call center's costs in order to induce the call center to exert enough effort to coordinate the outsourcing supply chain. Specifically, consider the following contract. Let $\alpha = b/r$ be the ratio between what the call center gets from each resolved call and what the user gets. The user modifies the PPCR contract by sharing $(1 - \alpha)$ proportion of the call center's staffing and effort costs. Moreover, the call center

pays a penalty of αc_g for each call served but not resolved, as well as α proportion of the waiting and abandonment costs. In sum, the user and the call center share each other's costs. Therefore, the contractual payment is

$$\Psi = bp(e)T(s) + (1-\alpha)\left(c_s\mu s + c_e e\right) - \alpha\left[c_g\left(1-p(e)\right)T(s) + c_a L(s) + \frac{c_w}{f(0)}L(s)\right]. \tag{13}$$

We call this contract a "pay-per-call-resolved plus cost sharing" contract (PPCR+CS). As in the PPCR contract, b is the unit rate for each call served and resolved, and it (or, equivalently, α) is the only decision variable the user has to consider in designing the contract. Under this contract, the outsourcing supply chain parties' expected profits are:

$$\pi^c_{PPCR+CS}(s,e) = \alpha\left\{\left[r\overline{p}(e) - c_g\left(1-\overline{p}(e)\right)\right]T(s) - c_s\mu s - c_e e - c_a L(s) - \frac{c_w}{f(0)}L(s)\right\}, \tag{14}$$

$$\pi^u_{PPCR+CS}(s,e) = (1-\alpha)\left\{\left[r\overline{p}(e) - c_g\left(1-\overline{p}(e)\right)\right]T(s) - c_s\mu s - c_e e - c_a L(s) - \frac{c_w}{f(0)}L(s)\right\}. \tag{15}$$

Clearly, under the PPCR+CS contract, the call center's share of the total system profit is α, and the user's share is $1-\alpha$. Therefore the call center's incentive is completely aligned with that of the outsourcing supply chain. So it will take the system-optimal actions in staffing and effort, and the supply chain will be coordinated.

Proposition 4 *Under the pay-per-call-resolved with cost sharing (PPCR+CS) contract, the call center will both staff and exert effort at the system-optimal level. That is,* $s_{PPCR+CS} = s^I$ *and* $e_{PPCR+CS} = e^I$.

Equations (14) and (15) suggest that the two parties each take a fixed portion of the total supply chain profit, so by carefully selecting b (hence α), the user can arbitrarily split the system profit with the call center.

Corollary 2 *Under the pay-per-call-resolved with cost sharing (PPCR+CS) contract, an arbitrary split of the expected outsourcing supply chain profits can be achieved by varying the pay-per-call-resolved rate* b. *The call center's share of the total profit is* $\alpha = b/r$.

Proposition 4 and Corollary 2 together demonstrate the importance of outsourcing supply chain coordination: by coordination, the outsourcing supply chain can achieve the maximum profit. Next, the two parties can always choose b (hence α) to split the profit so that each party's profit is higher than it would have been without the chain coordination.

The PPCR+CS contract seems quite cumbersome because the two parties need to share information about many cost items. Note, however, that the abandonment and waiting costs in (13), $c_a L(s)$ and $\frac{c_w}{f(0)} L(s)$, serve the same purpose: to penalize the call center for under-staffing. As long as one of them is in the contract, the contract will continue to coordinate the system. Therefore, by removing either one of them, we can simplify the PPCR+CS contract.

Other types of cost-sharing contracts can also coordinate the supply chain. For example, the user can share the effort cost, $(1-\alpha)\, c_e e$, but not the staffing cost. If α is set to be $b/\left(r + c_g\right)$, then the call center would still be induced to staff and exert effort at the system-optimal level. However, this contract cannot achieve an arbitrary split of profit because the expected profit of the call center is no longer exactly α proportion of the total system profit.

6.2 Non-observable Efforts

Oftentimes the call center's effort cannot be observed, or verified. For example, it is often difficult to measure a call center manager's effort in supervising her staff and solving day-to-day problems. If a contract relies on the call center to truthfully report its managers' supervision and training effort, then the call center has an incentive to over-state its effort.

In the face of this challenge, we propose the following contract, which consists of two parts.

- First, the call center has to pay a fee to serve each call (excluding the abandoned calls), but gets to keep all the revenue from served and resolved calls. This unit 'usage fee' is $(1-\alpha)\left[r\overline{p}(e^I) - c_g\left(1-\overline{p}(e^I)\right)\right]$, where $\alpha \in [0,1]$ is the only contract parameter the user needs to decide. For each call served but not resolved, however, the call center must pay a penalty of c_g to the user.

- Second, the user shares the call center's staffing and effort costs by paying the call center $(1-\alpha)\, c_s \mu s + (1-\alpha)\, c_e e^I$. No cost will be shared when the call center staffs no servers (i.e., $s = 0$). Moreover, the call center pays α proportion of the waiting and abandonment costs.

It is important to note that because the call center's effort can not be observed and verified, all the payments in the contract cannot be based on the call center's real effort level. Instead, they are based on the call center's real staffing level (which can be observed) and the chain-optimal effort level e^I (which can be calculated). Part of the user's payment to the call center, $(1-\alpha)\, c_e e^I$ is fixed. When α is too small, the fixed payment may be so high that it induces the call center to keep a minimal staffing level and exert no effort. In this case, the sole purpose of the call center's operation is to collect the fixed payment. This solution is clearly impractical. To avoid this trivial situation, we assume in the following discussion that the user will choose a sufficiently large α.[4]

The contract has the flavor of that of a franchise, in the sense that it lets the call center earn revenue directly from the customer, and pay a user fee ('franchise fee') based on call volume. But it is more than a common franchise contract, because the user also shares part of the call center's costs. Therefore it is more like a partnership. Under this partnership contract (which we denote as PART), the transfer payment is

$$\begin{aligned} \Psi &= rp(e)T(s) - (1-p(e))c_g T(s) - (1-\alpha)\left[rp(e^I) - c_g\left(1-p(e^I)\right)\right]T(s) \\ &\quad + (1-\alpha)\, c_s \mu s + (1-\alpha)\, c_e e^I - \alpha\left(c_a + \frac{c_w}{f(0)}\right) L(s), \end{aligned} \tag{16}$$

and the expected profits for both parties in the outsourcing supply chain are

$$\begin{aligned} \pi^c_{PART}(s,e) &= \left[r\overline{p}(e) - c_g\left(1-\overline{p}(e)\right)\right]T(s) - (1-\alpha)\left[r\overline{p}(e^I) - c_g\left(1-\overline{p}(e^I)\right)\right]T(s) \\ &\quad -\alpha c_s \mu s + (1-\alpha)\, c_e e^I - c_e e - \alpha\left(c_a + \frac{c_w}{f(0)}\right) L(s), \end{aligned} \tag{17}$$

$$\begin{aligned} \pi^u_{PART}(s,e) &= (1-\alpha)\left[r\overline{p}(e^I) - c_g\left(1-\overline{p}(e^I)\right)\right]T(s) - (1-\alpha)\left(c_a + \frac{c_w}{f(0)}\right) L(s) \\ &\quad - (1-\alpha)\, c_s \mu s - (1-\alpha)\, c_e e^I. \end{aligned}$$

[4]Specifically, we assume $\alpha \geq \frac{(r+c_g)\left(\overline{p}(e^I) - \overline{p}(0)\right)}{(r+c_g)\overline{p}(e^I) - (c_g + c_s)}$. For details, see the proof of Proposition 5 in the on-line appendix.

Proposition 5 *Under the partnership (PART) contract, the call center will both staff and exert effort at the system-optimal level. That is,* $s_{PART} = s^I$ *and* $e_{PART} = e^I$.

This partnership contract can also use the parameter α to divide the total profit. To see this, we note that at s_{PART} and e_{PART}, the call center's expected profit, from (17), becomes

$$\pi^c_{PART}(s_{PART}, e_{PART}) = \alpha \left\{ \left[r\overline{p}(e^I) - c_g \left(1 - \overline{p}(e^I)\right) - c_s \right] \lambda - c_e e^I \right\} = \alpha \cdot \pi^I(s^I, e^I). \qquad (18)$$

Corollary 3 *Under the partnership (PART) contract, the call center's share of the total profit is* α.

In order to achieve the full efficiency in call center outsourcing, a close relationship between the user company and the call center must be forged. It is not surprising that there is a direct correlation between the information content of a contract and its effectiveness in achieving system efficiency. In the PM contract, all the user needs to know is the call center's throughput T in order to calculate its contractual payment. This information is very easy to get. In most cases it is directly obtainable from the call center's automatic call distributor. In the PPCR contract, the user needs to have information on pT, the call center's resolved calls, or equivalently the revenue generated by the call center. It requires slightly more effort to get this information than just the throughput T. In the PPCR+CS contract, the user also needs to know the cost information of the call center in order to share a part of it. Moreover, the user has to share her own costs with the call center so that she could hold the call center responsible for suboptimal actions.

Collaboration and information sharing is especially important when the service quality is not directly observable and contractible. As demonstrated by the PART contract, the user first needs to share the staffing cost of the call center. Moreover, in order to induce the call center to expend more effort to improve service quality to meet system-optimality, the call center's profit margin needs to be adjusted to match that of the whole outsourcing supply chain. This requires both parties to share all of their cost information. This 'open book' approach demands a close collaboration between the two parties. Also, because the transactions are more complicated, each party needs to

keep close track of their bookkeeping. In short, coordination in outsourcing cannot be done in a simple 'hands-off' fashion, by some simple-term contract. The higher the efficiency an outsourcing supply chain wants to achieve, the closer the collaboration between outsourcing supply chain parties there needs to be. This also requires companies to commit themselves to administrative resources in outsourcing. Companies should go into outsourcing prepared for the costs involved in order to realize the anticipated gains from outsourcing. This is in contrast to the popular view that the objective of outsourcing is to cut as much cost as possible. In fact, according to a recent Deloitte Consulting study of 25 large organizations that have outsourced, 70% of them expressed, "Their costs of administering outsourcing contracts in many cases have been multiples of what they expected them to be."[5]

In practice we see companies such as General Electric and American Express choosing to open and operate their own ('captive') call centers in India (Sisk 2003) instead of completely outsourcing their call center operations. These call centers operate independently but maintain very close connections to the parent company in the United States. Other examples, though not in the call center industry, include major companies opening software development and research centers in India and China instead of completely outsourcing them. A major reason is the concern for quality. Our model's recommendation for close collaboration confirms these managerial decisions.

Finally in terms of social welfare, the PPCR+CS contract Pareto-dominates the PPCR contract, which in turn dominates the PM contract. The PART contract achieves the same social welfare as the PPCR+CS contract because they both coordinate the outsourcing supply chain.

Proposition 6

$$\begin{aligned}
&\pi^c_{PM}(s_{PM}, e_{PM}) + \pi^u_{PM}(s_{PM}, e_{PM}) \leq \pi^c_{PPCR}(s_{PPCR}, e_{PPCR}) + \pi^u_{PPCR}(s_{PPCR}, e_{PPCR}) \\
&\leq \pi^c_{PPCR+CS}(s_{PPCR+CS}, e_{PPCR+CS}) + \pi^u_{PPCR+CS}(s_{PPCR+CS}, e_{PPCR+CS}) \\
&= \pi^c_{PART}(s_{PART}, e_{PART}) + \pi^u_{PART}(s_{PART}, e_{PART}) = \pi^I(s^I, e^I). \qquad (19)
\end{aligned}$$

[5] http://www.thechannelinsider.com/article2/0,1759,1788202,00.asp

7 Numerical Analysis

In this section we conduct numerical analysis to gain more insights on the exact mechanism of the contracts studied in Sections 5-6. Without loss of generality, we set service rate μ at 1, and we let $\lambda = 200$ to simulate a medium to large size call center. We also use the following revenue and cost parameters: $r = 6, c_s = 1, c_g = 5, c_e = 90$.

Throughout the examples, we use the following function of call resolution capability p:

$$p\left(e\right) = \varepsilon \cdot \left(1 - \left(a + e\right)^{-k}\right), \tag{20}$$

where ε is a random variable with mean 1. Because $\overline{p}(0) = 1 - a^{-k}$, a measures the base level of call resolution. We will let $a = 1.5$.

We choose this function because the expected call resolution probability $\bar{p}(e)$ is increasing and concave in e. Moreover, by varying the value of k we can get a family of functions with different marginal effectiveness of effort. We have tested different values of k and have obtained similar results, indicating that the results are robust. To simplify exposition, we only present the case of $k = 2$ below. It should be noted that we have also conducted our numerical analysis using other function forms, e.g., $p\left(e\right) = \varepsilon\left(1 - a \cdot \exp(-k\, e)\right)$, and obtained similar results.

7.1 The integrated system

First, we investigate the shape of the expected profit of the integrated supply chain. In Figure 1 we plot an example of the total profit as a function of staffing s and effort e.

The expected profit is jointly concave in effort and staffing. The gradient is much greater when either effort or staffing is small (below optimum), and becomes much smaller when effort and staffing level become large (above optimum). In particular, for any fixed effort level, the profit curve is rather steep when $s < s^I$, but becomes much flatter when $s > s^I$. This is consistent with the numerical findings in Whitt (2004). Whitt found that over-staffing is generally more desirable than under-staffing, because the profit decreases slower when a call center is over-staffed than when it is under-staffed. In addition, our example reveals that over-exertion of effort is more desirable

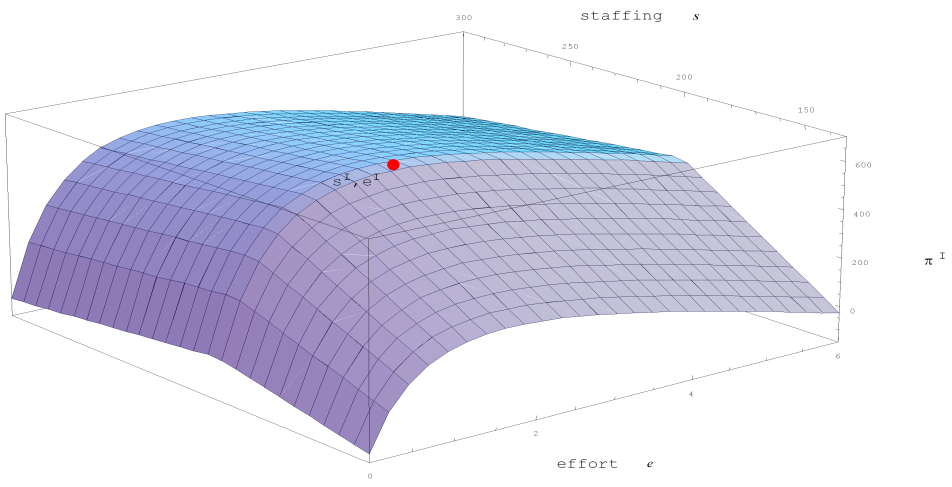

Figure 1: Expected profit for the integrated system

than under-exertion: for any staffing level, the expected profit curve is much steeper when $e < e^I$ than when $e > e^I$.

Next, we study the sensitivity of optimal system profit to the parameters. At the system-optimal s^I and e^I, there is no customer waiting or abandonment, so c_w and c_a are irrelevant. Because $\partial \pi^I(s^I, e^I)/\partial c_s = -\lambda$, the relationship between $\pi^I(s^I, e^I)$ and c_s is linear. The relationships between $\pi^I(s^I, e^I)$ and the other parameters λ, r, c_e, and c_g, however, are not so obvious. We have conducted a comprehensive numerical analysis, and the results are presented in Figure 2. Each panel in the figure varies one variable at a time, while keeping the others fixed. Other combinations of parameter values give similar graphs.

Figure 2 confirms the directional results that one would predict from the model. Specifically, the expected profit is increasing in demand rate λ and revenue rate r, but decreasing in effort cost c_e and loss of goodwill cost c_g. Even though analytically π^I is non-linear in all the four parameters, we note that numerically π^I is fairly linear in λ, r, and c_g. It is visibly more convex in c_e.

7.2 The decentralized system

In this section we study how the four contracts - PM, PPCR, PPCR+CS, and PART - succeed (or fail) to induce the call center to achieve system-optimal staffing and effort. In Figure 3 we

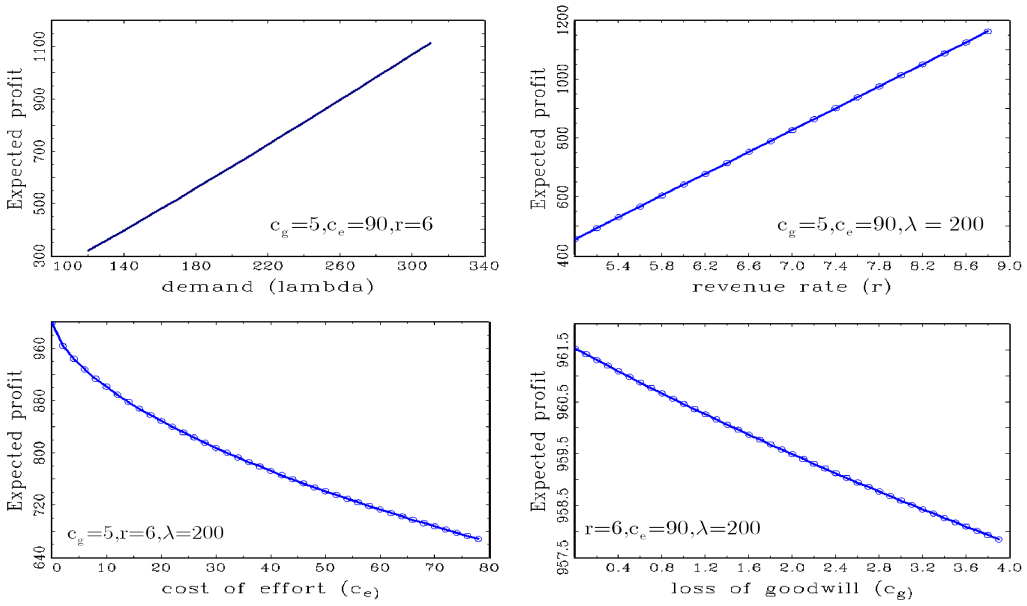

Figure 2: Effect of λ, r, c_e and c_g. ($c_s = 1$)

have plotted the call center's expected profit against effort level. The specific parameter values are: $\lambda = 200$, $c_s = 1$, $r = 6$, $c_g = 5$, $a = 1.5$, $k = 2$, and $c_e = 90$. For the PM, PPCR, and PPCR+CS contracts, the unit transfer price is $b = 3$. The corresponding profit split ratio for the PPCR+CS and PART contracts is $\alpha = b/r = 0.50$.

All the contracts induce the call center to staff at $s^I = 200$. The effort they induce differ, however. Table 1 shows their effort levels and profit breakdown.

	CC effort	Call resolution $\overline{p}$	CC profit	User profit	Total profit
PM	0.00	55.56%	400.00	−377.78	22.22
PPCR	0.87	82.22%	214.88	315.45	530.33
PPCR+CS	2.16	92.52%	320.69	320.69	641.37
PART	2.16	92.52%	320.69	320.69	641.37

Table 1: Expected Profit for user and outsourcer

In this example, the system-optimal effort level is $e^I = 2.16$, and the corresponding call resolution probability is 92.52%. Just as we have analytically shown, the PM contract fails to induce any

effort from the call center, which leads to an inferior service quality with only 55.56% of the served calls resolved. Because the user has to pay the call center $b = 3$ for each call served, resolved or not, its profit is negative. In practice, of course, the user would either not outsource the call center function at all, or choose a different b to render a positive profit.

The user can do better by using a PPCR contract. This contract induces the call center to exert a positive effort, at $e_{PPCR} = 0.87$ (still less than half of the system optimum). Service quality is improved to 82.22% calls resolved. Both the call center and the user now enjoy a positive profit. They can do even better, however, if they use either the PPCR+CS or the PART contract, both of which coordinate the system effort. Under these two contracts, the overall 'pie' (system profit) is bigger ($641.37 > 530.33$) and each party gets a bigger piece than before ($320.69 > 214.88$ for the call center, and $320.69 > 315.45$ for the user).

Even though both the PPCR+CS and PART contracts coordinate the system, they do so with different mechanisms. We show that in Figure 3.

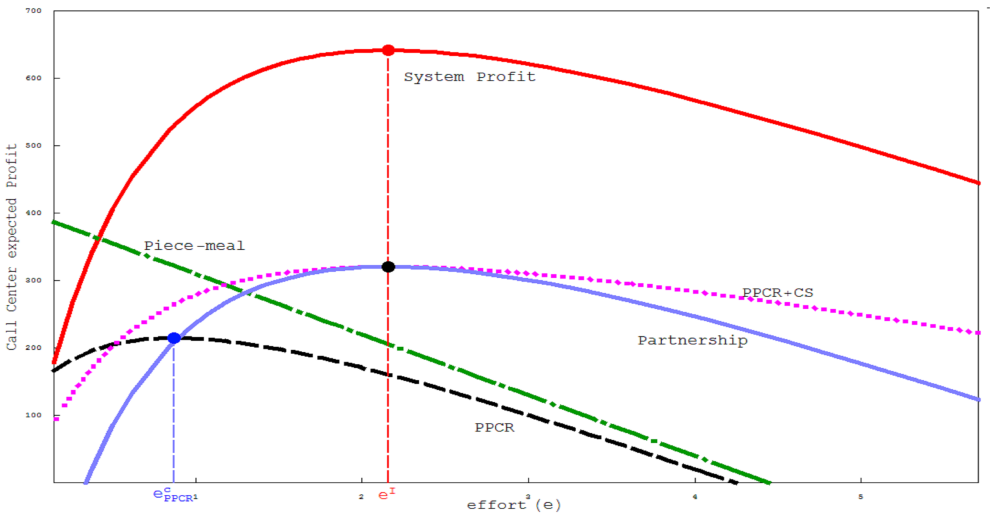

Figure 3: Call center expected profit

The curve for the PART contract is parallel to that for the integrated system. This is not surprising because under the PART contract the user basically gets a constant profit. In comparison,

under the PPCR+CS contract, the user shares a fixed portion of the total system profit. Therefore, when the effort level is off the system-optimal level, e^I, the user's (the call center's) profit is always lower (higher, resp.) under the PPCR+CS contract than under the PART contract. This observation is important because in practice, estimates of costs and call volumes are rarely precise. Therefore, the *a priori* system-optimal levels can be different from the *a posteriori* ones. Figure 3 suggests that from the user's point of view, the PART contract is more robust and should be preferred. However, it requires closer collaboration and information sharing between the user and the call center. This presents an interesting trade-off.

For each contract, there is a parameter that needs to be negotiated and determined (the decision variable). For the PM and PPCR contracts, the only contract parameter is the unit transfer price b, and for the PART contract it is the profit share ratio α. For the PPCR+CS contract, either b or α can be the decision variable, but not both, because they are related as $\alpha = b/r$.

Up to this point, we have fixed the contract parameter (b or α) to study the various types of contract. Next, we will study the effect of the contract parameter. Specifically, we are interested in the effect of b or α on the following four quantities under each contract: (1) call center's profit; (2) user's profit; (3) supply chain profit; and (4) the gap between the decentralized supply chain's profit and that of the integrated system. The results are given in Table 2 and illustrated in Figure 4.

		PM			PPCR			PPCR+CS			PART		
b	α	cc	user	system	cc	user	system	cc	user	system	cc	user	system
1.80	30.0%	160.00	-137.78	22.22	25.00	380.00	405.00	192.41	448.96	641.37	192.41	448.96	641.37
2.20	36.7%	240.00	-217.78	22.22	86.32	375.08	461.41	235.17	406.20	641.37	235.17	406.20	641.37
2.60	43.3%	320.00	-297.78	22.22	149.79	351.31	501.10	277.93	363.44	641.37	277.93	363.44	641.37
3.00	50.0%	400.00	-377.78	22.22	214.88	315.45	530.33	320.68	320.68	641.37	320.68	320.68	641.37
3.40	56.7%	480.00	-457.78	22.22	281.24	271.32	552.56	363.44	277.93	641.37	363.44	277.93	641.37
3.80	63.3%	560.00	-537.78	22.22	348.63	221.24	569.88	406.20	235.17	641.37	406.20	235.17	641.37
4.20	70.0%	640.00	-617.78	22.22	416.88	166.73	583.61	448.96	192.41	641.37	448.96	192.41	641.37
4.60	76.7%	720.00	-697.78	22.22	485.86	108.80	594.66	491.72	149.65	641.37	491.72	149.65	641.37

Table 2: Contract parameters and profits.

Under a PM contract, the call center will always staff at s^I and exert no effort, regardless of the value of b. Therefore, b, as a transfer from the user to the call center, affects how the total system profit is divided, but not the total profit itself. This is clearly reflected in the first panel

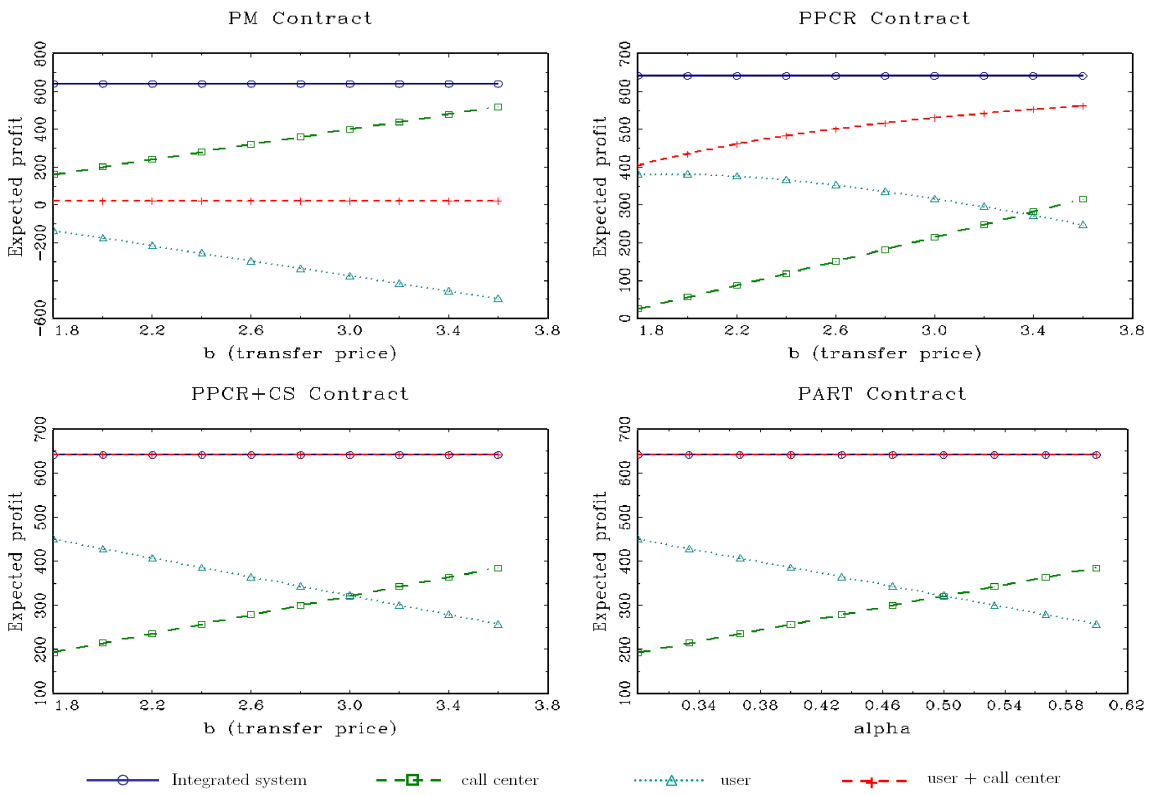

Figure 4: Contract parameters and profits

of Figure 4. In this example, to avoid negative profit, the call center should not be willing to pay a unit transfer fee of $b > 1.11$. It is important to notice the large total system profit gap between the PM contract and coordinated system (22.22 vs. 641.37).

Under the PPCR contract (upper right panel), the unit transfer fee b is for each call served and resolved. This induces the call center to exert positive effort, and consequently the profit gap between the centralized and the decentralized systems is greatly reduced. Because the user bears all the costs associated with low resolution probability under the PM contract, it stands to reap most of the benefit in narrowing the gap. As b increases, two phenomena take place: On the one hand, the incentive to resolve calls is higher, so the call center's effort level gets closer to the integrated system-optimal level, and the system profit gap narrows further. On the other hand, the transfer price from the user to the call center is higher, so the user's profit decreases, even as the system

profit increases. Therefore it is beneficial for the user to choose a small b in the PPCR contract, even though it hurts the call center as well as the whole chain.

System optimality is achieved by the PPCR+CS and PART contracts. The choice of b or α affects the division of the system profit. The user obviously prefers a smaller b or α, but there is a limit to how low it can go, because in order for the call center to cooperate, its profit should be no less than that under an alternative contract. In the above example, the call center's profit is 214.88 for $b = 3$ under the PPCR contract. If the user switches to a PPCR+CS or PART contract, it needs to choose at least $b = 2.01$, or $\alpha = 33.5\%$.

8 Conclusions and Future Research

When a user company outsources its call center operations to an outside company, they form an outsourcing supply chain. In this paper we study contract design issues in such an outsourcing supply chain. Specifically, we focus on chain coordination, where a contract is designed to induce the call center to both staff and exert effort at the system-optimal levels. The user and the call center should also be able to negotiate the parameters of a coordinating contract so that each can achieve a higher profit than when a non-coordinating contract is used.

We model the call center as a multi-server queue with customer abandonment, and use a fluid approximation to describe the steady-state behavior of the call center. Four contracts are studied: piecemeal, pay-per-call-resolved, pay-per-call-resolved plus cost sharing, and partnership contracts, and we show that all of them coordinate the staffing level, but induce different effort levels from the call center. The piecemeal contract induces the call center to exert no effort, while the pay-per-call-resolved contract induces positive, but still below system-optimal, effort from the call center. When effort is observable and contractible, we propose a pay-per-call-resolved plus cost sharing contract that can coordinate the outsourcing supply chain. When the call center's efforts are not observable, we propose a partnership contract that coordinates the supply chain. We argue that in order to achieve coordination, the user company and the call center must collaborate closely.

Our model has its limitations. First, because we use fluid approximation in the call center multi-server queueing system, our model is best applied to large call centers. Second, we have assumed constant system parameters such as arrival rate, unit costs, and unit revenue. It would be interesting to see how the contracts would perform or how the contract parameter should be determined when there is uncertainty about these system parameters. The comparison of PPCR+CS and PART contracts in Figure 3 is a starting point; much more work is needed. Third, we have assumed that the arrival rate is time-invariant, and we have used a linear cost structure, both for analytical tractability. It would also be interesting to see how our model can be extended along these dimensions, and how our results would change. Finally, we have assumed that the user company outsources all of its call center operations. There are certainly cases where partial outsourcing is possible, or even desirable (e.g., Akşin et al. 2004, Gans and Zhou 2005). It would be useful to analyze the contracts we propose in these contexts.

Acknowledgement

The authors thank Steve Graves, Sameer Hasija, Rob Shumsky, Ward Whitt, two anonymous referees, and participants at the MIT Sloan OM reading group for their helpful comments and suggestions. This research was supported by a junior faculty research grant from the Boston University School of Management, and the Center for International Business Education and Research (CIBER) at the University of Washington Business School.

References

[1] Aguir M.S., O.Z. Akşin, F. Karaesmen, and Y. Dallery (2004). "On the Interaction Between Retrials and Sizing of Call Centers." Working Paper, Koç University.

[2] Akşin, Z., F. de Véricourt, and F. Karaesmen (2004). "Call Center Outsourcing Contract Design and Choice." Working paper, Koç University.

[3] Andrews, B. and H. Parsons (1993). "Establishing Telephone-Agent Staffing Levels Through Economic Optimization." Interfaces. 23(2):14-20.

[4] Bassamboo A., J. M. Harrison, and A. Zeevi (2005). "Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits". Queueing System. 51(3-4):249-285.

[5] Brandt A. and M. Brandt (1999). "On a Two-Queue Priority System with Impatience and its Application to a Call Center." Methodology and Computing in Applied Probability. 1(2):191-210.

[6] Brown L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao (2005). "Statistical Analysis of a Telephone Call Center: a Queueing Science Perspective." Journal of the American Statistical Association. 100(469):36-50.

[7] BusinessWeek (2003). "The Hidden Costs of Outsourcing." Business Week (October 27, 2003).

[8] Cachon, G. and M. Lariviere (2005). "Supply chain coordination with revenue sharing contracts." Management Science 51(1):30-44.

[9] Cachon, G. (2003). "Supply Chain Coordination with Contracts." Handbook of Operations Management. Edited by S. Graves and T. De Kok. North Holland.

[10] Chevalier P., Y. De Ronge, N. Tabordon, L. Talbot (1998). "Service Subcontracting: An Economic Analysis." Proceeding CEMS Conference, 1998.

[11] Cohen, M. A., T. H. Ho, Z.J. Ren, C. Terwiesch (2003). "Measuring Imputed Costs in the Semiconductor Equipment Supply Chain." Management Science 49(12):1653-1670.

[12] de Véricourt F. and Y.-P. Zhou (2005). "Managing Response Time in a Call-Routing Problem with Service Failure." Operations Research 53(6):968-981.

[13] Falin, G. I. and J.G.C. Templeton (1997). "Retrial queues." Chapman & Hall. London.

[14] Gans, N. (2002). "Customer Loyalty and Supplier Quality Competition." Management Science 48(2):207-221.

[15] Gans, N., G. Koole, and A. Mandelbaum. (2003). "Telephone call centers: Tutorial, review, and research prospects." Manufacturing & Service Operations Management 5(2).

[16] Gans, N. and Y.-P. Zhou (2005). "Call-Routing Schemes for Call-Center Outsourcing." Working paper, University of Washington Business School.

[17] Garnett, O., A. Mandelbaum, and M. Reiman (2002). "Designing a Call Center with Impatient Customers." Manufacturing and Service Operations Management, 4(3):208-227, 2002.

[18] Gurvich I., M. Armony, and A. Mandelbaum (2004). "Staffing and Control of Large-Scale Service Systems with Multiple Customer Classes and Fully Flexible Servers." Working Paper, New York University.

[19] Harrison, J. M., and A. Zeevi (2005). "A method for staffing large call centers based on stochastic fluid models." Manufacturing and Service Operations Management. 7(1):20-36.

[20] Helber, S. (2004). "Profit-Maximizing Multi-Period Agent Allocation in Different Types of Inbound Call Centers." Call Center Workshop at the Centre de Recherches Mathematiques, Universite de Montreal, Montreal, Canada. Presentation available at `http://www.crm.umontreal.ca/Stochastic2004/Level2/horaireSto23-25CCW.html`

[21] Hoffman, K. L. and C. M. Harris (1986). "Estimation of a caller retrial rate for a telephone information system." European Journal of Operational Research, 27: 207-214.

[22] Koole, G. and A. Pot (2004). "Profit maximization and monotonicity results for inbound call centers." Working paper, Vrije Universiteit.

[23] Mandelbaum, A. and N. Shimkin (2000). "A model for rational abandonments from invisible queues." Queueing Systems 36(1-3): 141-173.

[24] Parasuraman A., V.A. Zeithaml, and L.L. Berry (1990). "Delivering Quality Services." Free Press, New York, 1990.

[25] Pasternack, B. A. (1985). "Optimal pricing and returns policies for perishable commodities." Marketing Science 4:166-176.

[26] Shimkin, N. and A. Mandelbaum (2004). "Rational abandonment from tele-queues: Nonlinear waiting costs with heterogeneous preferences." Queueing Systems 47(1-2): 117-146.

[27] Shumsky, R. and E. Pinker (2003). "Gatekeepers and Referrals in Services." Management Science 49(7): 839–856.

[28] Sisk M. (2003). "U.S. Banks' Cost Cutting is Lucrative Gig for India." USBanker 113(9):32.

[29] Taylor, T. A. (2002). "Supply Chain Coordination under Channel Rebates with Sales Effort Effects." Management Science 48(8):992-1007.

[30] Tsay, A. A. (1999). "The quantity flexibility contract and supplier-customer incentives." Management Science 45(10): 1339-1358.

[31] United States Government Accountability Office (2004). "Current Government Data Provide Limited Insight into Offshoring of Services". Report to Congressional Requesters, Sept. 2004. Available at `http://www.gao.gov/new.items/d04932.pdf`

[32] Whitt, W. (1999). "Improving Service by Informing Customers about Anticipated Delays." Management Science. 45(2):192-207.

[33] Whitt, W. (2004a). "Efficiency-Driven Heavy-Traffic Approximations for Many-Server Queues with Abandonments", Management Science 50(10): 1449-1461.

[34] Whitt, W. (2004b). "Staffing a Call Center with Uncertain Arrival Rate and Absenteeism." To appear in Production and Operations Management.

[35] Whitt, W. (2006). "Fluid Models for Many-Server Queues with Abandonments." Operations Research. 54(1):37-54.

[36] Zeltyn, S., A. Mandelbaum, and M. Reiman (2004). "Call Centers with Impatient Customers: Many-Server Asymptotics of the M/M/n+G queue." Working paper, Technion, Israel.

[37] Zohar, E., A. Mandelbaum, and M. Reiman (2002). "Adaptive behavior of impatient customers in tele-queues: Theory and empirical support." Management Science 48(4): 566-583.

Appendix for

Call Center Outsourcing: Coordinating Staffing Level and Service Quality

Appendix

In this appendix, we prove Propositions 1, 2, 3, 4, and 5. They derive the call center's optimal staffing and effort levels, s^* and e^*, in the integrated system, and under the PM, PPCR, PPCR+CS, and PART contracts respectively. The proofs are similar; each follows these three steps:

Step 1 For any given effort level e, show that when $s \geq \lambda/\mu$, the profit function is decreasing in s.

Step 2 For any given effort level e, show that when $s \leq \lambda/\mu$, the profit function is increasing in s. Together with step 2, this implies $s^* = \lambda/\mu$.

Step 3 For $s^* = \lambda/\mu$, the profit function is concave in e, so e^* can be found via the first order condition.

Proof of Proposition 1. The expected profit function, (6), can be simplified to:

$$\pi^I(s,e) = [r\overline{p}(e) - c_g(1-\overline{p}(e))]\,\lambda - \left[r\overline{p}(e) + c_a + \frac{c_w}{f\,(0)} - c_g\,(1-\overline{p}(e))\right] L(s) - c_s\mu s - c_e e.$$

Step 1 When $s \geq \lambda/\mu$, there is no abandonment, $L(s) = 0$, and

$$\pi^I\,(s,e) = [r\overline{p}(e) - c_g(1-\overline{p}(e))]\,\lambda - c_s\mu s - c_e e.$$

This is decreasing in s.

Step 2 When $s \leq \lambda/\mu$, $L(s) = \lambda - s\mu \geq 0$, and

$$\pi^I\,(s,e) = -\left[c_a + \frac{c_w}{f\,(0)}\right]\lambda + \left[r\overline{p}(e) + c_a + \frac{c_w}{f\,(0)} - c_g\,(1-\overline{p}(e)) - c_s\right]\mu s - c_e e.$$

This is increasing in s.

Step 3 When $s^I = \lambda/\mu$, and the expected profit reduces to

$$\pi(s^I,e) = [r\overline{p}(e) - c_g(1-\overline{p}(e)) - c_s]\,\lambda - c_e e = (r+c_g)\,\lambda\overline{p}(e) - c_e e - (c_g+c_s)\,\lambda. \tag{21}$$

Note that one necessary condition for the expected profit to be positive is that $r\overline{p}(e)+c_a+\frac{c_w}{f(0)}-c_g\,(1-\overline{p}(e))-c_s > 0$, which is satisfied by the assumption $r\overline{p}\,(0) > c_s+(1-\overline{p}\,(0))c_g$.

Because $\overline{p}(e)$ is concave in e, it is easy to see from (21) that $\pi\left(s^I, e\right)$ is concave in e. The first order derivative of (21) with respect to e is:

$$(r + c_g)\,\lambda\overline{p}'(e) - c_e. \tag{22}$$

Therefore, when $\frac{c_e}{(r+c_g)\lambda} \geq \overline{p}'_0$, the marginal cost is too big relative to the marginal revenue to be worth spending any effort. Otherwise, the optimal effort can be found by $\overline{p}'\left(e\right) = \frac{c_e}{(r+c_g)\lambda}$. That is,

$$\overline{p}'\left(e^I\right) = \frac{c_e}{(r + c_g)\,\lambda}, \text{if } \frac{c_e}{(r + c_g)\,\lambda} < \overline{p}'_0;\ e^I = 0 \text{ o.w.}$$

■

Proof of Proposition 2. Using $T(s) = \lambda - L(s)$, we reduce (11) to:

$$\pi^c_{PM}(s, e) \quad = \quad b\lambda - bL(s) - c_s\mu s - c_e e.$$

Step 1 When $s \geq s^I = \lambda/\mu$, $L(s) = 0$ and $\pi^c_{PM}(s, e) = b\lambda - c_s\mu s - c_e e$. This is decreasing in s.

Step 2 When $s \leq s^I = \lambda/\mu$, and $\pi^c_{PM}(s, e) = (b - c_s)\,\mu s - c_e e$. This is increasing in s.

Step 3 Using $s_{PM} = s^I = \lambda/\mu$, we find that $\pi^c_{PM}(s_{PM}, e)$ is decreasing in e. Therefore $e_{PM} = 0$.

■

Proof of Proposition 3. Again, using $T(s) = \lambda - L(s)$, we reduce (12) to:

$$\pi^c_{PPCR}(s, e) = b\overline{p}(e)\lambda - b\overline{p}(e)L(s) - c_s\mu s - c_e e.$$

Step 1 When $s \geq s^I = \lambda/\mu$, $L(s) = 0$ and $\pi^c_{PPCR}(s, e) = b\overline{p}(e)\lambda - c_s\mu s - c_e e$. This is decreasing in s.

Step 2 When $s \leq s^I = \lambda/\mu$, $\pi^c_{PPCR}(s, e) = (b\overline{p}(e) - c_s)\,\mu s - c_e e$. There are two cases:

1. $b\overline{p}(e) < c_s$. In this case, $\pi^c_{PPCR}(s, e)$ is decreasing in s. It is best to set $s = 0$.

2. $b\overline{p}(e) \geq c_s$. In this case, $\pi^c_{PPCR}(s, e)$ is increasing in s. It is best to set $s = s^I = \lambda/\mu$.

Step 3 Therefore, the optimal staffing level is either 0 or s^I. It is easy to see that when $s = 0$, the optimal e is zero as well. When $s = s^I$, we substitute this back into the profit function, and get

$\pi^c_{PPCR}(e) = (b\overline{p}(e) - c_s)\,\lambda - c_e e$. This is concave in e so the first order condition yields the optimal effort level: $\overline{p}'\,(e_{PPCR}) = \frac{c_e}{b\lambda}$.

Our condition on b indicates that the call center gets a higher profit at (s^I, e^I) than zero (the alternative). So $s_{PPCR} = s^I$ and $e_{PPCR} = e^I$. ■

Proof of Proposition 4. We can rewrite the call center's expected profit under the PPCR+CS contract as:

$$\pi^c_{PPCR+CS}(s,e) = \alpha\left[r\overline{p}(e) - c_g\left(1-\overline{p}(e)\right)\right]\lambda - \alpha\left[r\overline{p}(e) - c_g\left(1-\overline{p}(e)\right) + c_a + \frac{c_w}{f\,(0)}\right]L(s) - \alpha c_s\mu s - \alpha c_e e. \quad (23)$$

Step 1 When $s \geq s^I = \lambda/\mu$, $\pi^c_{PPCR+CS} = \alpha\left[r\overline{p}(e) - c_g\left(1-\overline{p}(e)\right)\right]\lambda - \alpha c_s\mu s - \alpha c_e e$. This is decreasing in s.

Step 2 When $s \leq s^I = \lambda/\mu$, $L(s) = \lambda - \mu s$, and

$$\pi^c_{PPCR+CS}(s,e) = -\alpha\left[c_a + \frac{c_w}{f\,(0)}\right]\lambda + \alpha\left[r\overline{p}(e) - c_g\left(1-\overline{p}(e)\right) + c_a + \frac{c_w}{f\,(0)} - c_s\right]\mu s - \alpha c_e e. \quad (24)$$

This is increasing in s.

Step 3 Substituting $s_{PPCR+CS} = s^I = \lambda/\mu$ in (23) we get

$$\pi^c_{PPCR+CS}(e) \;=\; \alpha(r+c_g)\lambda\overline{p}(e) - \alpha c_e e - \alpha(c_g+c_s)\lambda,$$

Since this is concave in e, $e_{PPCR+CS}$ can be found via its first order derivative: $\alpha\left[(r+c_g)\lambda\overline{p}'(e) - c_e\right]$, which differs from the first order derivative for the integrated system, (22), only by a constant multiplicative factor α. Therefore, $e_{PPCR+CS} = e^I$. ■

Proof of Proposition 5.

The call center's expected profit under the partnership contract is

$$\begin{aligned}\pi^c_{PART}(s,e) \;=\; & \left[(r+c_g)\left(\overline{p}(e) - (1-\alpha)\,\overline{p}(e^I)\right) - \alpha c_g\right]T(s) - \alpha c_s\mu s \\ & + (1-\alpha)\,c_e e^I - c_e e - \alpha\left(c_a + \frac{c_w}{f(0)}\right)L(s).\end{aligned} \quad (25)$$

Step 1 When $s \geq s^I = \lambda/\mu$, $L(s) = 0$, and $T(s) = \lambda$. The only term in (25) that depends on s is $-\alpha c_s\mu s$ and it's decreasing in s.

Step 2 When $s \leq s^I = \lambda/\mu$, $T(s) = \mu s$, and

$$\pi^c_{PART}(s,e) = \left[(r+c_g)\left(\overline{p}(e) - (1-\alpha)\,\overline{p}(e^I)\right) - \alpha(c_g+c_s)\right]\mu s + (1-\alpha)\,c_e e^I - c_e e.$$

Our assumptions on the parameters, $r\overline{p}(0) > c_s + (1-\overline{p}(0))c_g$ and $\alpha \geq \frac{(r+c_g)\left(\overline{p}(e^I)-\overline{p}(0)\right)}{(r+c_g)\overline{p}(e^I)-(c_g+c_s)}$, imply that the profit is increasing in s.

Step 3 Therefore, the optimal staffing level is s^I. We substitute this back into the profit function (25), and get

$$\pi^c_{PART}(s^I,e) \quad = \quad \left[(r+c_g)\left(\overline{p}(e) - (1-\alpha)\,\overline{p}(e^I)\right) - \alpha(c_g+c_s)\right]\lambda + (1-\alpha)\,c_e e^I - c_e e.$$

Since this is concave in e, it suffices to examine its first order derivative with respect to e: $(r + c_g)\lambda\overline{p}'(e) - c_e$. Because it is identical to (22), it is best to set $e = e^I$.

Clearly between the two options, it is better to have $s_{PART} = s^I$ and $e_{PART} = e^I$. ■