Call Center Outsourcing Contracts Under Information Asymmetry

Sameer Hasija
School of Management, Binghamton University, State University of New York, Binghamton, NY, 13902, shasija@binghamton.edu

Edieal J. Pinker
W. E. Simon Graduate School of Business Administration University of Rochester, Rochester, NY, 14627, pinker@simon.rochester.edu

Robert A. Shumsky
Tuck School of Business Administration, Dartmouth College, Hanover, NH, 03755, robert.shumsky@dartmouth.edu

In this paper we examine contracts to coordinate the capacity decision of a vendor who has been hired by a client to provide call center support. We consider a variety of contracts, all based on our observations of contracts used by one large vendor. We examine the role of different contract features such as pay-per-time, pay-per-call, service level agreements, and constraints on service rates and abandonment. We show how different combinations of these contract features enable client firms to better manage vendors when there is information asymmetry about worker productivity. In particular we focus on how different contracts can *coordinate* by yielding the system-optimal capacity decision by the vendor and consider how profits are allocated between the client and the vendor.

Key words: Call Center; Outsourcing; Contracts; Service Supply Chains

History: This paper was first submitted October, 16, 2006, and has been with the authors for 95 days for 2 revisions.

1. Introduction

Call centers are essential components of many large businesses. While some firms choose to create internal call centers, many now purchase call center support services from other firms. In a typical outsourc-

ing arrangement, a firm (the *client*) hires a call-center specialist (the *vendor*) to provide sufficient technology and staff to serve the client's customers. The client specifies the quality of service and the financial terms in a detailed contract, which may include queueing performance criteria (e.g. 80% of callers wait less than 20 sec.), customer satisfaction requirements (as measured by surveys or observed by monitoring calls), and financial rewards and penalties. Motivated by contracts used by one large vendor, this paper examines how the terms of real-world call center outsourcing contracts affect the capacity investment decisions of the vendor as well as the financial performance of the client, vendor, and the system as a whole.

In this paper we will sometimes refer to the system as a *service supply chain* or simply as a supply chain, for the relationship between the client and the vendor is analogous to the relationship between a retailer and its supplier. The client, like the retailer, purchases capacity from the vendor/supplier. On one dimension, however, the relationship between our client and vendor reverses the typical retailer/supplier relationship. In our environment the vendor interacts directly with the customer, while the supplier in a supply chain does not. Therefore, our client does not directly order an observable volume of service from the vendor, as a retailer would order a specific number of units. Instead the vendor serves the client's customers by performing a variety of functions that are often unobservable to the client, such as hiring, training and investing in technology. Payment is usually contingent on the vendor serving realized demand according to criteria specified in the contract. The client uses the contract to influence the unobservable behavior, and poor contract design can lead to vendor actions that reduce client profits and supply-chain performance (see the introduction to Ren and Zhou, 2006, for additional comparisons between call-center outsourcing and the traditional supply chain).

In this paper we model the vendor's actions as two decisions, a staffing level and a service rate that are chosen to maximize its profits under a given contract. Our model of the vendor's service system is a Markovian queuing system with exponential abandonment ($M/M/N+M$). We assume that the client designs and proposes the contract while the vendor may accept or refuse the offer. Because poor service

can lead to lost future sales and, in the case of an inbound direct sales channel, immediate lost sales, the client loses revenue as its customers wait in line and abandon the queue. The vendor does not incur these costs but must pay the staffing costs. We say that the service supply chain is *coordinated* when the vendor chooses a staffing level that maximizes the total supply chain profit, the sum of the vendor and client profits.

The contractual terms modeled in this paper were motivated by contracts signed by a vendor with 15,000 employees that provides call center support to Fortune 500 technology and financial services firms. Table 1 displays a representative sample of these contracts. The rows in the table represent different contracts with clients (A) - (F). The second column lists the waiting-time target, or service level agreement (SLA) for each contract, such as $\Pr\{\text{wait} \le 20\,\text{sec.}\} \ge 0.75$ for client (E). The third column lists financial incentives and penalties. The term "SLA penalty" implies that the vendor pays a financial penalty for not meeting an SLA. AHT (average handle time) is the average service time per customer and "AHT penalty" means that the vendor pays a financial penalty for going over an AHT target (or going under a service-rate target) set by the outsourcer. Table 2 summarizes the definitions of these abbreviations, which will be used throughout the paper.

Client	Service level agreement (SLA)	Financial Incentives
(A)	90% of calls answered within 360 sec	PPC; SLA penalty; Monthly payment limit
(B)	70% of calls answered within 60 sec; 70% of calls resolved without escalation.	PPC; SLA penalty
(C)	80% of calls answered within 120 sec	PPT; SLA penalty; AHT penalty
(D)	80% of calls answered within 180 sec	PPT; SLA penalty; AHT penalty; Guaranteed base payment.
(E)	75% of calls answered within 20 sec	Pay per resolution; SLA penalty; Guaranteed base payment.
(F)	80% should be attempted	PPT; Abandonment probability penalty; AHT penalty

Table 1: Sample contracts

Abbreviation	Contract Term
PPC	Pay per call
PPT	Pay per time
SLA	Service level agreement
W	Linear penalty for waiting
Ab	Abandonment probability constraint
AHT	Penalty for not meeting average handle time (service-rate constraint)

Table 2: Abbreviations for contract terms

The third column of Table 1 also describes two payment mechanisms: pay per call (PPC) and pay per time (PPT). Under a PPC mechanism the vendor earns a fixed fee from the client for each customer it serves. PPT schemes compensate the vendor per unit time that it spends serving customers (this contact time is easily monitored by the client, who can observe the telecommunications switch that is shared by the client and vendor). In our sample of contracts PPT schemes are always accompanied with penalties

for not meeting AHT targets. An explanation of this last observation is that the PPT compensation scheme provides an incentive to the vendor to increase the AHT, and the AHT penalty limits this behavior.

On first glance, a PPC or PPT term in the contract seems to be superfluous. Given the significant penalties for failure to meet the SLA, if all other aspects of the contract are eliminated then the vendor's staffing rule is to assign a sufficient number of agents so that the SLA constraint is tight. As long as the SLA is set appropriately by the client and sufficient payment passes between the client and the vendor (most simply as a lump-sum), the vendor will accept the contract, the vendor will staff at the level desired by the client, and the client will retain any excess profits. Indeed, contracts for (D) and (E) in the table have guaranteed base payments.

However, the vendor's managers stated to us that under most contracts the vendor is compensated by PPC or PPT, either as the sole payment method or as a supplement to the base payment. There are a variety of plausible explanations for this. For example, it may be a convenient method for spreading out a lump-sum payment, following the principle that the vendor is paid when it does the work. Furthermore, if the level of demand is uncertain, both PPC and PPT contract mechanisms reduce the vendor's risk of large losses, for the vendor will be compensated if a demand surge requires it to add expensive capacity to meet the SLA. While in some environments this may be the reason for PPC or PPT terms, in our work we assume that the mean demand rate can be determined accurately and is known to both the vendor and the client.

In this paper we explore an alternate role for PPC and PPT contracts: they allow the client to overcome information asymmetry with respect to the vendor's potential productivity. When clients negotiate terms of the contract with the vendor, the vendor may have significantly more information on the maximum possible service rates of its own agents. This information asymmetry may be caused by a variety of factors. For example, the vendor hires and trains agents and

thus can better assess their potential productivity. Often the vendor provides similar services to other clients and therefore has more experience and data that can be used to forecast productivity. The latter explanation becomes increasingly plausible as more firms outsource their call-center operations and retain less knowledge about their own customer-service processes.

We show that under this information asymmetry, when the client is restricted to certain types of contracts (PPC or PPT without AHT constraints), the vendor may invest in the supply-chain optimal capacity but the vendor also extracts information rents – it captures a significant portion of supply-chain profits. By offering both PPC and PPT-based contracts rather than a single contract type, the client can reduce these information rents by screening the vendors without a significant loss in overall supply chain performance. We also show that when the client has complete information about the vendor's productivity then there is no need to include an average handle time constraint (AHT) in an optimally-designed PPT contract. When there is information asymmetry on vendor productivity then an AHT constraint increases the client's profits and improves chain performance. Therefore, the existence of AHT constraints in the PPT contracts signed by our vendor is consistent with our model of information asymmetry.

Our results on PPC and PPT contracts are analogous to basic results from labor economics which suggest that variable pay can be used by a firm to sort low and high-productivity workers (Lazear, 1995). Specifically, a PPT contract is like an hourly wage; it specifies payment for input. A PPC contract is similar to a piece-rate contract on outputs. In general, our model is a monopolistic screening model with precontractual asymmetric information (see, for example, Mas-Colell et al., 1995, pg. 500), in which a worker's productivity type is unobservable to the principal before a contract is signed, and the output (but not the productivity type) of the worker can be observed after the contract is signed. We emphasize, however, that our model is not a

simple extension of the monopolistic screening model described in Mas-Colell et al. Our application has an additional layer of complexity, the vendor's stochastic queueing/staffing problem.

Another significant difference between our model and standard models from labor economics is that in our case the PPT and PPC contracts are not used to weed out (or avoid hiring) inefficient vendors. Because of the large fixed cost to select and establish a service relationship with a vendor, clients are reluctant to switch vendors and instead must design contracts to extract the best performance possible from a favored vendor. Using a well-designed contract to 'get it right the first time' has significant value in these settings when compared with costly alternatives such as careful onsite monitoring of the vendor or renegotiation after a probationary period. Onsite monitoring of call centers to determine if workers are as productive as they could be is fraught with challenges. Knowledge of the local labor pool and training methods is necessary to set realistic performance goals, and both local labor conditions and training regimes are difficult to monitor. When the client firm has outsourced the function it is less able to effectively benchmark performance measures. When call centers are off-shore, monitoring is more expensive. While clients often monitor for quality in the customer interactions, simultaneous productivity monitoring may lead to conflicting motivations. As a result, information asymmetry about agent productivity often persists after the contract is signed and operations commence. Despite all the challenges of accurate monitoring, inevitably over time, the client will learn more about the ability of the vendor and can take that into account when renegotiating contract terms. The analysis of renegotiation across multiple contracting periods is an interesting area for research but beyond the scope of this paper.

Our model does not address two performance criteria that are seen in some call center outsourcing contracts: (i) escalation behavior and (ii) customer quality measures besides waiting-time. In some busi-

ness environments service requests are vertically differentiated, and we see multi-tiered service centers. In such systems, a lower-tier (less skilled) agent can escalate a service request to a higher-tier agent if he cannot resolve the issue. Sometimes this higher-tier agent is employed directly by the client. In such cases, contracts include terms to influence the vendor's escalation policy (see the contracts for clients B, E, and F in Table 1). Shumsky and Pinker (2003) show how such incentives can induce system-optimal escalation policies within a single firm. Ren and Zhou (2006) show how a pay per resolution contract affects the effort exerted by the vendor towards increasing the call resolution rate. The models in this paper are limited to systems without call escalation, for we assume that all calls are successfully resolved by the vendor.

The call center managers that we interviewed emphasized the importance of quality measures beyond waiting-times, and their clients administered frequent customer satisfaction surveys and pushed the vendor to keep the CSAT (customer satisfaction) scores high. It is true that certain terms of the contracts shown in Table 1 may have an impact on service quality. For example a pay-per-time contract may lead agents to spend more time with each customer. In some cases this may lead to a perception of better service while in others it may be viewed as a degradation of service quality. In this paper, however, we focus on how the contract terms affect productivity rather than quality. Our focus on productivity was motivated by the fact that none of the contracts we examined included explicit terms based on quality measures such as the CSAT. We believe the primary reason for this is that compared to waiting time measures quality is difficult to measure reliably and so it is managed differently. Exploring this hypothesis will be an interesting area for further research.

In the next section we review the related literature. In section 3 we examine contracts under complete information. We show that contracts based only on PPC or PPT terms are generally not favorable for the client but that PPC or PPT contracts with SLA, waiting-time or abandonment penalties coordinate the service supply chain and allow for arbitrary allocation of the supply

chain profits between the client and the vendor. In section 4 we assume that the client only knows that the vendor is one of two productivity types: high or low. High (low) productivity corresponds to high (low) agent service rates. We show how the client can use PPC and PPT terms to screen the vendor type, coordinate the chain, and maximize client profits. In Section 5 we extend these results to a model in which the client has an arbitrary prior distribution on the service rate. We show that a single PPC contract can coordinate the chain for vendors that fall within a certain range of service rates, but the client must pay information rents to vendors with high productivity. We then show that by offering the choice of PPC or PPT contracts with waiting-time penalties the client can reduce the information rents and raise its profits, although the chain may not always be coordinated. In Section 6 we illustrate the use of PPC and PPT contracts with a numerical example. The numerical example shows how the contracts reduce information rents paid by the client and also show how these contracts can improve overall supply chain performance over a single PPC contract by expanding the range of vendors that accept the contract. Finally, in Section 7 we discuss possible future areas of research.

2. Literature Review

Considerable attention has been given to outsourcing contracts in manufacturing supply chains (see Cachon 2003 and the references therein). The literature on outsourcing contracts for service supply chains is more limited. Gans and Zhou (2007) and Aksin et al. (2006) consider a client who can outsource some fraction of service calls to a vendor. Gans and Zhou (2007) study the centralized capacity decision and queuing control problem. Aksin et al. (2006) compare the equilibrium performance of service systems in which the client either outsources a steady stream of calls or outsources peak demand. We assume here that all calls are routed to the vendor.

While we also assume that the client has already decided to outsource its call center to the vendor, Allon and Federgruen (2006) focus on the outsourcing decision. They study retailers who are locked in price and waiting-time competition and have the option to outsource call-center service to a common vendor. They present conditions under which outsourcing is profitable for the clients. A portion of their analysis describes the effects of "volume based" contracts on supply chain coordination and their results parallel our analysis of PPC contracts with full information in Section 3.

Aron and Liu (2003) examine coordination of the quality-of-service decisions made by vendors. They study governance systems where the client actively participates in the managerial process of monitoring and controlling the vendor's agents to ensure a desired quality level. They find that when the outsourced service process is complex so that the cost of measuring output quality is high then the clients can increase the efficiency and scope of outsourcing by combining the efficiency of the price mechanism (market control) with managerial control. They also show that for a low-complexity process there is no significant advantage for the client to exert managerial control over the vendor's agents.

Our work is closely related to the work of Ren and Zhou (2006). They study a service supply chain consisting of a single client and a single vendor and also consider contracts that induce the vendor to choose supply-chain optimal staffing levels. They assume that vendor productivity is common knowledge and focus on the vendor's level of effort, where higher effort increases the probability that a call earns revenue for the client. In their analysis, they use a fluid model that ignores the queueing phenomena, and the use of this approximation has a number of consequences. First, the optimal staffing level is equal to λ/μ, in other words a load factor of 1. If waiting and or abandonment costs are large, a fluid model distorts the staffing requirements. Second, they find that the client can use a contract with only a PPC component, a "piecemeal contract," to coordinate staffing levels (they also show that such a contract will not coordinate the effort level). In this paper we find that a PPC-only contract cannot be used by the client to coordinate staffing levels. The difference in our results is due to the underlying queueing model. Ren and Zhou use a fluid approximation, but when we include a stochastic component in the chain's prof-

it function a coordinating PPC-only contract produces negative revenue for the client. In practice, we have not observed instances of call center contracts based solely on PPC terms, indicating that a stochastic model is appropriate here. Throughout this paper, the stochastic model also allows us to examine the effects of contract terms related to waiting time, terms that are irrelevant in the fluid model.

3. Contracting with full information

3.1 The Model

In this section we describe the profit functions of the client and the vendor and determine which contracts coordinate the supply chain while allowing for arbitrary allocation of the supply chain profits. We assume that both firms are risk neutral, and in this section we also assume that all information on the system is shared by the client and vendor. The client offers a contract to the vendor and, if the vendor accepts the contract, the vendor makes its profit-maximizing capacity choice. The client's reservation value is M, i.e., the client can earn profit rate M if it chooses to enter a contract with a different vendor. The parameter M includes the search cost and the cost of building a new relationship and can possibly be very low. Therefore, the client offers the vendor a contract only if the client's expected profit rate under that contract is greater than M. The vendor accepts the contract if its expected profit is greater than its reservation value, V. If the vendor accepts the contract it then invests in capacity, a staffing level N. Finally, demand is realized and is served according to the dynamics of an $M/M/N+M$ queueing system.

Each served customer generates a value R for the client and the client also incurs a cost P per unit time the customer waits in queue. In a sales environment this is direct revenue. In other cases such as technical support the value, R, is a proxy for the net expected long term cash flow generated by a satisfied customer. We consider waiting time, rather than system time, as the measure of the customer's experience. Such an assumption is common in the literature, for waiting time is usually perceived by customers as a waste of productive time while the time spent in service may not be perceived as a cost (see, for example, Gans et al., 2003, Ren and Zhou, 2006, and Hasija et al., 2005). The vendor incurs a cost c for each agent

staffed and customers arrive according to a Poisson process with rate λ. The vendor's agents serve customers at rate μ, and each agent's service times are exponentially distributed. The customers of the client abandon the queue at rate θ, and the customer's abandonment time is exponentially distributed. $F(N)$ is the equilibrium probability of abandonment, given staffing level N. $F(N)$ is a non-increasing function in N. By Little's Law, the average queue length is $(\lambda/\theta)F(N)$. Also let $G(N,t) = \Pr\{wait < t\}$.

Throughout the paper we assume that N is continuous and we describe $F(N)$ and $G(N,t)$ using the diffusion approximations of Garnett et al. (2002). See the Appendix for the appropriate expressions. Garnett et al. have shown that their approximations work extremely well for calculating waiting-time tail probabilities and abandonment probabilities in systems with as few as 20-30 servers (see Appendix A of Garnett et al., 2002). Note, however, that results similar to all of the Propositions in Sections 3 through 4.2 also hold for Markovian *M/M/N+M* systems with integer values for N.

For the initial analysis in this Section we describe the contract between the client and vendor as a per-unit-time transfer payment T. Throughout this paper we use the subscript c for the client's profit, v for the vendor's profit, and s for the supply-chain profit. The profit per unit time for the client (π_c) and the vendor (π_v) are,

$$\pi_c(N) = R\,\lambda(1-F(N)) - P\frac{\lambda}{\theta}F(N) - T, \tag{1}$$

$$\pi_v(N) = T - cN\,. \tag{2}$$

The service supply chain profit function is,

$$\pi_s(N) = R\,\lambda(1-F(N)) - P\frac{\lambda}{\theta}F(N) - cN\,. \tag{3}$$

Let N^* be the supply chain profit maximizing capacity, $\pi_s^* = \pi_s(N^*) = \max_{N \geq 0} \pi_s(N)$. To avoid trivial cases, we assume parameter values such that there exists N^*, $0 < N^* < \infty$, and $\pi_s(N^*) \geq V$, where V is the vendor's reservation value.

The client wishes to design a contract that maximizes its profits. This can be achieved if the vendor's profit-maximizing capacity decision N_v^* is equal to N^* (the supply chain is *coordinated*) and if the supply chain profits can be arbitrarily allocated between the client and the vendor. Under such a contract the client can choose contract parameters so that the vendor earns its reservation value. In that case, the client earns a maximum of $\pi_c^* = \pi_s(N^*) - V$. Because M is the client's reservation value, the client offers the vendor such a profit-maximizing contract if $\pi_c^* \geq M$. We now focus on contract terms that we have observed in practice (see Tables 1 and 2): payment per call from client to vendor, payment for talk time, penalty per unit of the customer's waiting time and penalty for not meeting service level agreements.

3.2 Contracts with only PPC or PPT components

Given a PPC contract, the client pays the vendor r for each customer served by the vendor, $T = r\lambda(1-F(N))$. The vendor's expected daily profit under this contract is,

$$\pi_v(N) = r\lambda(1 - F(N)) - cN . \quad (4)$$

Expressions (3) and (4) lead to the following Proposition.

Proposition 1 A contract with only a PPC term can coordinate the chain so that $N_v^* = N^*$, but if the chain is coordinated the client earns negative profits.

Proof All proofs are included in the on-line appendix, Hasija et al., 2007.

In particular, we find that under a PPC-only contract, the client earns $-P\lambda/\theta < 0$ when the supply chain is coordinated. Therefore, a client will not offer a coordinating PPC-only contract. In addition, if the client offers a PPC-only contract with terms that maximize its profits, the chain is not coordinated

($N_v^* \neq N^*$) so that both client and chain profits are suboptimal. Similar results hold for contracts with only a PPT component.

These PPC-only and PPT-only contracts fail because they do not penalize the vendor when customers wait in the system; therefore the vendor has an incentive to underinvest in capacity. Therefore the client must provide an incentive to the vendor to invest in extra capacity. If the client's only lever to encourage this investment is to offer more pay per call (or more pay for time spent on the call), then the necessary level of pay is larger than the client's revenue. As we see in the following two sections, this problem is eliminated by adding a contract component that is based on waiting time.

3.3 Pay per call with penalty for not meeting the service level agreement (PPC+SLA)

Under this contract the client pays the vendor *r* for each call served and charges a penalty *p* if the vendor does not meet the service level agreement. We assume that the time period chosen by the client to observe the performance of the vendor is long enough so that the customer queue is essentially in equilibrium during the period. Not meeting the SLA means that the vendor makes a staffing decision *N* such that $G(N,t) < \alpha$ where (t, α) are specified in the contract. The vendor's expected daily profit is

$$\pi_v(N) = \begin{cases} r\lambda(1-F(N)) - cN & \text{if } G(N,t) \geq \alpha \\ r\lambda(1-F(N)) - p - cN & \text{if } G(N,t) < \alpha \end{cases} \tag{5}$$

Proposition 2 *The client can maximize its profit with a* PPC+SLA *contract by choosing the following contract parameters:*

(i) $p = r\lambda$ *so that the vendor maximizes profit when the SLA is met. Note that any large penalty (e.g., over* $r\lambda$*) is sufficient.*

(ii) (t, α) *such that the SLA constraint is tight at* N^* *i.e.,* $Pr\{wait < t\} = \alpha$ *at* N^*.

(iii) $$r = \frac{V + cN^*}{\lambda(1-F(N^*))}.$$

Under this contract it is possible to arbitrarily allocate the supply chain profits between the client and the vendor because the client can transfer more money to the vendor by raising the value of V when calculating r. Similar results can be obtained for a pay per time contract with an SLA constraint (PPT+SLA). Note, however, that an AHT constraint is not needed, for the appropriate choice of parameters in a PPT+SLA contract will coordinate the chain and maximize client profits. Finally, note that the client will only offer this contract if its maximum profit $\pi_c^* \geq M$.

3.4 Pay per call and penalty for waiting (PPC+W)

With this contract the client pays the vendor r for each call served and charges a penalty p for each unit of customer waiting time. The vendor's expected daily profit is

$$\pi_v(N) = r\lambda(1 - F(N)) - p\frac{\lambda}{\theta}F(N) - cN \tag{6}$$

Proposition 3 *The client can maximize profit with a* PPC+W contract *by choosing the following contract parameters:*

(i) $$r = \frac{V + R\lambda - \pi_s^*}{\lambda}.$$

(ii) $$p = \left(R + \frac{P}{\theta} - \frac{V + R\lambda - \pi_s^*}{\lambda}\right)\theta.$$

where π_s^* *is the supply chain optimal profit.*

The client will only offer this contract if $\pi_c^* \geq M$. Results similar to those of Proposition 3 can be obtained for a pay per time contract with a penalty for waiting (PPT+W), and again, an AHT constraint is not needed.

Finally, nearly identical results demonstrate that the client can coordinate the chain and maximize profits by using a constraint on the abandonment probability instead of waiting-time penalties or SLAs (PPC+Ab and PPT+Ab contracts).

4. Information Asymmetry with Two Productivity Types

In this Section we consider environments in which the vendor may have more information about agent productivity, μ, than the client. As we stated in the introduction to this paper, there are many reasons for this information asymmetry, e.g., the vendor's training and information technology investment decisions may not be visible to the client, and the vendor may have accumulated experience with other clients that allows it to produce superior productivity forecasts.

In this section we assume that the vendor's agents may be one of only two productivity types: a high type with μ_H and a low type with μ_L, where $\mu_H > \mu_L$. While this model is quite stylized, the results in this Section are necessary building blocks for assessing the performance of contracts when μ may have any positive value. We assume that the vendor knows the productivity type but the client does not. Given this information asymmetry, the client can maximize its profits if the contract leads to the following three conditions: (1) The vendor self-selects and reveals its productivity type, (2) Given that the vendor selects the appropriate contract, the vendor chooses the system-optimal capacity, and (3) given that the system is coordinated profits can be arbitrarily allocated between client and vendor. We say that a contract that satisfies these conditions *screens* the vendor. Guided by the revelation principle, we assume that the client offers pairs of contracts to the vendor, where each contract corresponds to a productivity type.

Table 3 defines abbreviations for all of the contracts that we will consider. In general, the letter 'C' indicates a pay per call contract, 'T' a pay per time contract, 'S' an SLA term, 'W' a penalty for waiting-time, 'A' an AHT term, 'H' the high-type vendor and 'L' the low-type vendor.

We first show that pairs of pay-per-call contracts cannot screen the vendor, even though each contract would be optimal in the full-information case. We then show that screening is possible by offering either a pair of PPT contracts with AHT constraints or by offering one contract based on PPC and another based on PPT.

Abbreviation	Contract
CS-H	PPC+SLA contract designed for a high productivity vendor (see Proposition 2).
CS-L	PPC+SLA contract designed for a low productivity vendor (see Proposition 2).
CW-H	PPC+W contract designed for a high productivity vendor (see Proposition 3).
CW-L	PPC+W contract designed for a low productivity vendor (see Proposition 3).
TS-L	PPT+SLA contract designed for a low productivity vendor.
TW-L	PPT+W contract designed for a low productivity vendor.
TSA-H	PPT+SLA+AHT contract designed for a high productivity vendor.

Table 3: Abbreviations to describe contracts

4.1 Independently optimal contracts that do not screen

To screen the vendors, it would seem reasonable to offer pairs of tailored PPC or PPT contracts that are optimal for the client under complete information. Let N_i^* be the service supply chain profit maximizing staffing level for $i = H, L$. Define $F_i(X)$ to be the equilibrium abandonment probability in a system staffed with X servers of type i, for $i = H, L$.

First we define the tailored PPC+SLA contracts,

CS-H: (i) Pay $r_H = \dfrac{V + cN_H^*}{\lambda(1 - F_H(N_H^*))}$ per call served.

(ii) SLA: $\Pr\{wait < t\} \geq \alpha_H$

(iii) Set t and α_H such that the SLA constraint is tight at N_H^* i.e., $G_H(N_H^*, t) = \alpha_H$.

CS-L: (i) Pay $r_L = \dfrac{V + cN_L^*}{\lambda(1 - F_L(N_L^*))}$ per call served.

(ii) SLA: $\Pr\{wait < t\} \geq \alpha_L$

(iii) Set α_L such that the SLA (with same *t* as in contract CS-H) constraint is tight at N_L^* i.e., $G_L(N_L^*, t) = \alpha_L$.

Proposition 4 *A contract that allows the vendor to choose between CS-H and CS-L will not screen the vendor productivity type.*

The proof of Proposition 4 shows that when these contracts are offered to the high-productivity vendor, it prefers the contract designed for the low-productivity vendor, resulting in lower profits for the client. Therefore, this pair of contracts is not incentive compatible. Similar results apply to PPT contracts with SLA constraints (but without AHT constraints) and to contracts with abandonment constraints: two PPT+SLA contracts, two PPC+Ab contracts, or two PPT+Ab contracts will not screen the vendor type.

Now consider pairs of tailored PPC+W contracts:

CW-H: (i) Pay $r_H = \dfrac{V + R\lambda - \pi_{H,s}^*}{\lambda}$ per call served

(ii) Charge $p_H = \left(R + \dfrac{P}{\theta} - \dfrac{V + R\lambda - \pi_{H,s}^*}{\lambda} \right)\theta$ per unit of customer waiting time.

CW-L: (i) Pay $r_L = \dfrac{V + R\lambda - \pi_{L,s}^*}{\lambda}$ per call served

(ii) Charge $p_L = \left(R + \dfrac{P}{\theta} - \dfrac{V + R\lambda - \pi_{L,s}^*}{\lambda} \right)\theta$ per unit of customer waiting time,

where $\pi_{H,s}^*$ and $\pi_{L,s}^*$ are respectively the optimal profits for the supply-chain when the vendor is high or low type.

Proposition 5 *A contract that allows the vendor to choose between CW-H and CW-L will not screen the vendor productivity type.*

The proof of Proposition 5 is similar to that of Proposition 4: again, the high-productivity vendor chooses the contract tailored for the low-productivity vendor, and the contract is not incentive compatible. Similar results apply to PPT+W contracts (without AHT constraints) as well.

We will see in Section 5 that a high-productivity vendor who chooses a PPC+W contract designed for a lower-productivity vendor *will* choose the system-optimal capacity (see Proposition 12). Therefore, a choice between CW-H and CW-L does coordinate the chain even though it does not screen. In other words, this pair of contracts produces optimal performance for the supply chain but the client does not maximize profits.

4.2 Screening contracts

Here we describe two pairs of incentive-compatible contracts that screen the vendors. The first is a choice between a PPC and a PPT contract while the second is a pair of PPT contracts in which an AHT constraint ensures that the high-productivity vendor chooses the system-optimal service rate. While these two pairs of contracts produce identical performance when there are just two productivity types, we will see in Section 5 that the PPC/PPT choice generates higher supply-chain profits when there is an arbitrary distribution of vendor productivity.

Note that if we ignore the stochastic elements of the system (e.g., by using a fluid model to describe the queues as in Ren and Zhou, 2006), it is easy to show that the two types of servers can be screened by offering PPC and PPT contracts. In a fluid model, incentive compatibility follows directly from average service times: slow vendors take longer to complete each call and therefore prefer to be compensated by time, while fast vendors create more rapid throughput and prefer to be paid accordingly. Thus, the vendors reveal their types when accepting the contracts.

This argument, however, is not sufficient to demonstrate incentive compatibility for our stochastic model. Here we must show that the contracts simultaneously motivate the vendors to reveal their types and create incentives to staff optimally with the appropriate capacity buffer (no such capacity buffer is needed in the fluid model which always staffs at λ/μ). This staffing problem complicates the contract de-

sign problem, for changes in the staffing level change the number of customers served (via the abandonment process), which then changes the payment rate to the vendor. In fact, the following properties of the queueing system and the contract terms will guarantee that successful screening is possible.

4.2.1 Properties of the queueing system that guarantee screening

Let N_H be the high type vendor's profit maximizing decision under contract TS-L, a PPT+SLA contract designed for a low productivity vendor. Define the following properties,

Property 1: $\frac{\mu_H}{\mu_L} N_H \geq N_L^*$,

Property 2: $F_H(N_H) \geq F_L(\frac{\mu_H}{\mu_L} N_H)$.

These two properties are mathematical statements of a commonly observed characteristic of queueing systems: the average waiting time for many slow servers is lower than the average waiting time for fast servers with equal total capacity. Specifically, Property 1 implies that to provide the same waiting time standard as a low type vendor, a high type vendor has to invest in a higher total capacity than the low type vendor. Property 2 implies that a low type vendor with the same total capacity as a high type vendor has a lower abandonment probability than the high type vendor. In Propositions 7 and 8, below, these two properties ensure that, for example, a high-type vendor earns less than its reservation value when accepting a pay per time contract designed to coordinate the supply chain with a low type.

Numerical experiments with a Markovian *M*/*M*/*N*+*M* model satisfy both properties, but we have not been able to prove that the Markovian model satisfies the conditions in general. In the following Proposition 6 we show that the two conditions hold under the diffusion approximation of Garnett et al. (2002). Lemma 1 describes a property of the hazard function $h(x)$ (defined in the appendix) that will be useful in the proof of Proposition 6.

Lemma 1 *$ah(x)$-$h(ax)$ is increasing in a for all x.*

Proposition 6 *Properties 1 and 2 are satisfied under the diffusion approximation.*

Now we are ready to describe the specific screening contracts. Without loss of generality, in cases where the vendor is indifferent between two contracts, we assume that the vendor chooses the contract that leads to the higher supply-chain profit. It is easy to relax this assumption by allowing the client to offer an infinitesimal fraction of the excess profit to the vendor.

4.2.2 Screening with a PPC/PPT contract

In this section we show that the client can successfully screen by offering the vendor a choice between a PPC+SLA and a PPT+SLA contract or a choice between a PPC+W and a PPT+W contract. First assume that the client offers the vendor a choice between CS-H and a PPT+SLA contract, TS-L. For brevity we exclude an analysis of abandonment penalties but it can be shown that an Ab component can substitute for SLA or W components in all the results of this section.

TS-L: (i) Pay $r_L = \dfrac{\mu_L(V + cN_L^*)}{\lambda(1 - F_L(N_L^*))}$ per unit time of vendor's service.

(ii) SLA: $\Pr\{wait < t\} \geq \alpha_L$.

(iii) Set α_L such that the SLA (with same t as in contract CS-H) constraint is tight at N_L^* i.e.,

$G_L(N_L^*, t) = \alpha_L$.

Lemma 2 $\dfrac{\mu_L(V + cN)}{\lambda(1 - F_L(N))} \geq r_L \ \forall\ N \geq N_L^*$.

Proposition 7 *A contract that allows the vendor to choose between CS-H and TS-L screens the vendor.*

Now we consider contracts with waiting time penalties. The following proposition states that the client can also use a PPC+W and a PPT+W contract to successfully screen the vendor and maximize profits. First define contract TW-L:

TW-L: (i) Pay $r_L = \dfrac{(V + R\lambda - \pi_{L,s}^*)\mu_L}{\lambda}$ per unit time spent on customer service.

(ii) Charge $p_L = \left(R + \frac{P}{\theta} - \frac{V + R\lambda - \pi^*_{L,s}}{\lambda} \right)\theta$ per unit of customer waiting time.

Proposition 8 *A contract that allows the vendor to choose between CW-H and TW-L screens the vendor.*

Note that in Proposition 7 and 8 no AHT constraint is needed for the PPT contract because the parameters of the contracts ensure that (i) the low-productivity vendor will choose the PPT contract and (ii) given that the low-productivity vendor chooses the PPT contract, its parameters ensure coordination of the chain and maximization of client profits. In Section 4.2.3 we will see how an AHT constraint can be used to screen two types of vendors with PPT contracts. In Section 5, when the client's prior distribution of vendor productivity contains more than two values, the AHT constraint will be used with these PPC/PPT contracts to maximize the client's benefits from screening.

4.2.3 Screening with AHT constraints

We saw in Proposition 4 that offering a pair of independently optimal PPC contracts does not screen, because the high-type vendor chooses the contract designed for the low-type. Once the high type accepts the low-type contract, it then earns more profit, and reduces the client's profit, by operating at a high rate. This problem may be corrected by again offering two PPC contracts, but with an AHT constraint on the low-type contract, specifically, an *upper-bound* AHT constraint that states that the vendor *cannot go faster* than the low rate μ_l.

Not surprisingly, we have never observed such a contract in practice. We also find that an AHT constraint as a more traditional lower bound, when applied to the high-type contract, can screen. Specifically, the client offers the vendor a choice between a PPT+SLA (or PPT+W) and a PPT+SLA+AHT (or PPT+W+AHT) contract. As was true for the SLA constraint, the client can ensure that the AHT constraint is always satisfied by the vendor by associating a high penalty for not meeting the AHT constraint. First we define the TSA-H contract:

TSA-H: (i) Pay $r_H = \dfrac{\mu_H(V + cN_H^*)}{\lambda(1 - F_H(N_H^*))}$ per unit time of vendor's service.

(ii) SLA: Pr{*wait* < *t*} ≥ α_H .

(iii) Set α_H such that the SLA (with same *t* as in contract TS-L) constraint is tight at N_H^* i.e.,

$G_H(N_H^*, t) = \alpha_H$.

(iv) AHT: Vendor pays a penalty P_{AHT} greater than $r_H \dfrac{\lambda}{\mu_H}$ if the vendor's service rate is less than μ_H .

Proposition 9: *A contract that allows the vendor to choose between TS-L and TSA-H screens the vendor.* Here the PPT+SLA contract is tailored for a low-productivity vendor and the PPT+SLA+AHT contract is tailored for a high-productivity vendor. A low-type vendor cannot satisfy the AHT constraint and therefore will accept the PPT+SLA contract. A high- type vendor will earn a profit rate that is less than its reservation value under the PPT+SLA contract (refer to Proposition 7) and therefore will accept the PPT+SLA+AHT contract under which the vendor will earn its reservation profit rate. Therefore the two PPT contracts screen the vendor type and maximize the chain and vendor profits.

Note that for all screening contracts (those in Sections 4.2.2 and this section), the client's profits must be compared to its reservation value *M*. Let $\pi_{L,c}^*, \pi_{H,c}^*$ be the maximum client profit rates when the vendor is the low and high productivity type, respectively. If $\pi_{L,c}^* < M \le \pi_{H,c}^*$, then the client will not be profitable under any contract with the low-productivity vendor and therefore the client will only offer the vendor a contract tailored for the high-type. If $\pi_{H,c}^* < M$, then the client will not be profitable with either vendor and therefore will not offer either contract. If $\pi_{L,c}^* \ge M$, then the client can maximize its profit by offering the screening contracts described above.

The screening contracts in sections 4.2.2 and this section lead to identical client and supply-chain profits when the client's prior distribution of the vendor's productivity contains two values. In the next section we assume that the client has an arbitrary prior distribution of the vendor's productivity, and we show that the two types of screening contracts can generate very different supply-chain profits.

5. Information Asymmetry with an Arbitrary Productivity Distribution

In this section we assume that the vendor's maximum potential service rate $\mu \in (0,\infty)$. The vendor knows its own potential service rate μ, and the client has a prior probability distribution of μ in the domain $(0,\infty)$. Our analysis in this section does not require any specific distributional assumption about the prior. We assume that the vendor chooses to operate at a service rate $\mu_v \in (0,\mu]$ such that the vendor's profit is maximized, given a contract. If slowing down its agents implies higher profits, then the vendor will choose to operate at a service rate $\mu_v < \mu$. However, we assume that the vendor cannot make its agents work at a service rate higher than the service rate μ. As in previous sections, the client's reservation value is M and the vendor's reservation value is V.

In this section we describe contracts that coordinate the service supply chain for certain productivity values μ, and we describe contracts that are non-coordinating for some values of μ. We show that a scheme similar to the screening contracts described above increases the client's profit above the profit achieved with a single PPC contract.

We now present some preliminary results:

Proposition 10: *The vendor's profit is increasing in μ_v under a PPC+SLA or PPC+W contract.*

Proposition 11: *The vendor's profit is decreasing in μ_v under a PPT+SLA or PPT+W contract.*

Proposition 10 states that PPC+SLA or PPC+W contracts create an incentive for the vendor to work as fast as possible. On the other hand, Proposition 11 states that PPT+SLA or PPT+W contracts create an

incentive for the vendor to decrease the service rate. Therefore the client must include a constraint on the service rate to ensure a minimum service rate (PPT+SLA+AHT or PPT+W+AHT), and in practice we find that pay-per-time contracts are accompanied by a constraint on the service rate. From Proposition 11 we infer that under such contracts the vendor chooses to operate at a service rate such that the constraint is tight.

5.1 A single PPC contract

We can also infer from Proposition 11 that a PPT+W contract will not coordinate the supply chain. We now show that the client can coordinate the chain over a certain range of μ by offering a PPC+W contract.

Proposition 12: *The PPC+W contract with* $r + \frac{p}{\theta} = R + \frac{P}{\theta}$ *coordinates the service supply chain if the vendor's productivity* μ *is high enough to satisfy the participation constraint.*

The expression $R + P/\theta$ represents the non-staffing costs generated by the queueing system (costs related to waiting and abandonment). Equating $r + p/\theta$ to this expression ensures that the vendor's profit function is identical to the supply chain profit function minus a constant. Therefore the condition $r + \frac{p}{\theta} = R + \frac{P}{\theta}$ is sufficient to coordinate the service supply chain. Note that contracts CW-H and CW-L described in Proposition 5 satisfy this condition, so that both low and high-productivity vendors will make system-optimal capacity decisions.

Therefore, a single PPC+W contract coordinates the chain, given any vendor with μ sufficiently high to meet its participation constraint. Let μ_l be the cut-off productivity where the maximum vendor profit $\pi_v^*(\mu_l) = V$. Vendors with $\mu \geq \mu_l$ will accept the contract and will invest in the system-optimal capacity. If $\mu < \mu_l$, the vendor does not accept the contract and hence the client makes M. The value of μ_l depends upon the contract parameters r and p, and the client's optimal choice of r, p and μ_l depends upon the prior distribution of μ, for it may be optimal for the client to design a contract that excludes a vendor with par-

ticularly low productivity. The results in this section do not depend on the particular values of r, p and μ_l. In Section 6 we will find these parameters numerically, given a distribution for μ.

Although the single PPC+W contract coordinates the chain over all participating vendors, it is not necessarily a good contract for the client. The client's expected profit under the single PPC+W contract is maximized only when the vendor's service rate $\mu = \mu_l$. When $\mu > \mu_l$ then the service supply chain profit is greater than when $\mu = \mu_l$. However, the extra supply chain profits are earned by the vendor as information rent (see Figure 1).

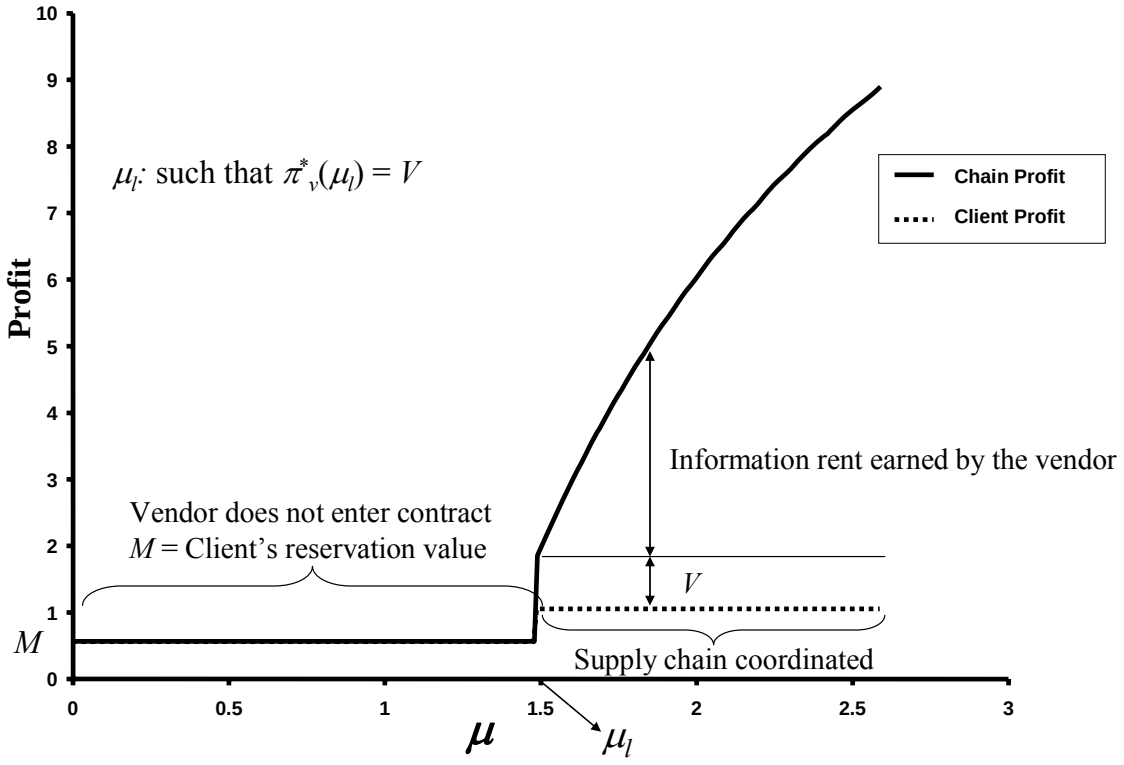

Figure 1: Client and supply chain profits under a single PPC+W contract

In the proof of Proposition 12 we see that for $\mu > \mu_l$, $\pi_v(\mu, N) = \pi_s(\mu, N) - (R - r)\lambda$. Therefore, under the PPC+W contract the client profit is independent of μ, and is equal to $(R\text{-}r)\,\lambda$.

5.2 A PPC/PPT screening contract increases the client's profits

In this section we present a screening contract that does not coordinate the service supply chain over all participating vendors, but does lead to higher profits for the client compared to the single PPC+W contract presented in Proposition 12.

Proposition 13: *The client's profit when it offers the vendor a choice between a PPT+W+AHT (or PPT+SLA+AHT) and a PPC+W contract is always greater than or equal to the profit when it offers the vendor a PPC+W contract.*

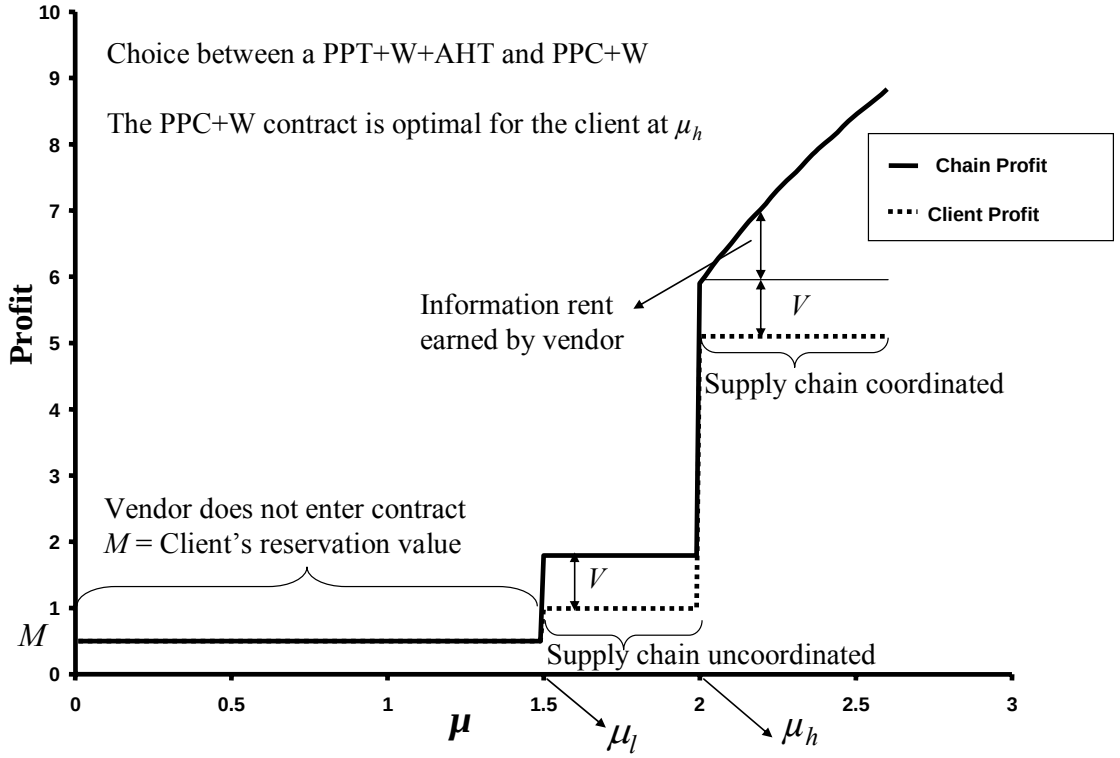

Figure 2: Client and supply chain profits under the PPT/PPC screening contract

Offering both PPT+W+AHT and PPC+W contracts induces a partition of the set of possible μ. The PPC+W contact is constructed so that for some $\mu_h > \mu_l$ if $\mu = \mu_h$ the vendor would set $\mu_v = \mu_h$ and the supply chain would be coordinated with the client maximizing profit. If $\mu > \mu_h$ the vendor can in-

crease μ_v and collect larger information rents, but these information rents are lower than they would be under the single PPC+W contract with the same lower threshold μ_l. For $\mu_l < \mu < \mu_h$ the supply chain will not be coordinated and the vendor will operate at $\mu_v = \mu_l$ but the client will be not be worse off than under the single PPC+W contract with the same μ_l. The AHT constraint places a lower bound on how much a vendor can slow down under the PPT contracts. This is illustrated in Figure 2 where for $\mu \le \mu_h$, the profit earned by the client is the same in both Figure 1 and 2. In Figure 2, however, the client earns a higher profit in the range $\mu \ge \mu_h$ than in Figure 1. Further in the range (μ_l, μ_h) the service supply chain is not coordinated in Figure 2. Therefore Figures 1 and 2 illustrate how a (noncoordinating) screening contract can be used by the client to increase its profit.

Note that in the proof of Proposition 13 we show that the client's profit is higher under a screening contract with a lower threshold μ_l than under any single PPC+W contract with the same lower threshold μ_l. Therefore the client's profit will also be higher with an optimal (from the client's perspective) screening contract than the optimal single PPC+W contract. In Section 6 we compare the optimal single PPC+W contract with the optimal PPT+W+AHT and PPC+W contracts for a given a priori distribution of μ.

Finally, in the proof of Proposition 13 and in the examples shown in Figures 1 and 2, the range of participating vendors is the same ($\mu > \mu_l = 1.5$). These ranges, however, depend upon the terms of the contracts. In Section 6 we will see that for certain prior productivity distributions the optimal ranges may differ, and this difference can produce better supply-chain results under the screening contract as well as higher client profits.

5.3 A PPT/PPT+AHT screening contract reduces the chain profits

In this section we show how a screening contract using two PPT contracts leads to lower chain profits than the screening contract presented in Proposition 13.

Proposition 14: *The client can use two PPT+W+AHT (or PPT+SLA+AHT) contracts to earn the same profit as the screening contract in Proposition 13, however such a contract leads to a lower supply chain profit.*

The two PPT contracts have a similar partitioning effect as the screening contract in Proposition 13. Vendors with $\mu < \mu_l$ will not participate. Vendors with $\mu \in [\mu_l, \mu_h)$ will accept the PPT+W+AHT contract tailored for a vendor with a service rate equal to μ_l and will choose to operate with $\mu_v = \mu_l$. Vendors with $\mu \geq \mu_h$ will accept the PPT+W+AHT contract tailored for a vendor with a service rate equal to μ_h and will choose to operate with $\mu_v = \mu_h$. Therefore we can see that the supply chain will not be coordinated for all μ except $\mu = \mu_l$ and $\mu = \mu_h$. Thus the chain profits are lower than under the screening contract in Proposition 13. This is shown in Figure 3. Note that although the chain profits are lower, the client profits are the same under the PPT based screening contract as under screening contract in Proposition 13.

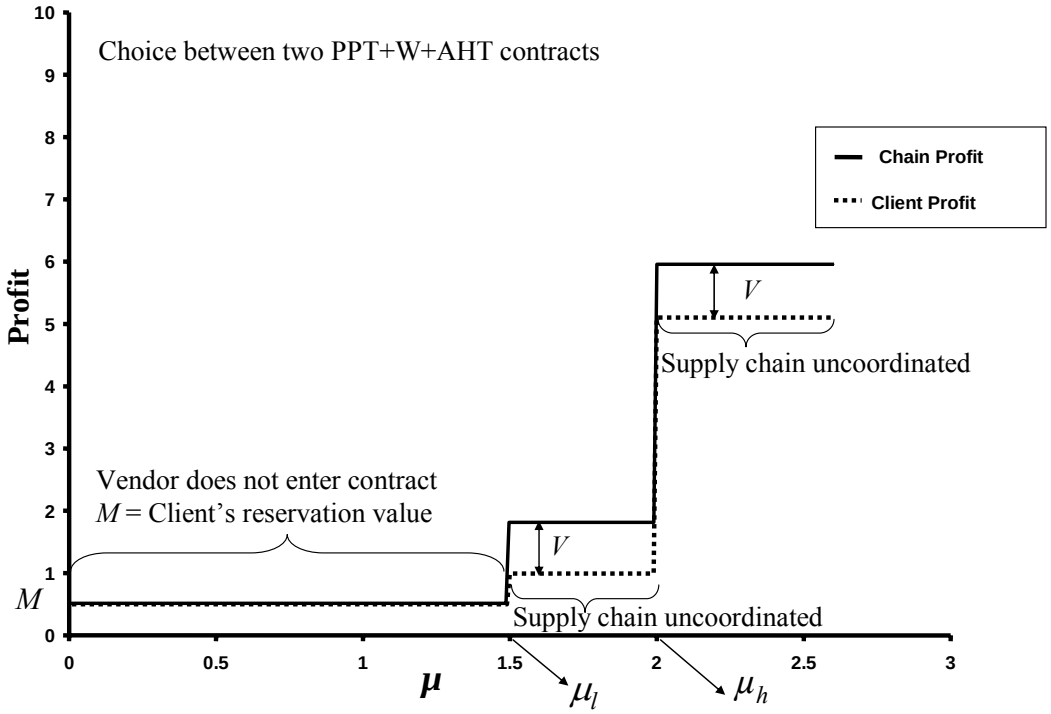

Figure 3: Client and supply chain profits under the PPT/PPT-AHT screening contract

6. Numerical Example

In this section we present a numerical example to illustrate the performance of the primary contracts described in Section 5 under various levels of information asymmetry, where the level of asymmetry is described by the standard deviation of the client's prior distribution of μ. Table 4 summarizes the parameter values used for all experiments.

Parameter	Value
R	0.4 per customer (revenue earned by the client for serving each customer)
P	0.1 per min. per customer (waiting cost incurred by the client)
c	0.5 per min. per employee (staffing cost incurred by the vendor)
V	0.8 per min. (vendor's reservation profit rate)
M	0.5 per min. (client's reservation profit rate)
λ	50 customers/min. (arrival rate)
θ	1/3 per min. (rate of customer's queue abandonment)

Table 4: Summary of parameter values used in numerical experiments

For these parameter values the service supply chain is feasible for $\mu > 1.45$, i.e., the maximum service supply chain profit when $\mu \le 1.45$ is less than $M+V$. In all of our examples we assume that $\mu \ge 1.50$. We also assume that the client's prior distribution on the vendor's productivity is a discrete uniform distribution. In our experiments we change the standard deviation of this distribution by changing its width while keeping the mean constant at 1.75. In particular, the tightest distribution is {1.70, 1.71,…, 1.80} for a standard deviation of approximately 0.03 while the widest is {1.50,1.51,…, 2.00} for a standard deviation of 0.15.

In the following experiments we compare the performance of the PPC+W contract described in Proposition 12 with the performance of the screening contracts described in Proposition 13 (PPC/PPT) and Proposition 14 (PPT/PPT+AHT). In the previous section these contracts were based on unspecified parameters μ_l and μ_h. Here we optimize the contracts over μ_l and μ_h such that the client's profits are maximized. We found the optimal parameters by using a grid search, although plots of these curves indicate

that the client's profit function is unimodal in these parameters so that in theory more efficient methods could be used. The resulting parameters are a single value of μ_l^* for the single PPC contract and values μ_l^* and μ_h^* for the screening contracts. Note that the two screening contracts use identical values of μ_l^* and μ_h^* because the client profits are identical for each realized value μ (see Figures 2 and 3).

We will call the contracts using these parameters the *optimal PPC contract*, the *optimal PPC/PPT screening contract* and the *optimal PPT/PPT+AHT screening contract*. Table 5 lists the contract parameters, and Figure 4 shows the expected chain and client profits under each contract.

		contract parameters		
		PPC	screening contracts	
prior distribution	standard deviation	μ_l^*	μ_l^*	μ_h^*
[1.70,1.80]	0.03	1.70	1.70	1.75
[1.67,1.83]	0.05	1.67	1.67	1.75
[1.65,1.85]	0.06	1.65	1.65	1.75
[1.62,1.88]	0.08	1.66	1.62	1.75
[1.60,1.90]	0.09	1.66	1.60	1.75
[1.57,1.93]	0.11	1.68	1.60	1.76
[1.55,1.95]	0.12	1.69	1.60	1.77
[1.52,1.98]	0.14	1.70	1.61	1.79
[1.50,2.00]	0.15	1.71	1.62	1.80

Table 5: Contract parameters in the numerical experiments

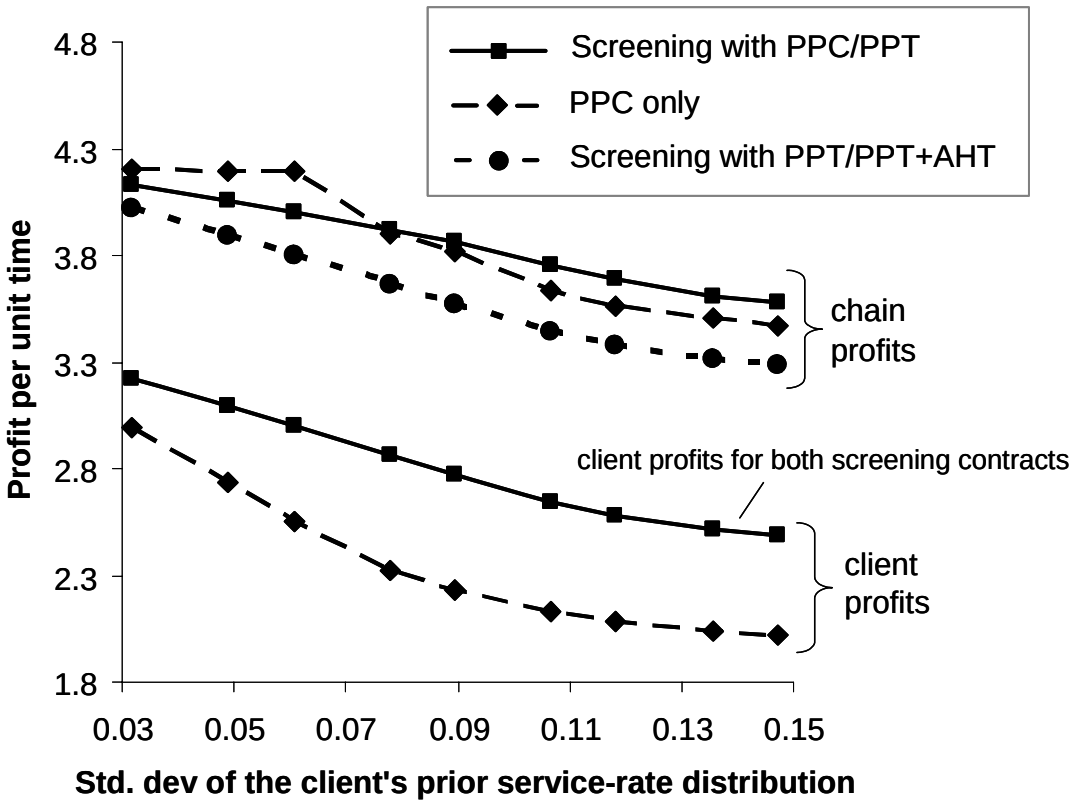

Figure 4: Chain and client profit vs measure of information asymmetry

In Figure 4, both chain and client profits fall as the information asymmetry increases (note that under both screening contracts the client profits are identical). We also see that the benefits of the screening contracts for the client can be significant. Given a standard deviation of 0.06, for example, the client captures 75% of supply chain profits under the optimal PPC/PPT screening contract but only captures 61% of profits under the optimal PPC contract.

In Figure 4 we also compare the total supply chain profits under each contract. First, screening with PPT/PPT+AHT consistently underperforms both the PPC/PPT screening contract and the PPC-only contract. This is because the supply chain is uncoordinated for high values of μ (see Figure 3). We also see that when the standard deviation is large, the supply chain profit for the optimal PPC/PPT screening contract can be higher than the supply chain profit for the optimal PPC contract. The screening contract produces these results even though it does not coordinate the chain for certain realized values of μ while the

PPC contract does coordinate the chain for any value μ, given that a vendor with productivity μ chooses to participate. The reason for the screening contract's advantage is that under higher standard deviations, the screening contract allows the client to profitably include a wider range of vendor types in the supply chain. Specifically, under the screening contract the client can design the PPT portion of the contract for vendors with low productivity. The optimal PPC contract does not include these vendors in the chain because a single PPC contract designed for a low productivity vendor produces higher information rents paid to vendors with higher productivity. Under the optimal screening contracts, these higher-productivity vendors will choose the PPC contract designed for them.

As an example consider the case represented in row 4 of Table 5. If the client offers the vendor an optimal PPC contract, the client's profit maximizing $\mu_l^* = 1.66$. This implies that a vendor with a service rate $\mu < 1.66$ will not enter the contract while a vendor with a service rate $\mu \geq 1.66$ will enter the contract and the service supply chain will be coordinated. However, if the client offers optimal PPC/PPT screening contract, the profit maximizing $\mu_l^* = 1.62$ and $\mu_h^* = 1.75$. This implies that a vendor with a service rate $\mu < 1.62$ will not enter the contract. A vendor with a service rate $1.62 \leq \mu < 1.75$ will choose the PPT portion of the contract and the service chain will not be coordinated in that range. A vendor with a service rate $\mu \geq 1.75$ will choose the PPC portion and the service supply chain will be coordinated in that range. Although the chain is not coordinated for all possible vendors, the chain earns higher profits, in expected value, because the client can profitably employ vendors in the range $1.62 \leq \mu < 1.66$. The same is true for the contracts shown in rows 5-9 of Table 5, and these rows correspond to the points in Figure 4 for which the screening contract produces higher chain profits than the PPC contract. In rows 1-3, however, the optimal PPC contract is designed for the lowest possible vendor productivity and therefore the PPC contract yields higher supply chain profits.

7. Conclusions and Future Research

In order for a firm to obtain all of the benefits of outsourcing it is important to choose contracts so that the vendor acts in the best interests of the client. In this paper we present a brief survey of contracts signed by one large call-center vendor, and we study the performance of these contracts. Specifically, we examine how different contract terms influence the vendor's capacity decision. Given that there is no information asymmetry between the client and the vendor, we identify single contracts that enable the client to maximize its profits by coordinating the service supply chain and allowing for arbitrary allocation of profit. These contracts have terms that force the client to internalize waiting and abandonment costs: constraints on the service level, a waiting-time penalty, or a penalty for each abandoning customer.

In business environments where there is information asymmetry about the maximum productivity of the vendor, it is also necessary to include contract terms related to waiting and abandonment costs, and all of the contracts we describe include such terms. Given information asymmetry about productivity, however, we find that no single contract can maximize the client's profit, for the client will pay information rent to the most productive vendors. When the client's prior belief about the vendor's productivity can be described by an extreme value distribution (high and low-productivity agents), then the client can screen the vendor productivity type, maximize supply chain profits, and maximize its own profit by offering either (i) a choice between two pay per time contracts, where the contract designed for the high-productivity vendor has an average handle time constraint, or (ii) a choice between pay per time and pay per call contracts. We then study the case when the client's prior belief about the vendor's productivity is given by a general distribution. Instead of using a single contract, the client can again increase its profit by using either of the screening contracts, although choice (ii), above, achieves higher supply-chain and vendor profits than choice (i).

We have focused on the relationship between a single vendor and a single client and assumed that each contract is accepted or refused and is not renegotiated as new information becomes available to the client. Future research includes analyzing the performance of contracts in a multi-period setting where produc-

tivity information revealed by the vendor during one period (say, by the choice of the service rate μ_v) enables the client to update its prior on the vendor's productivity and change the contract terms in subsequent periods. In addition, we may examine the impact of information asymmetry in the reliability of the vendor, perhaps as measured by the variability of the vendor's average service time. We may also examine vendor decisions that affect the quality of the customer experience, where quality includes, but is not limited to, waiting time. In general, vendors make many decisions besides staffing levels that affect the degree to which they are aligned with the client's objectives. Specifically, the vendor makes decisions about training and hiring processes, investments in information technology, and agent and manager incentives. Future research may consider models that incorporate these decisions.

ACKNOWLEDGEMENTS

The authors would like to thank Noah Gans, Serguei Netessine, Terry Taylor, two anonymous reviewers and an associate editor for their helpful comments.

REFERENCES

Allon G. and A. Federgruen. 2006. Outsourcing service processes to a common service provider under price and time competition, Working paper, Kellogg School of Management, Northwestern University, Evanston, IL.

Aron, R. and Y. Liu, 2003. Offshore Outsourcing of services: A model of the extended organizational form and survey findings, Working paper, Wharton School, University of Pennsylvania, Philadelphia, PA.

Askin, O. Z., F. de Véricourt, and F. Karaesmen. 2006. Call center outsourcing contract analysis and choice, Forthcoming in *Management Science* and Working paper, Koc University, Turkey.

Borst, S., A. Mandelbaum, and M. Reiman. 2004. Dimensioning large call centers, *Operations Research,* Vol. 52, No. 1, pp.17–34.

Cachon, G. 2003. *Supply chain coordination with contracts*, Handbooks in Operations Research and Management Science: Supply Chain Management.

Gans, N., G. Koole and A. Mandelbaum.2003. Telephone call centers: Tutorial, review, and research prospects, *Manufacturing & Service Operations Management*, Vol. 5, No. 2, pp. 79-141.

Gans, N. and Y.-P. Zhou. 2007. Call-routing schemes for call-center outsourcing, *Manufacturing & Service Operations Management*, Vol. 9, No. 1, pp. 33-50.

Garnett, O., A. Mandelbaum and M. Reiman. 2002. Designing a call center with impatient customers, *Manufacturing & Service Operations Management,* Vol. 4, No. 3, pp. 208-227.

Hasija, S., E. Pinker and R. Shumsky. 2005. Staffing and routing in a two-tier call center, *International Journal of Operational Research,* Vol. 1, No. 1, 8-29.

Hasija, S., E. Pinker and R. Shumsky. 2007. Call center outsourcing contracts under information uncertainty: on-line appendix, Working paper, Simon School of Business, Rochester, NY.

Lazear, E. P. 1995. *Personnel Economics*. The MIT Press, Cambridge, MA.

Mas-Colell, A., M. Whinston and J. Green. 1995. *Microeconomic Theory*. Oxford University Press, New York.

Ren, Z. J., and Y-P. Zhou. 2006. Call center outsourcing: Coordinating staffing levels and service quality, *Management Science*, forthcoming.

Shumsky, R. and E. Pinker. 2003. Gatekeepers and referrals in service, *Management Science,* Vol. 49, No. 7, pp. 839–856.

Appendix: Diffusion Approximation

The following approximation is due to Garnett et al. (2002).

$$N = \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} \tag{7}$$

$$F(N) = \left(1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)})}\right) w\left(-\beta, \sqrt{\frac{\mu}{\theta}}\right) \tag{8}$$

$$G(N,t) = 1 - \left(\frac{h(\beta\sqrt{\mu/\theta})}{\Psi(\beta\sqrt{\mu/\theta}, \sqrt{N\mu\theta}t)}\right) w\left(-\beta, \sqrt{\frac{\mu}{\theta}}\right) e^{-\theta t} \tag{9}$$

Where

$$h(x) = \frac{\phi(x)}{1-\Phi(x)}$$

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \Phi(x) = \int_{-\infty}^{x} \phi(y)dy$$

$$w(x,y) = \left[1 + \frac{h(-xy)}{yh(x)}\right]^{-1}$$

$$\Psi(x,y) = \frac{\phi(x)}{1-\Phi(x+y)}.$$