

Outsourcing a Two-Level Service Process

Hsiao-Hui Lee • Edieal Pinker • Robert A. Shumsky

School of Business, University of Hong Kong, Hong Kong

The Simon School of Business, University of Rochester, Rochester, New York 14627, USA

Tuck School of Business, Dartmouth College, Hanover, NH 03755, USA

hhlee@hku.hk • ed.pinker@simon.rochester.edu • shumsky@dartmouth.edu

November 29, 2011

This paper studies outsourcing decisions for a two-level service process in which the first level serves as a gatekeeper for a second level of experts. The objective of the system operator (the client) is to minimize the sum of staffing costs, customer waiting costs, and mistreatment costs due to unsuccessful attempts by a gatekeeper to solve the customer's problem. The client may outsource all or part of the process to a vendor, and first-best contracts exist when the client outsources only gatekeepers or experts. When the client outsources the entire system as a two-level process, a client-optimal contract may not exist unless the exogenous system parameters satisfy a particular (and unlikely) coordination condition. In addition, optimal incentive-compatible contracts exist when the vendor's structure choice (one-level or two-level) can deviate from the client's preference. Finally, we numerically examine how vendor structure choice and labor cost advantages influence the client's optimal outsourcing option.

1 Introduction

When managing a service process a firm must specify both the workflow and how much capacity to allocate to each part of the process. When a process is outsourced the firm cedes control over these decisions. In this paper we investigate the implications when workflow decisions are outsourced for a particular two-level customer service process.

When a customer enters our process, he is screened by a gatekeeper (level one) who may then attempt to solve his problem. If the gatekeeper does not attempt to solve the problem, or if the attempt fails, the problem is solved by an expert (level two) who is more expensive for the firm. The customer request could be a call to a technical support center or to a health care triage service such as NHS Direct, a helpline operated by the National Health Service of England and Wales that employs hundreds of gatekeepers and expert providers (Taylor, 2010). In both the technical support and health care environments, the customer must directly participate in the entire service. Our model, however, also applies to systems in which the customer does not participate directly in all parts of the service, such as the processing of credit applications.

In this paper we will use the language of the health care setting to describe some components of the system, but the analysis will only apply to large systems such as call centers. We refer to the initial assessment by the gatekeeper as a *diagnosis*, the resolution of the customer's problem as a *treatment*, and a gatekeeper's unsuccessful attempt at resolution as a *mistreatment*. When a customer is mistreated he experiences both additional delay as well as the direct disutility of being mistreated. This poor service experience may lead to a loss in a customer's expected lifetime value to the firm, and we call this loss the *mistreatment cost*. Our model incorporates three decisions faced by managers of such two-level systems: staffing quantities for both levels and referral rules between them. We assume that managers are minimizing the sum of staffing, mistreatment, and customer waiting costs.

When a firm is outsourcing its process, we call the firm the *client* and we call the external service provider the *vendor*. We focus on four outsourcing options. First, the client can outsource the process as a one-level system in which the gatekeeper is eliminated. If she decides to break the system into two components, we consider three other outsourcing options for the client: outsourcing only the expert (system S_e), outsourcing only the gatekeeper (system S_g), or outsourcing both to the same vendor (system S_b). One can observe all four of these options in practice, e.g., Infosys, one of the largest outsourcing companies in India, offers the technical support services of gatekeepers and/or experts.¹ The client may also keep the process in-house as a one or two-level system.

In our model, the vendor maximizes his profits when choosing staffing levels and referral rates. Initially, we will assume that the client specifies the vendor's process design (e.g., one-level or two-level). Then, we will incorporate vendor process choice into the model. In this paper, we consider the following questions: (1) How does the client write effective contracts for each outsourcing option? (2) Which outsourcing option should the client choose, and how do the costs, such as

¹<http://www.infosys.com/global-sourcing/case-studies>

the staffing, waiting, and mistreatment costs, affect her choice? (3) If the vendor is allowed to choose the process, how does vendor choice affect the client's outsourcing decision? (4) How does outsourcing affect the customer experience?

2 Literature Review

This paper contributes to a relatively new stream of operations management research on services outsourcing and contracting. To date, this literature has focused on processes with a single interaction between the customer and agent (a one-level system). This stream includes Ren and Zhou (2008), who study contracts that coordinate the vendor's staffing and service quality decisions. Ren and Zhang (2009) examine coordinating contracts under unknown and correlated capacity and quality costs, while Hasija et al. (2008) examine screening contracts that may be used by a client when the vendor's service rate is uncertain. Allon and Federgruen (2008) examine contracts for common outsourced services under price and time competition. Akşin et al. (2008) assume that arrival rates are uncertain and study contracts that allow the client to outsource the base customer demand or to outsource the peak demand. The model in our paper incorporates two, rather than one, level of agents.

As in our model, Lu et al. (2009) allow customers or products to visit multiple workers, for their system allows for rework. They examine how wage and piece-rate compensation packages affect workers' efforts to reduce the need for rework. They do not consider outsourcing sub-components of the system, as we do here.

The two-level system described here is related to Shumsky and Pinker (2003) and Hasija et al. (2005). Shumsky and Pinker (2003) describe a similar model of gatekeepers and experts, and they derive the optimal referral rate in a two-level system with deterministic service times and deterministic customer inter-arrival times. Modeling the management of gatekeepers as a principal-agent problem, Shumsky and Pinker find that incentives with both pay-per-service and pay-per-solve components can induce the gatekeeper to choose the system-optimal referral rate. Hasija et al. (2005) extend this deterministic model to a stochastic setting. The model in this paper generalizes the models in these previous papers in a variety of ways. The most significant difference, however, is that this paper focuses on the outsourcing and contracting issues while Shumsky and Pinker (2003) focus only on the referral decision and Hasija et al. (2005) only consider a centralized system.

There is also a stream of economics literature about the role of gatekeepers, particularly in health care. Marinoso and Jelovac (2003) and Malcomson (2004) describe optimal contracts for

gatekeepers, given that gatekeepers choose a level of diagnosis effort, and then may choose to treat or refer patients. Brekke et al. (2007) focuses on the gatekeeper’s role in allocating patients to the most appropriate secondary care provider (what we call experts). These papers do not address our topic: the financial and operational implications of aggregating or disaggregating and outsourcing the gatekeepers and experts.

3 Model

Our model for the two-level structure (Figure 1) follows Shumsky and Pinker (2003). The gatekeeper spends time diagnosing every customer’s request and then decides whether to attempt treatment. The complexity of a customer’s problem is represented by a real number $x \in [0, 1]$, which is defined as the fractile of complexity and therefore is uniformly distributed. Because requests are ranked by complexity a customer request with $x < x_0$ is more likely to be solved by the gatekeeper than every request with $x \geq x_0$. The gatekeepers’ skill levels are described by a treatment function $f(x)$, the probability that a request at the x fractile of treatment complexity for a gatekeeper can be successfully treated by that gatekeeper. The function $f(x)$ is strictly decreasing, continuous and differentiable. From the definition, $f(x)$ is between 0 and 1, and $f'(x) < 0$ implies that if the service is more complex, the probability that a gatekeeper can successfully treat the problem is smaller. To simplify the exposition we initially assume that the gatekeeper can accurately diagnose the complexity x for each request. We relax this assumption in Section 8.

The function $F(k) = \int_0^k f(x)dx$ is the expected fraction of requests that are successfully treated by a gatekeeper who chooses to treat all customer requests with complexity up to k . We call k the *treatment threshold*, and it is a decision variable in the system, along with the staffing levels. From the definition, the function $F(k)$ lies between 0 and k . For now we assume that any customer not successfully treated by a gatekeeper is then referred to, and successfully treated by, an expert. In Section 8.2 we will extend the model so that experts may also mistreat customers. We also assume that $F(1) < 1$, otherwise, we do not need experts at all. The workflow is shown in Figure 1. The figure shows that the mean total arrival rate to the experts is $\lambda_e(k) = (1 - F(k))\lambda$.

The gatekeeper’s (expert’s) mean treatment times $t_g(x)$ ($t_e(x)$) are functions of the problem complexity x . Treatment time should rise as complexity rises, so we assume that t_e and t_g are increasing in x , as well as continuous and differentiable. Given that the gatekeeper chooses treatment threshold k , $T_g(k) = \int_0^k t_g(x)dx$ is the expected gatekeeper treatment time averaged over all customers, including those not treated by gatekeepers. The quantity $T_g(k)/k$ is the expected

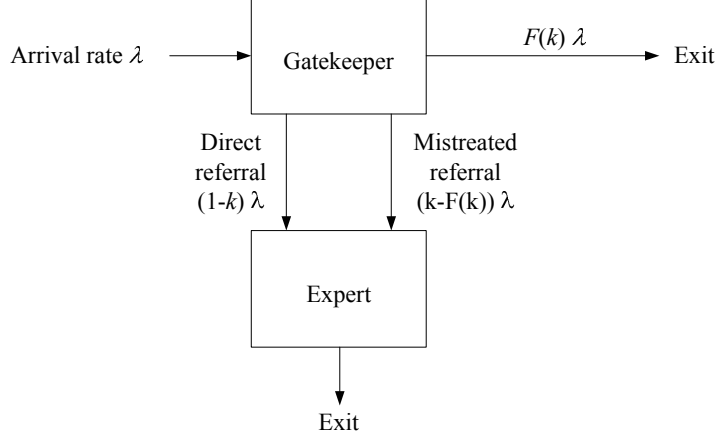


Figure 1: The two-level service process

treatment time of a customer, given that the customer is treated by a gatekeeper. Note that $\partial T_g(k)/\partial k = t_g(k) > 0$ and $\partial^2 T_g(k)/\partial k^2 = t'_g(k) > 0$. Let μ_d be the gatekeeper's service rate for a diagnosis, and therefore the service rate for a gatekeeper using treatment threshold k is $\mu_g(k) = [1/\mu_d + T_g(k)]^{-1}$.

If the gatekeeper chooses treatment threshold k , the expected expert treatment time averaged over all customers is $T_e(k) = \int_0^k t_e(x)(1-f(x))dx + \int_k^1 t_e(x)dx$. The quantity $T_e(k)/(1-F(k))$ is the expected treatment time, given treatment by an expert. The derivatives $\partial T_e(k)/\partial k = -t_e(k)f(k)$ and $\partial^2 T_e(k)/\partial k^2 = -V(k)$, where $V(k) = t'_e(k)f(k) + t_e(k)f'(k)$. The expert's service rate, given gatekeeper threshold k , is $\mu_e(k) = (1 - F(k))/T_e(k)$.

The client incurs costs for staffing, customer waiting and mistreatment. The gatekeeper and expert wage rates are C_g and C_e respectively, the customer waiting cost per unit time is C_w , and the cost per mistreatment by a gatekeeper is C_m . Note that wage rates C_g and C_e may differ between the client and the vendor. We will not add notation to distinguish between client and vendor wages, but the appropriate interpretation should be obvious from the context. For example, when outsourcing only gatekeepers, C_g refers to the vendor's gatekeeper wage rate while C_e refers to the client's expert wage rate.

We adopt the following notation when describing the client's costs and the vendor's profits: $\pi_c^j(\mathbf{y})$ = the client's cost when the service option j is selected, given that the client makes decision \mathbf{y} (a vector), and similarly $\pi_v^j(\mathbf{y})$ = the vendor's profits when the service option j is selected. Four outsourcing options are considered in our paper: $j \in \{e, g, b, 1\}$, in which e = system S_e with experts outsourced, g = system S_g with gatekeepers outsourced, b = system S_b with both outsourced, and

1 = the one-level structure, in which the customers are treated directly by an expert without being diagnosed by a gatekeeper. For example, $\pi_v^g(k, n_g)$ represents the vendor's profit function (subscript v) in system S_g (superscript g), when the vendor sets the treatment threshold to k and hires n_g gatekeepers. Finally, when the system is centralized with no outsourcing, the client's cost function has no superscript, e.g., $\pi_c(k, n_g, n_e)$.

4 Centralized System

We first analyze the two-level system when both subsystems are performed in-house by the client. The client's objective is to minimize expected cost by choosing the numbers of gatekeepers, n_g , and experts, n_e , as well as the treatment threshold for gatekeepers, k . The client pays for the gatekeepers, experts, and two indirect costs, the waiting and mistreatment costs. We use W_g and W_e to represent the expected waiting times in the gatekeeper and expert queues, respectively, and these are functions of the decision variables. The client's expected cost is,

$$\pi_c(k, n_g, n_e) = [C_g n_g + C_e n_e] + C_w [W_g \lambda + W_e \lambda_e(k)] + C_m (k - F(k)) \lambda. \quad (1)$$

From (1) we can see that for a given k , the mistreatment costs are completely determined, as are the workloads at each level of the system.

To derive expressions for the waiting time at each level, we make a few simplifying approximations. We first assume that the customer arrival process to the gatekeepers is Poisson. Second we approximate the service-time distributions at both the gatekeeper and expert levels as exponential random variables with rates $\mu_g(k)$ and $\mu_e(k)$, respectively. Therefore, we do not explicitly model separate diagnosis and treatment times. Our final approximation is that the arrival process to the expert queue is Poisson. The approximation is exact when, (i) diagnosis is instantaneous ($1/\mu_d = 0$) and, (ii) treatment times are exponentially distributed with mean values that do not depend on treatment complexity. The above approximations imply that for a fixed threshold k the two subsystems (gatekeeper and expert) can be analyzed independently as M/M/N queueing systems.

If we assume we are operating in one of the asymptotic regimes described in Borst et al. (2004), these approximations allows us to use square root rules to determine the optimal staffing. In the remainder of the paper, analytical results apply to any of the regimes described in Borst et al. For our tests of approximation accuracy and our numerical experiments, we will assume that we are in the QED regime of Halfin and Whitt (1981), who provide closed-form expressions for expected

waits. When diagnosis is not instantaneous or treatment times depend upon case complexity, clearly the assumptions of Borst et al (2004) and Halfin and Whit (1981) do not hold. In Appendix A we use simulation to show that the QED approximations are accurate for large systems, as measured by errors in the estimated total cost of the system. In the simulations we observe that the approximations are most accurate when the CV of the service time is close to one, as in the exponential distribution we are using as an approximation.

To apply the square root staffing rule we need both subsystems to operate in the asymptotic regime. For the gatekeeper subsystem, we assume that the arrival rate λ is large enough so that this is true. For the arrivals to the expert subsystem, for any k , $\lim_{\lambda \rightarrow \infty} \lambda_e(k)/\lambda = \lim_{\lambda \rightarrow \infty} \lambda(1 - F(k))/\lambda = \lim_{\lambda \rightarrow \infty} (1 - F(k)) \geq 1 - F(1) > 0$, where the last two steps follow from $F(k) \leq F(1) < 1$. Therefore, as $\lambda \rightarrow \infty$, $\lambda_e(k) \rightarrow \infty$ no slower than λ , for any k , and we can assume that the expert subsystem has a sufficiently large arrival rate as well.

To simplify notation, for the rest of the paper we suppress the dependence of $\mu_g(k)$, $t_g(k)$, $T_g(k)$, $\mu_e(k)$, $t_e(k)$, $T_e(k)$, and $\lambda_e(k)$ on k . It will also be useful to define $\rho_g = \lambda/\mu_g$ and $\rho_e = \lambda_e/\mu_e$, and again we will usually not express the dependence of these system loads on k .

4.1 Square Root Staffing Rule

Following Borst et al. (2004), consider a single-queue system with an arrival rate of λ , a service rate of μ , and N servers. The optimal staffing level is $N = \rho + \beta\sqrt{\rho}$, in which $\beta > 0$ can be seen as the standardized excess capacity to manage system variability. We can always find a β that meets any desired service requirement. In this paper, we quantify the service requirement by the expected waiting time, and the goal is to minimize the expected total cost

$$C_n N + C_w \lambda W, \tag{2}$$

where C_n is the unit cost per worker, C_w is the waiting cost per unit time, and W is the mean waiting time in the system. For an $M/M/N$ queue in the QED regime, the expected waiting time W is a function of β , λ , and μ (Borst et al., 2004):

$$W = \frac{\alpha(\beta)}{\sqrt{\lambda\mu}},$$

in which $\alpha(\beta) = [\beta + \beta^2\Phi(\beta)/\phi(\beta)]^{-1}$, while $\Phi(\beta)$ and $\phi(\beta)$ are the CDF and PDF, respectively, of a standard normal distribution.

The following lemma will be used to show that in a two-level system, the optimal staffing decisions and the optimal threshold are unique. All proofs are in Appendix B.

Lemma 1 *If $C_w > 0$, the optimal standardized excess capacity β^* satisfies $d(\alpha(\beta))/d\beta|_{\beta=\beta^*} = -C_n/C_w$ and is strictly increasing in C_w .*

4.2 The First-Best Solution

By applying the square root staffing rule, we staff the gatekeeper subsystem with $n_g^*(k, C_w) = \rho_g + \beta_g^*(C_w)\sqrt{\rho_g}$ and the expert subsystem with $n_e^*(k, C_w) = \rho_e + \beta_e^*(C_w)\sqrt{\rho_e}$, where β_g^* and β_e^* are the optimal standardized excess capacities for the gatekeeper and expert subsystems. Furthermore, from Lemma 1, we know that β_i^* satisfies

$$\frac{d}{d\beta}\alpha(\beta)\Big|_{\beta=\beta_i^*(C_w)} = -\frac{C_i}{C_w}, \quad (3)$$

for $i \in \{e, g\}$. The quantities β_g^* and β_e^* are parameterized by C_w because, later in the paper, the waiting costs for the gatekeeper or expert subsystem may depend on the contract terms.

Given expressions for $n_g^*(k, C_w)$ and $n_e^*(k, C_w)$, the client's cost function can be reduced to a function of k ,

$$\pi_c(k) = C_m\lambda(k - F(k)) + C_g\rho_g[1 + 2\Theta_g(k, C_w)] + C_e\rho_e[1 + 2\Theta_e(k, C_w)], \quad (4)$$

where $\Theta_i(k, y) = \eta_i(y)/(2\sqrt{\rho_i})$ and $\eta_i(y) = \beta_i^*(y) + (y/C_i)\alpha(\beta_i^*(y))$ for $i = \{e, g\}$. Note that as the arrival rate increases, the function $\Theta_i(k, y)$ approaches zero.

Lemma 2 guarantees the strict convexity of the cost function with respect to the treatment threshold and in addition the uniqueness of the optimal threshold and staffing because by Lemma 1 for a given k , $\beta_g^*(C_w)$ and $\beta_e^*(C_w)$ are unique.

Lemma 2 *$\pi_c(k)$ is a strictly convex function in k if (1) $\lambda > \tilde{\lambda}_c$ and (2) $\partial^2 T_e(k)/\partial k^2 > 0$ for all k , in which $\tilde{\lambda}_c = t_g^4 \mu_g^3 C_g^2 \eta_g^2(C_w)/16H_c^2$, and $H_c = -C_m f'(k^*) - C_e V(k^*) - C_e \eta_e(C_w) t_e^2 f^2(k^*)/4T_e^{3/2}$.*

The first condition places a lower bound on the arrival rate and is not related to the requirements of the QED regime. In our numerical examples, we find that this condition is easy to satisfy.

The second condition requires convexity of $T_e(k)$. Recall that $T_e(k)$ is the expert treatment time averaged over all customers, including customers treated by the gatekeeper who have an expert treatment time equal to 0. To examine the implications of this condition, assume that $t_e(x) = \bar{t}_e(1 + x\Delta_e)$, $\Delta_e > 0$, and $f(x) = 1 - bx$, $b > 0$. Then $\partial^2 T_e(k)/\partial k^2 > 0$ if $\Delta_e < b$, because

$$\frac{t'_e(k)}{t_e(k)} = \frac{\bar{t}_e \Delta_e}{\bar{t}_e(1 + k\Delta_e)} \leq \Delta_e < b \leq \frac{b}{1 - bk} = \frac{-f'(k)}{f(k)}.$$

This implies that the relative increment of the expert's treatment time (Δ_e) cannot be higher than the decrement of the gatekeeper's ability to successfully treat a customer (b). This would be consistent with situations in which lack of skill or knowledge cannot be completely compensated for by extra time spent with a customer.

Lemma 2 allows us to find the optimal threshold k^* by setting the first derivative of the client's cost to zero. The result is that $k^* = f^{-1}(r^*)$, in which

$$r^* = \frac{C_m + C_g t_g (1 + \Theta_g(k^*, C_w))}{C_m + C_e t_e (1 + \Theta_e(k^*, C_w))}. \quad (5)$$

Because the treatment function f is continuous, by the fixed-point theorem, the solution k^* exists. From the properties of $f(k)$, $r^* \in [0, 1]$.

Denote the optimal threshold in Shumsky and Pinker (2003) as $k_d^* = f^{-1}(r_d^*)$, in which $r_d^* = (C_m + C_g t_g) / (C_m + C_e t_e)$ with the subscript d indicating their paper's deterministic assumption for the arrival and service time distributions. The expression for our optimal threshold k^* has a form similar to k_d^* with the factor $C_i t_i$ replaced by $C_i t_i (1 + \Theta_i(k^*, C_w))$ for $i = \{e, g\}$. Because both $\Theta_g(k^*, C_w)$ and $\Theta_e(k^*, C_w)$ approach 0 as λ rises, $\lim_{\lambda \rightarrow \infty} k^* = k_d^*$; for large stochastic systems, the optimal treatment threshold is close to the optimal threshold for a deterministic system. Hasiija et al. (2005) find numerically that k_d^* can be an accurate approximation for k^* for systems of moderate to large size (e.g., 20 servers or more). We observed similar results when running the numerical experiments presented later in this paper.

5 Analysis of Outsourcing Contracts

We consider four outsourcing options: outsourcing the expert (S_e), outsourcing the gatekeeper (S_g), outsourcing both (S_b), and outsourcing a one-level system with no gatekeepers involved. The vendor can accept the contract or reject it; we assume here that the vendor's reservation level is 0. The client would like to design an enforceable contract that minimizes her costs, while providing a non-negative profit to the vendor. We are also interested in whether a particular contract achieves the *system optimum* or *first best*, the lowest cost achieved when decisions are centralized and when using the lowest wage rates of both client and vendor. When a contract achieves first best and the client incurs its minimum possible costs, we say that the contract *coordinates* the system.

We assume that the client has perfect knowledge of all static information such as model parameters. In some environments this is a limitation of our model, but clients often obtain estimates of the vendor's system parameters from third-party outsourcing consultants (e.g., Mackie, 2007)

or from business information firms (e.g., www.datamonitor.com) that examine the performance of peer-groups of competing vendors. The client may also obtain static information by closely monitoring the vendor during previous contracts or during a trial period before the final contract is signed. Certainly, obtaining accurate information is not always possible. As Mackie (2007) writes, a benchmarking study to obtain parameter estimates is possible for "any service that has a low level of variability, a maturity of specification and a strong market for competitive supply." Our model applies to such services, while in other environments there may be significant information asymmetry between the client and vendor (Hasija et al., 2008; Ren and Zhang, 2009).

When the client outsources a component of the system, or the entire system, we assume that the client only observes customer workflows into and out of the vendor, but does not observe the workflows or decisions inside the vendor's facility such as staffing levels, internal queue lengths, or internal referral rates. In other words, the contract terms are based only on data observable through standard, external technology such as a telecommunications switch. Our contracts are based upon the number of customers handled by the vendor (corresponding to a pay-per-service incentive), the customer's total time in system with the vendor (corresponding to a system-time penalty), and whether a customer has been satisfactorily served by the vendor (corresponding to a pay-per-solve incentive). The client, however, continues to bear the costs of customer delays and mistreatment. Finally, in this section we assume that the choice of the overall system structure (a one-level or two-level system) is observable to the client and is contractible. In Section 6 we allow the vendor to choose the structure.

5.1 Systems S_e and S_g and the one-level system

In system S_e , the client outsources the expert subsystem to the vendor, who receives a payment from the client for the treatments performed. Meanwhile, the client operates the gatekeeper subsystem and sets the gatekeeper staffing level, the treatment threshold, and the contract. Similarly to Hasija et al. (2008) it can be shown that a pay-per-service contract (P^e) with a system-time penalty (Q^e) will achieve the centralized solution. The specific contract terms are given in the following proposition.

Proposition 1 *A contract with system-time-penalty + pay-per-service components coordinates the system if the client offers the contract $(Q^e, P^e) = (C_w, C_e \rho_e [1 + 2\Theta_e(k^*, C_w)] / \lambda_e + C_w T_e(k^*) / (1 - F(k^*)))$.*

In system S_g , the client outsources the gatekeeper subsystem to the vendor, who receives a payment from the client and sets the number of gatekeepers and the treatment threshold. The client determines the staffing level for the expert subsystem, given the flow of referrals from the vendor. We consider a contract in which the client pays the vendor for each customer handled (pay-per-service, P^g) and a reward for the treatments that are successfully performed by the gatekeepers (pay-per-solve, R^g), and also applies a system-time penalty (Q^g). We assume that the vendor cannot deny that a customer was mistreated, block access to the client's experts, and collect the reward for treatment.

Lemma 3 $\pi_v^g(k)$ is strictly concave if $\lambda > \tilde{\lambda}_g$, in which $\tilde{\lambda}_g = t_g^4 \mu_g^3 C_g^2 \eta_g^2(Q^g)/16H_g^2$ and $H_g = -f'(k)R^g + Q^g t'_g + C_g t'_g$.

As noted before, the condition $\lambda > \tilde{\lambda}_g$ is very mild.

Proposition 2 A contract with system-time-penalty + pay-per-service + pay-per-solve components coordinates the system if the client offers $(Q^g, P^g, R^g) = (C_w, C_g \rho_g [1 + 2\Theta_g(k^*, C_w)] / \lambda - F(k^*)R^g + C_w(1/\mu_d + T_g), t_g[C_w + C_g(1 + \Theta_g(k^*, C_w))]/r^*)$.

Q^g ensures that the vendor staffs optimally when $k^g = k^*$, R^g ensures that $k^g = k^*$, and finally P^g ensures that the client extracts all the vendor's profit while the vendor is still willing to accept the contract.

Finally, under certain sets of parameters, a one-level system with experts only is optimal for example if the wage rate difference between experts and gatekeepers is not large. If the client outsources a one-level system, the contract is similar to the one in system S_e , i.e., a pay-per-service contract (P^1) with system-time penalty (Q^1). Corollary 1 describes the contract, which follows directly from Proposition 1.

Corollary 1 A contract with system-time-penalty + pay-per-service components coordinates the system if the client offers the contract $(Q^1, P^1) = (C_w, C_e \rho_e (1 + 2\Theta_e(0, C_w)) / \lambda + C_w T_e(0))$.

5.2 System S_b

The most interesting and challenging outsourcing problem is posed by S_b , the system in which both gatekeepers and experts are outsourced. In this case the vendor's choice of treatment threshold (or referral rate), gatekeeper staffing level, and expert staffing level are not directly observed by the client. In addition, customer mistreatments cannot be directly observed, for the client sees only a

stream of successfully treated customers leaving the vendor's system. Under these conditions, we find that there can be a significant coordination cost to the client.

In system S_b , the only two measures observed by the client are the rate of customers served and the time in system. Therefore, we consider here a pay-per-service contract with a linear system-time penalty. Denote P^b as the payment per complete treatment and Q^b as the system-time penalty (\$/time/customer), which penalizes the total time in service, with mean $t(k) = 1/\mu_d + T_e + T_g$, and the time in queue, with mean $W_g + (1 - F(k))W_e$. The transfer payment from the client to the vendor is proportional to the number of services completed less the system-time penalty, i.e., $\lambda [P^b - Q^b[t(k) + W_g + (1 - F(k))W_e]]$. Consequently, the vendor's profit is,

$$\pi_v^b(n_g, n_e, k) = P^b\lambda - Q^b\lambda t(k) - (C_g n_g + Q^b W_g \lambda) - (C_e n_e + Q^b W_e \lambda_e). \quad (6)$$

Let $\pi_v^b(k)$ be the vendor's profit after applying the square-root staffing rule.

Lemma 4 $\pi_v^b(k)$ is a strictly concave function in k if (1) $\lambda > \tilde{\lambda}_b$ and (2) $\partial^2 T_e(k)/\partial k^2 > 0$ for all k , in which $\tilde{\lambda}_b = t_g^4 \mu_g^3 C_g^2 \eta_g^2(Q^b)/16H_b^2$ and $H_b = -(Q^b + C_e)V(Q^b) + Q^b t'_g - C_e \eta_e(Q^b) t_e^2 f^2(k)/4T_e^{3/2}$.

As in Lemma 2, the conditions specify a minimum arrival rate as well as characteristics of the treatment functions. Again, we have found numerically that neither condition is restrictive.

By optimizing $\pi_v^b(k)$ over k , we find that when the vendor is offered contract (Q^b, P^b) , the vendor chooses threshold $k^b = f^{-1}(r^b)$,

$$r^b = \frac{Q^b t_g + C_g t_g (1 + \Theta_g(k^b, Q^b))}{Q^b t_e + C_e t_e (1 + \Theta_e(k^b, Q^b))}. \quad (7)$$

It is useful to compare this threshold condition with Equation (5), the condition that defines the first-best treatment threshold. When moving from r^* to r^b , the mistreatment cost C_m is replaced by $Q^b t_g$ and $Q^b t_e$, and the waiting cost C_w is replaced by Q^b . This implies that in system S_b , the system-time penalty Q^b must serve two roles for the client. Because the system-time penalty penalizes waiting times in both queues, Q^b serves as a congestion cost to the vendor. On the other hand, because mistreated requests are treated twice and therefore increase time in service, Q^b also serves as a mistreatment penalty. In this case, Q^b affects k^b directly. Lemma 5 describes the relationship between k^b and Q^b : if the conditions of the lemma are satisfied, the direct effect of the mistreatment penalty dominates and the threshold k^b decreases with Q^b .

Lemma 5 If (1) $\lambda > \tilde{\lambda}_b$ and (2) $\partial^2 T_e(k)/\partial k^2 > 0$ for all k , and (3) $t_g \geq \max(1, \tilde{\alpha})t_e$ for $0 \leq k \leq 1$, then k^b is strictly decreasing with Q^b , in which $\tilde{\alpha} = (f(k^*)/\sqrt{\mu_g T_e})(\alpha(\beta_e^*(C_w))/\alpha(\beta_g^*(C_w)))$.

Conditions 1 and 2 guarantee the concavity of the vendor's profit function; Condition 3 requires a system in which the gatekeeper's treatment time is longer than the expert's treatment time.

Proposition 3 *A contract with system-time-penalty + pay-per-service components, (Q^b, P^b) , coordinates the system if and only if*

$$Q^b = C_w = \frac{f(k^*)C_e t_e (1 + \Theta_e(k^*, C_w)) - C_g t_g (1 + \Theta_g(k^*, C_w))}{t_g - f(k^*)t_e}.$$

We refer to the condition on Q^b in Proposition 3, as the *coordination condition*. If it is satisfied then system S_b can achieve the system optimum. Corollary 2 describes how C_m and C_w relate to each other within the coordination condition. The corollary provides intuition as to what causes inefficiency in system S_b ; we will reinforce this intuition with numerical experiments in the next section.

Corollary 2 *The mistreatment cost C_m that satisfies the coordination condition increases with the waiting cost C_w if (1) $\lambda > \tilde{\lambda}_c$, (2) $t_g \geq \max(1, \tilde{\alpha})t_e$ for $0 \leq k \leq 1$, and (3) $t'_g \leq (t_g - t_e)(-f'(k^*)) / (1 - f(k^*)) + t'_e f(k^*)$.*

The corollary implies that to achieve coordination, mistreatment and waiting costs must rise together. The relationship between C_m and C_w follows only if all three conditions are satisfied. Specifically, Condition 1 again requires a large enough arrival rate; Conditions 2 and 3 require the gatekeepers to be slower than the experts. For example, if $t_g(x) = \bar{t}_g(1 + x\Delta_g)$, $t_e(x) = \bar{t}_e(1 + x\Delta_e)$, and $f(x) = 1 - bx$, Condition 2 ($t_g \geq \max(1, \tilde{\alpha})t_e$) is not sufficient and we need Condition 3, which reduces to $\bar{t}_g \geq \bar{t}_e(1 + b\Delta_e)$.

Next we discuss the contract the client offers when the coordination condition is not satisfied. The client pays the vendor for the services that he provides and incurs the waiting and mistreatment costs, yielding the following objective:

$$\pi_c^b(Q^b, P^b) = C_m \lambda (k^b - F(k^b)) + P^b \lambda - Q^b \lambda t(k^b) + (C_w - Q^b) \left(\alpha(\beta_g^*) \sqrt{\rho_g} + \alpha(\beta_e^*) \sqrt{\rho_e} \right),$$

in which $\beta_i^*(Q^b)$ is abbreviated to β_i^* for $i \in \{e, g\}$. From Equation (7), we know that the threshold k^b only depends on the value of Q^b but not P^b . Consequently, for a given Q^b , the client's cost increases with P^b , and thus she prefers to lower the value of P^b . However, P^b has to be sufficiently high so that the vendor will accept the contract. As a result, she offers a P^b that gives the vendor zero profit,

$$P^b = Q^b t(k^b) + \frac{1}{\lambda} \left\{ C_e \rho_e \left[1 + 2\Theta_e(k^b, Q^b) \right] + C_g \rho_g \left[1 + 2\Theta_g(k^b, Q^b) \right] \right\}.$$

The client's cost is therefore only a function of Q^b ,

$$\pi_c^b(Q^b) = C_m \lambda (k^b - F(k^b)) + \left(C_g n_g + C_w \alpha(\beta_g^*(Q^b)) \sqrt{\rho_g} \right) + \left(C_e n_e + C_w \alpha(\beta_e^*(Q^b)) \sqrt{\rho_e} \right).$$

Note that the congestion cost is C_w for the client, but is Q^b for the vendor.

Proposition 4 describes the complete client-optimal contract (Q^b, P^b) , given $\lambda > \tilde{\lambda}_*^b$. As for the other arrival rate lower bounds, this condition is not restrictive.

Proposition 4 *Assume that $\lambda > \tilde{\lambda}_*^b$. The client offers a contract $(Q^b, P^b) = (Q_*^b, P_*^b)$, in which*

$$Q_*^b = \left[C_e t_e \left[1 + \Theta_e(k_*^b, Q_*^b) \right] r_*^b - C_g t_g \left[1 + \Theta_g(k_*^b, Q_*^b) \right] \right] / [t_g - r_*^b t_e], \text{ and}$$

$$P_*^b = Q_*^b t(k_*^b) + \left[C_e \rho_e \left[1 + 2\Theta_e(k_*^b, Q_*^b) \right] + C_g \rho_g \left[1 + 2\Theta_g(k_*^b, Q_*^b) \right] \right] / \lambda,$$

and the vendor sets the threshold at $k^b = k_*^b = f^{-1}(r_*^b)$, in which $r_*^b = (C_m + C_g t_g (1 + \Psi_g(k_*^b))) / C_m + C_e t_e (1 + \Psi_e(k_*^b))$, $\Psi_i(k) = \tilde{\eta}_i(Q^b) / [2\sqrt{\rho_i}]$ and $\tilde{\eta}_i(Q^b) = \beta_i^*(Q^b) + C_w \alpha(\beta_i^*(Q^b)) / C_i$, for $i \in \{e, g\}$, $\tilde{\lambda}_*^b = t_g^4 \mu_g^3 C_g^2 \tilde{\eta}_g^2(Q_*^b) / 16 (H_*^b)^2$, and $H_*^b = -C_m f'(k_*^b) - C_e V(Q_*^b) - C_e \tilde{\eta}_e^2(Q_*^b) t_e^2 f^2(k_*^b) / 4T_e^{3/2}$.

Because $k_c^b \neq k^*$, the client's costs are higher than the system optimum, $\pi_c(k^*)$. If the vendor does not provide cost advantage(s) to the client, the client will not outsource both levels.

The following two propositions describe the relationships among the costs C_w and C_m and the optimal solutions k_*^b and Q_*^b . First, when we increase the waiting cost, it is intuitive that the client would raise the system-time penalty to enforce higher staffing levels. Therefore, when the client increases the system-time penalty, the optimal threshold k_*^b decreases as well, because to the vendor, the mistreatment cost increases.

Proposition 5 *Given the conditions of Lemma 2 and Proposition 4, Q_*^b increases with C_w and k_*^b decreases with C_w .*

Proposition 5 implies that as C_w increases, the total mistreatment cost in the system declines because a lower k_*^b implies fewer gatekeeper treatment attempts, and therefore lower mistreatment costs. Proposition 6 shows how the mistreatment cost parameter affects the optimal system-time penalty and the optimal threshold.

Proposition 6 *Given the conditions of Lemma 2 and Proposition 4, Q_*^b increases with C_m and k_*^b decreases with C_m .*

Proposition 6 implies that as C_m increases, the mean customer waiting time in the system declines because a higher Q_*^b will lead to greater staffing by the vendor and shorter queueing times. In addition, a lower k_*^b implies fewer gatekeeper treatment attempts, and therefore less time spent with the gatekeeper.

6 Vendor Process Choice

Thus far we have assumed that the vendor complies with the client's process choice (expert only, two-level, one-level, etc.). There may be cases in which it is difficult for the client to monitor the vendor's process choice and therefore it is possible for the vendor to deviate from the client's preference. In this section we will account for the possibility of such behavior by the vendor.

Assume that the client wants to outsource the entire system. If the client offers a contract designed for a one-level (two-level) system, and under that contract the vendor chooses to operate a one-level (two-level) system, then we say that the contract is *incentive compatible*. Here we describe the client's optimal incentive-compatible contract. Let (Q, P) be the system-time penalty and pay-per-service payment the client offers. The vendor's profit function if the vendor chooses a one-level process is:

$$\pi_v^1(Q, P) = P\lambda - QT_e(0)\lambda - C_e\rho_e(0)(1 + 2\Theta_e(0, Q)).$$

If he chooses a two-level process the vendor's profit is:

$$\pi_v^b(Q, P) = P\lambda - Qt(k(Q))\lambda - C_g\rho_g(k(Q))(1 + 2\Theta_g(k(Q), Q)) - C_e\rho_e(k(Q))(1 + 2\Theta_e(k(Q), Q)),$$

in which

$$f(k(Q)) = \frac{t_g(k(Q))Q + C_g(1 + \Theta_g(k(Q), Q))}{t_e(k(Q))Q + C_e(1 + \Theta_e(k(Q), Q))}.$$

This equation implicitly defines $k(Q)$ and only depends on Q but not P .

We first construct the vendor's individual rationality (IR) constraints for both process choices. A necessary condition for the vendor to accept the contract and choose a one-level process is that the pay-per-service payment satisfies $\pi_v^1(Q, P) \geq 0$, or equivalently,

$$P \geq P^1(Q) = QT_e(0) + C_e\rho_e(0)(1 + 2\Theta_e(0, Q)) / \lambda > 0. \quad (8)$$

A necessary condition for the vendor to accept the contract and choose a two-level process, is that the pay-per-service payment satisfies the IR constraint $\pi_v^b(Q, P) \geq 0$, or equivalently,

$$P \geq P^b(Q) = Qt(k(Q)) + \frac{C_g\rho_g(k(Q))}{\lambda}(1 + 2\Theta_g(k(Q), Q)) + \frac{C_e\rho_e(k(Q))}{\lambda}(1 + 2\Theta_e(k(Q), Q)) > 0. \quad (9)$$

The following lemma describes the space of possible contracts that the client may offer.

Lemma 6 *For a given Q , the client only offers contract $(Q, \min(P^1(Q), P^b(Q)))$ to the vendor.*

In Proposition 7 below we will describe the client's optimal incentive-compatible contract in terms of the points of intersection between the two IR constraints. But first, consider the case where the two IR constraints do not intersect. If $P^1(Q) < P^b(Q)$ for all Q , the vendor will always choose a one-level system and therefore an incentive compatible contract does not exist for a two-level system. If $P^1(Q) > P^b(Q)$ for all Q , the vendor will always choose a two-level system and an incentive compatible contract does not exist for a one-level system.

Now assume that the IR constraints do cross, and let N be the number of intersections. At each intersection, $P^1(Q) = P^b(Q)$, so that the vendor is indifferent between a one- and a two-level system, and we assume that the vendor chooses the process preferred by the client. Denote the intersections as $(Q_1, P_1), (Q_2, P_2), \dots, (Q_N, P_N)$, in which $Q_1 < Q_2 < \dots < Q_N < \infty$. Recall that the optimal contract for system S_b in Proposition 4 is (Q_*^b, P_*^b) , and for the one-level system in Corollary 1 is (Q^1, P^1) . Thus, we can find the two nearest intersections to Q_*^b and Q^1 , and we label these intersections with $L, L + 1, J$, and $J + 1$. That is, $Q_L < Q_*^b < Q_{L+1}$ and $Q_J < Q^1 < Q_{J+1}$.

Proposition 7 (A) *The client's optimal and incentive compatible contract for system S_b is either (Q_*^b, P_*^b) , if $P_*^b \leq P^1(Q_*^b)$, or is (Q_L, P_L) or (Q_{L+1}, P_{L+1}) , if $P_*^b > P^1(Q_*^b)$. (B) *The client's optimal and incentive compatible contract for the one-level system is either (Q^1, P^1) , if $P^1 \leq P^b(Q^1)$, or is (Q_J, P_J) or (Q_{J+1}, P_{J+1}) if $P^1 > P^b(Q^1)$.**

This proposition implies that for system S_b , if the client anticipates that the vendor will not deviate from the client's choice, then the client will choose the contract in Proposition 4; if the vendor is expected to deviate from his choice, then it is optimal for the client to choose an incentive compatible contract that generates the lowest cost. This incentive compatible contract will be one of the two intersections that are nearest to (Q_*^b, P_*^b) . A similar logic applies when outsourcing a one-level system: either offer the contract in Corollary 1 or a contract located on one of the two nearest intersections of the IR constraints.

7 Numerical Examples of Outsourcing Decisions

We now use numerical experiments to explore the costs and benefits of each outsourcing contract. We first examine how outsourcing both components can lead to system inefficiency, higher costs for the client, and changes to the customer experience in terms of waiting and mistreatment costs. We first assume that the vendor does not deviate from the client's preferred system structure. Then

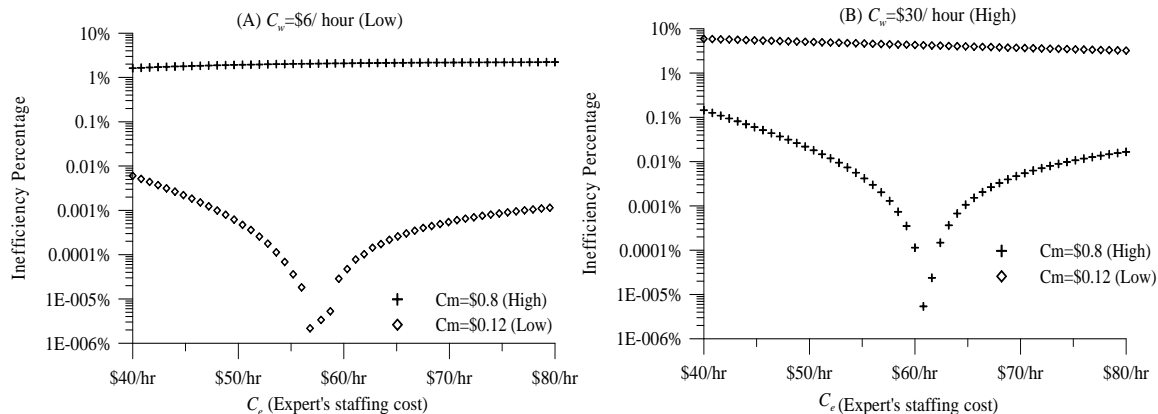


Figure 2: Inefficiency in system S_b

we explore the client’s optimal outsourcing strategy in the presence of both vendor process choice and vendor staffing cost advantages.

We use the following parameters: $\lambda = 100$ customer requests per minute, $\mu_d = 2$ per minute, $1/\bar{t}_g = 2/3$ per minute, $\Delta_g = 0$, $1/\bar{t}_e = 1.5$ per minute, $\Delta_e = 0$, $C_g = \$10$ per hour, $C_w = \$6$ or $\$30$ per hour, $C_m = \$0.12$ or $\$0.8$ per mistreatment, C_e from $\$40$ to $\$80$ per hour, and $f(x) = 1 - x$. The parameters are chosen so that (1) a two-level structure outperforms a one-level structure for a majority of the experiments; and (2) the size of the system in each experiment is reasonably within the QED regime. The results described here are consistent with the results from experiments with many other parameter sets, including those with $\Delta_g > 0$ and $\Delta_e > 0$.

To illustrate the potential costs of outsourcing both gatekeepers and experts, for each parameter set we calculate the optimal thresholds for the centralized system and system S_b (k^* and k_*^b) and the optimal costs for both systems. Define the system *inefficiency* as the relative cost difference between the system optimum and the client’s smallest possible cost in S_b , i.e., $|\pi_c^b(Q_*^b, P_*^b) - \pi_c(k^*)|/\pi_c(k^*)$. Figure 2A displays the system inefficiencies when $C_w = \$6$. Note that the plot has a log scale on the y-axis. When $C_m = \$0.12$, the parameters are close to satisfying the coordination condition defined in Proposition 3, and contracts under these parameters are almost efficient. When $C_m = \$0.8$, the parameters are further from the coordination condition and inefficiency is higher. Figure 2B shows a similar effect: the pair $\{C_w = \$30, C_m = \$0.8\}$ allows for an efficient contract, while $\{C_w = \$30, C_m = \$0.12\}$ does not. In the remainder of this section we will refer to $C_w = \$6$ and $C_m = \$0.12$ as (relatively) ‘low’ parameter values and to $C_w = \$30$ and $C_m = \$0.8$ as ‘high’ values. Therefore, contracts for the {low, high} and {high, low} parameter pairs are inefficient.

Now we decompose the system costs and see how each component changes when we outsource

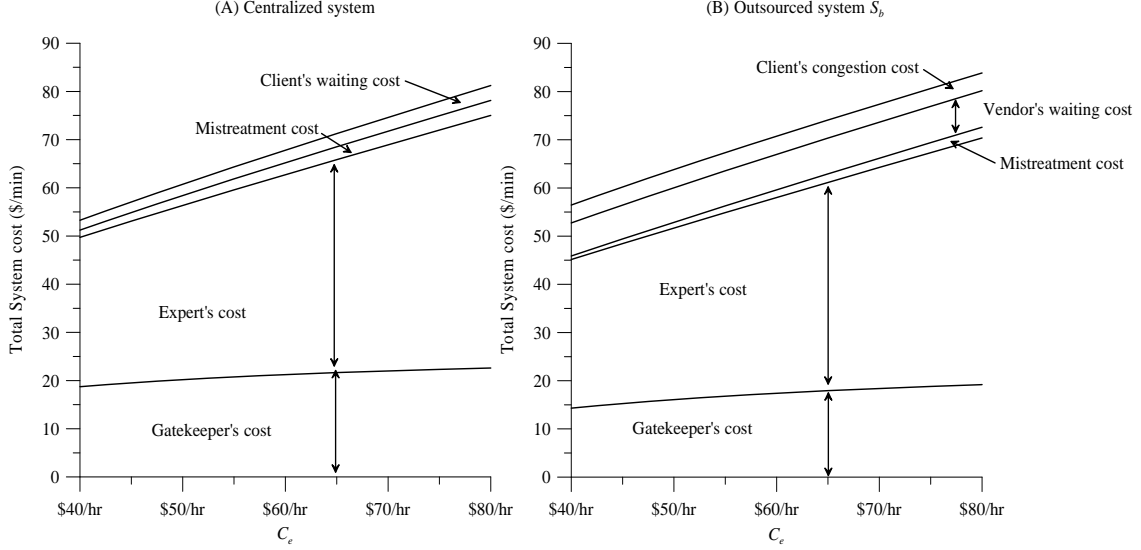


Figure 3: Cost decomposition for (A) the centralized system and (B) system S_b

the system. Consider the (high, low) case, with $C_w = \$30$ per hour and $C_m = \$0.12$. Figure 3A shows, for the centralized system, the gatekeeper's staffing cost, the expert's staffing cost, the mistreatment cost, and the client's expected total waiting cost, $C_w(W_g\lambda + W_e\lambda_e)$. Figure 3B shows similar costs in system S_b , with the client's total waiting cost separated into two components. The first component is the 'vendor's waiting cost,' $Q^b(W_g\lambda + W_e\lambda_e)$, the expected payment from the vendor to the client due to the system-time penalty. The second component is the 'client's congestion cost,' $(C_w - Q^b)(W_g\lambda + W_e\lambda_e)$, the cost of customer waiting-time net the system time penalty.

When looking from Figure 3A to B, we see visually that the primary increase in cost is due to additional waiting time. On the right-hand sides of plots A and B, when $C_e = \$80$, the quantity $C_w(W_g\lambda + W_e\lambda_e)$ rises by 263% when switching from the centralized system (plot A) to the outsourced system (plot B). On the left-hand sides of the plots, when $C_e = \$40$, the relative increase is even larger, 412%. Figure 3B shows that most of this increase is paid for by the vendor, in the form of the system-time penalty, but that the residual client's congestion cost remains substantial as well. Staffing costs decline by 6%-9% while mistreatment costs decline by 29%-52%.

To explain these changes, note that because the mistreatment cost is low, the client prefers a high threshold. To force the vendor to set the threshold at a high level, she offers a system-time penalty Q^b that is smaller than the high waiting cost in the centralized system, C_w . Lowering Q^b essentially lowers the mistreatment penalty to the vendor and pushes the vendor towards the client's

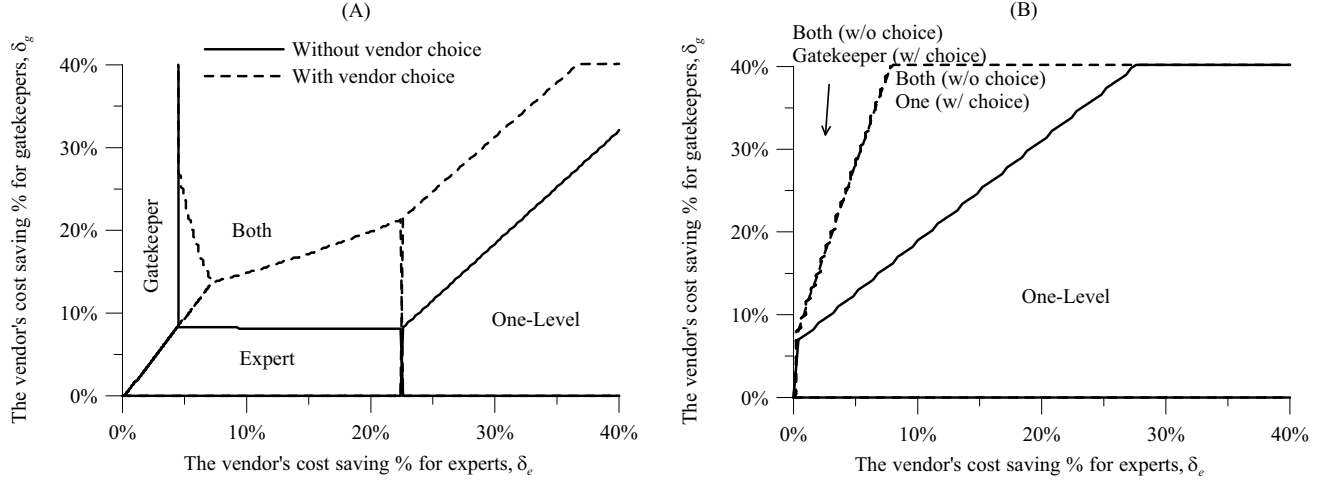


Figure 4: The outsourcing decision when (A) $C_m = \$0.12$, and (B) $C_m = \$0.8$

desired k . Lowering Q^b , however, also reduces the cost of waiting for the vendor, and therefore the vendor reduces staffing levels and the client incurs higher congestion costs. We also find that when $C_w = \$6$ and $C_m = \$0.8$ (low, high), the reverse is true: mistreatment and staffing costs rise and waiting costs fall as compared to the centralized system.

Now consider a vendor who offers a gatekeeper's cost of $C'_g = C_g(1 - \delta_g)$ and an expert's cost of $C'_e = C_e(1 - \delta_e)$, in which C_g and C_e are the client staffing costs and δ_g and δ_e represent the cost savings provided by the vendor. Given these cost advantages, we compute the optimal total cost for the client under each outsourcing scenario and then make the outsourcing decision for her by selecting the system that provides the lowest cost. The results are shown in Figure 4, for which we set $C_e = \$60/\text{hour}$ and $C_w = \$30/\text{hour}$.

In Figure 4, the solid lines delineate regions over which each process is optimal for the client when the process choice is contractible. The dotted lines delineate regions over which each process is optimal when vendors can deviate, and therefore the client uses the optimal incentive-compatible contracts described in the previous section. For example, when vendors can deviate, Figure 4A shows that an expert-only system should be outsourced over the region from $\delta_e = 0\%$ to 23% along the x-axis, up to the slanted dotted line above. In Figure 4B, the area specifying a one-level design is larger under vendor choice. In the remaining area in the upper left of the graph the client should outsource a two-level system when vendors do not deviate (the areas labeled "Both (w/o choice)"), and with deviation the client should outsource gatekeepers only ("Gatekeeper (w/ choice)").

Now we examine the financial impact of vendor deviation from the client's optimal process

choice. Suppose that the process choice is not contractible and that the client does not anticipate deviation; she only offers the contracts derived in Section 5. We find that the loss in client profits can be quite large. In Figure 4A, the vendor will choose to deviate from S_b over the entire region labeled "Both." In this region, if the client offers the contracts defined by Proposition 4 and the process choice is not contractible, the vendor will choose to operate a one-level system, resulting in costs to the client that are up to 36% larger than the system with no deviation. Similar deviations in Figure 4B can be even more costly, up to 47% larger.

If the client anticipates deviation, the two-level contracts must be adjusted to be incentive compatible, as in Proposition 7. Therefore, the coordination cost for system S_b rises, and the regions over which S_b is optimal shrinks. In Figure 4B, for example, the S_b region disappears under vendor process choice. The difference between the client costs with no vendor deviation and the client costs under the incentive-compatible contracts is the break-even cost of monitoring and enforcing the vendor's process choice. In both Figures 4A and B, we find that this break-even monitoring cost ranges from 0 to as high as 6%. Offering incentive-compatible contracts significantly reduces, but does not eliminate, the cost of vendor process deviation.

8 Model Extensions

8.1 Gatekeeper Misdiagnosis

We have thus far assumed that the gatekeeper can perfectly diagnose the complexity of a case. Relaxing that assumption will not alter our main results. Suppose that the gatekeeper diagnoses a case as having complexity \hat{x} but because of misdiagnoses the true complexity, x , has a probability density of $g(x, \hat{x})$. In that case the probability of being treated correctly is given by:

$$\hat{f}(\hat{x}) = \int_0^1 f(x)g(x, \hat{x})dx.$$

If we make the reasonable assumption that for $\hat{x}_1 > \hat{x}_2$, $g(x, \hat{x}_1)$ stochastically dominates $g(x, \hat{x}_2)$, then $\hat{f}(\hat{x})$ is decreasing in \hat{x} , and we can replace $f(x)$ and $F(x)$ in our model accordingly. To preserve all of our analytical results, we revise the expected treatment time functions:

$$T_g(k) = \int_0^k \int_0^1 t(y)g(y, \hat{x})dyd\hat{x}.$$

8.2 Expert Mistreatment

In some settings both gatekeepers and experts may mistreat customers, so that a customer may leave the system without successful treatment. Let $f_i(x)$, $i \in \{g, e\}$, be the probability that a type i

worker successfully treats a problem with complexity x . Assume that functions $f_i(x)$ are continuous and differentiable, $f_g(x) \leq f_e(x)$, and $f'_i(x) < 0$. Let $\bar{f}_i(x) = 1 - f_i(x)$, and let C_m^i be the cost per mistreatment caused by worker type i . To describe the relationship between f_g and f_e for each problem of complexity x , define the random variable z_i , $i \in \{g, e\}$, where $z_i = 0$ if i successfully treats a problem and $z_i = 1$ if i mistreats. Therefore, $E(z_i|x) = \bar{f}_i(x)$, and the covariance between z_g and z_e for a problem with complexity x is,

$$\text{cov}(z_g, z_e|x) = \Pr\{z_g = 1|z_e = 1, x\} \bar{f}_e(x) - \bar{f}_g(x) \bar{f}_e(x).$$

Define $\psi(x) = \Pr\{z_g = 1|z_e = 1, x\}$. The quantity $\psi(x)$ is the probability that the gatekeeper mistreats a customer of complexity x , given that an expert would mistreat the customer if the customer is referred to the expert.

In the client's cost function (Equation 4), the expected cost of mistreatment per customer, given treatment threshold k , is $C_m(k - F(k))$. Given that experts can mistreat as well, we can rewrite this cost as,

$$\begin{aligned} & C_m^e \left(1 - \int_0^1 f_e(x) dx \right) + C_m^g \left(k - \int_0^k f_g(x) dx \right) \\ & + C_m^e \int_0^k (\psi(x) + (1 - \psi(x)) f_e(x) - 1) dx. \end{aligned} \quad (10)$$

Note that the first term of this expression is a constant. Now redefine $C_m = C_m^g$ and

$$f(x) = f_g(x) - \frac{C_m^e (\psi(x) + (1 - \psi(x)) f_e(x) - 1)}{C_m^g}. \quad (11)$$

As before, let $F(k) = \int_0^k f(x) dx$. Given these definitions, the mistreatment cost $C_m(k - F(k))$, from Equation (4) differs from (10) only by the constant first term. Therefore, by applying $C_m = C_m^g$ and (11), all of the previous results continue to hold when experts can also mistreat.

8.3 Additional Variations and Extensions

Depending on the business environment, the structure of gatekeeper systems and outsourcing decisions may be driven by factors that are not included in the model described above. We have considered three variations that yield similar results: (i) impatient customers, (ii) clients penalized for total system time, and (iii) waiting costs that differ between the gatekeeper and expert levels. For all of these the S_e and S_g systems can be coordinated and, given non-trivial mistreatment costs, coordinating contracts do not exist for S_b systems. With negligible mistreatment costs and system time penalties a coordinating contract does exist for S_b systems. Thus, for example, the

model would predict that back-office processes with insignificant mistreatment costs are generally outsourced as a whole, rather than in parts.

Finally, we have assumed that for a problem of complexity x , both the treatment time and probability of success are exogenous, so that there is no endogenous trade-off between the gatekeeper's probability of success $f(x)$ and the mean gatekeeper treatment time $t_g(x)$. Thus we are assuming that once all relevant information is collected from the customer extra time will not improve the gatekeeper's chances. We also assume that an agent does not have the option to adjust a particular customer's service time in response to system-time penalties. These assumptions are appropriate for applications where the agents follow a clear script when attempting to solve a problem. In general, the trade-off between treatment time and service success is an interesting area of study (e.g., see de Véricourt and Sun, 2009), but is beyond the scope of this paper.

9 Conclusions

In this paper we investigate what happens when a firm outsources a service process to a vendor and gives up control over how the work is done. We formulate a model of a two-stage process where the vendor decides what type of workers do each stage of the process, in addition to determining the staffing levels. We find that when a firm relinquishes process control it cannot create efficient coordinating contracts based upon easily observable measures. We also identify contracts that, while inefficient, are optimal for the client. Although these analytical results are derived under various assumptions and conditions, we are able to use an extensive numerical study to show that they are not restrictive. Numerical experiments also show how the loss of control through outsourcing can change the customer experience, leading to longer customer waiting time or longer interactions with incompetent workers. The loss of control also introduces inefficiency into the process, which can negate the lower marginal labor costs that the vendor might offer. The cost of inefficiency can also increase if the client cannot ensure that the vendor conforms to a specific system design. Therefore, the firm may be better off limiting the scope of the outsourcing arrangement.

References

- Allon, G., A. Federgruen. 2008. Outsourcing service processes to a common service provider under price and time competition. Working paper. Accessed September 16, 2011, http://www.kellogg.northwestern.edu/faculty/allon/htm/research/outsourcing_dec_13.pdf

- Akşın, Z., F. de Véricourt, F. Karaesmen. 2008. Call center outsourcing contract analysis and choice. *Management Science* **54**(2) 354–368.
- Borst, S., A. Mandelbaum, M. I. Reiman. 2004. Dimensioning large call centers. *Operations Research* **52**(1) 17–34.
- Brekke, K.R., R. Nuscheler, O.R. Straume. 2007. Gatekeeping in health care. *Journal of Health Economics* **26** 149–170.
- Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29**(3) 567–588.
- Hasija, S., E. J. Pinker, R. A. Shumsky. 2005. Staffing and routing in a two-tier call centre. *International Journal of Operational Research* **1**(1/2) 8–29.
- Hasija, S., E. J. Pinker, R. A. Shumsky. 2008. Call center outsourcing contracts under information asymmetry. *Management Science* **54**(4) 793–807.
- Lu, L. X., J. A. Van Mieghem, R. C. Savaskan. 2009. Incentives for quality through endogenous routing. *Manufacturing & Service Operations Management* **11**(2) 254–273.
- Mackie, J. 2007. Benchmarking. Working Paper, EquaTerra. Accessed June 28, 2010, <http://www.equaterra.com/fw/main/Benchmarking-243.html>.
- Malcomson, J.M. 2004. Health service gatekeepers. *RAND Journal of Economics* **35**(2) 401–421.
- Marinoso, B.G., I. Jelovac. 2003. GPs’ payment contracts and their referral practice. *Journal of Health Economics* **22**(4) 617–635.
- Ren, Z. J., F. Zhang. 2009. Service outsourcing: Capacity, quality, and correlated Costs. Working paper.
- Ren, Z. J., Y-P. Zhou. 2008. Call center outsourcing: Coordinating staffing level and service quality. *Management Science* **54**(2) 369–383.
- Shumsky, R. A., E. J. Pinker. 2003. Gatekeepers and referrals in services. *Management Science* **49**(7) 839–856.
- Taylor, J.W. 2010. Exponentially weighted methods for forecasting intraday time series with multiple seasonal cycles. *International Journal of Forecasting* **26** 627–646.
- de Véricourt, F., P. Sun. 2009. Judgement accuracy under congestion in service systems. *Working paper, Fuqua School of Business, Duke University*. Accessed July 19, 2010, http://faculty.fuqua.duke.edu/~psun/bio/AccuracyCongestion_7.9.09.pdf.

Acknowledgments

The authors thank Mor Armony, Sameer Hasija, Avi Mendelbaum, Guillaume Roels, three anonymous reviewers, and an associate editor for their helpful comments.

Appendix A: Accuracy of The Approximation

We use simulation to test the accuracy of the $M/M/N$ - QED approximation described in Section 4. In our simulation we assume that arrivals to the gatekeeper are Poisson and that all diagnosis and treatment times, given x , are exponential. Therefore, a gatekeeper's total service is the sum of the exponential diagnosis time and the complexity-dependent treatment time, if needed. Let $v(x)$ be the simulated total gatekeeper service time for a problem with complexity x , let κ be an exponential random variable with mean $1/\mu_d$, and let $\tau_g(x)$ be an exponential random variable with mean $t_g(x)$. Therefore,

$$v(x) \sim \begin{cases} \kappa + \tau_g(x) & \text{if } x < k^* \\ \kappa & \text{if } x \geq k^*. \end{cases}$$

Likewise, expert treatment time is distributed as an exponential random variable with mean $t_e(x)$.

Given each set of system parameters, we calculate the treatment threshold and staffing decisions k^* , $n_e^*(k^*, C_w)$ and $n_g^*(k^*, C_w)$ from the $M/M/N$ - QED approximation. We then implement the optimal threshold in the simulation. To find the optimal staffing level in the simulation, we search around n_e^* and n_g^* and identify the staffing levels that generate the lowest total cost in the simulation.

We consider the following parameters in the simulation: $\lambda = 100, 200, \text{ and } 500$ customer requests per minute; $C_e = \$50, \$75, \text{ and } \$100$ per hour; $C_g = 0.5C_e, 0.25C_e, \text{ and } 0.1C_e$; $C_m = \$0.02, \$0.2, \text{ and } \$1.0$ per mistreatment; and $C_w = \$10, \$30, \text{ and } \$90$ per hour. We use treatment time functions $t_g(x) = \bar{t}_g(1 + x\Delta_g)$ and $t_e(x) = \bar{t}_e(1 + x\Delta_e)$. We consider the following parameters for the service times: $\mu_d = 2$ per minute; $\bar{t}_g = 0.5, 0.83, \text{ and } 1.5$ minutes; $\bar{t}_e = 0.9\bar{t}_g, 0.75\bar{t}_g, \text{ and } 0.5\bar{t}_g$ minutes; $\Delta_e = 0.1, 0.3, \text{ and } 0.5$; and $\Delta_g = 0.1, 0.3, \text{ and } 0.5$. Finally, we use a linear treatment function, $f(x) = 1 - bx$, with $b = 1$. To ensure that experiments generated a full range of values for k^* (there were relatively few experiments with $k^* > 0.9$ with the parameters described above), we also ran experiments with $b = 0.9$ and a subset of the other parameters. This generated 15,743 two-level systems that satisfy the convexity constraints of Lemma 2.

For each set of parameters we focus on the approximation's accuracy in terms of the total cost. As a measure of accuracy, we use the absolute value of the difference between the total

Table 1: Total cost error (Average/Max error) in %.

min load	k^*					total
	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1	
0-50	0.2 / 0.9	0.2 / 0.9	0.2 / 1.0	0.6 / 2.2	1.8 / 6.0	0.7 / 6.0
50-75	0.2 / 0.9	0.2 / 0.8	0.3 / 1.2	0.9 / 2.8	1.7 / 4.3	0.5 / 4.3
75-100	0.2 / 0.8	0.2 / 0.9	0.3 / 1.1	0.9 / 2.5	1.8 / 3.7	0.6 / 3.7
100-150	0.2 / 0.8	0.2 / 0.6	0.4 / 1.4	0.9 / 2.4	1.4 / 3.2	0.5 / 3.2
150-225	0.2 / 0.7	0.2 / 0.7	0.5 / 1.1	0.9 / 2.2	1.3 / 2.5	0.5 / 2.5
>225	0.3 / 0.7	0.3 / 0.8	0.5 / 1.1	0.9 / 1.7	1.2 / 2.1	0.5 / 2.1
total	0.2 / 0.9	0.2 / 0.9	0.4 / 1.4	0.8 / 2.8	1.6 / 6.0	0.6 / 6.0

cost from simulation and the total cost from the $M/M/N$ - QED approximation, expressed as a percentage of the total cost from simulation. Table 1 shows the average and maximum percentages for various groups of experiments. The column category k^* is the optimal threshold, given the parameters. The row category *min load* is the smaller of the gatekeeper and expert loads: $\text{min load} = \min [\lambda/\mu_g(k^*), (1 - F(k^*))\lambda/\mu_e(k^*)]$. For example, for the 929 experiments with $0.2 \leq k^* < 0.4$ and a minimum load from 50 – 75, the average approximation error for the total cost was 0.2% and the maximum was 0.8%. The number of experiments associated with each entry in the table varies from approximately 150 to 1000, and the experiments are roughly evenly distributed over the entire ranges of minimum loads and k^* . Overall, the average cost error was 0.6%, with a maximum of 6%. We see that the accuracy of the approximation depends both upon the size of the load placed on each resource pool, as well as the value of k^* . Because the approximations are based on limiting behavior as the system size increases, it is no surprise the performance is best for large loads: for systems with minimum loads above 50, the average error is 0.5% and the maximum error is 4.3%.

We believe that the growth in approximation error as k^* rises is due to the gatekeeper service-time distribution. Recall that the coefficient of variation (CV) of an exponential random variable is equal to 1. When k^* is small, gatekeepers mostly diagnose, so that the CV of their service times is close to 1. When k^* is large, most gatekeeper service times are the sums of two independent exponential random variables, so that the CV of gatekeeper service time falls below 1. When k^* is large, however, the dependence of treatment times on treatment complexity *increases* gatekeeper service time variability, bringing the CV back closer to 1. These effects and others due to the

interactions among k^* and staffing levels help to explain the overall accuracy of the approximation.

Appendix B: Proofs

Lemma 1 Proof. Substitute for N and W in (2), writing the total cost in terms of the standardized excess capacity β , $\widehat{C}(\beta) = C_n(\rho + \beta\sqrt{\rho}) + C_w\sqrt{\rho}\alpha(\beta)$. Then $d\widehat{C}(\beta)/d\beta = C_n\sqrt{\rho} + C_w\sqrt{\rho}(d\alpha(\beta)/d\beta)$ and $d^2\widehat{C}(\beta)/d\beta^2 = C_w\sqrt{\rho}(d^2\alpha(\beta)/d\beta^2)$. Because $d(\alpha(\beta))/d\beta < 0$ and $d^2(\alpha(\beta))/d\beta^2 > 0$ (Borst et al., 2004), $\widehat{C}(\beta)$ is convex in β . As a result, when $C_w > 0$, a solution satisfying $\partial\widehat{C}(\beta)/\partial\beta = 0$ exists, i.e., the optimal standardized excess capacity β^* , and it satisfies $d(\alpha(\beta))/d\beta|_{\beta=\beta^*} = -C_n/C_w$.

Now let $h(\beta, C_w) = d(\alpha(\beta))/d\beta + C_n/C_w$. By the definition of β^* , $h(\beta^*, C_w) = 0$. Using implicit differentiation:

$$\frac{d\beta^*}{dC_w} = -\frac{\partial h(\beta, C_w)/\partial C_w}{\partial h(\beta, C_w)/\partial \beta} \Big|_{\beta=\beta^*}.$$

The numerator $\partial h(\beta, C_w)/\partial C_w = -1/C_w^2$, while the denominator is, $\partial h(\beta, C_w)/\partial \beta|_{\beta=\beta^*} = \alpha''(\beta)|_{\beta=\beta^*}$. Therefore, $d\beta^*/dC_w = 1/[C_w^2\alpha''(\beta^*)] > 0$, in which the inequality follows from $C_w > 0$ and the fact that $\alpha(\beta)$ is strictly convex in β . ■

Lemma 2 Proof. $\pi_c(k) = \widehat{C}_m(k) + \widehat{C}_e(k) + \widehat{C}_g(k)$, in which $\widehat{C}_m(k) = C_m\lambda(k - F(k))$, $\widehat{C}_e(k) = C_e\rho_e(1 + 2\Theta_e(k, C_w))$, and $\widehat{C}_g(k) = C_g\rho_g(1 + 2\Theta_g(k, C_w))$.

$$\frac{\partial^2\widehat{C}_m(k)}{\partial k^2} = -C_m f'(k)\lambda \geq 0, \quad (\text{A1})$$

$$\frac{\partial^2\widehat{C}_e(k)}{\partial k^2} = -C_e\lambda(1 + \Theta_e(k, C_w))V(k) - C_e\lambda\Theta_e(k, C_w)\frac{t_e^2 f^2(k)}{2T_e}, \quad (\text{A2})$$

$$\frac{\partial^2\widehat{C}_g(k)}{\partial k^2} = C_g\lambda t'_g + \frac{C_g\lambda\eta_g(C_w)t'_g}{2\sqrt{\lambda/\mu_g}} - \frac{C_g\lambda t_g^2\eta_g(C_w)}{4\sqrt{\lambda/\mu_g}^3}. \quad (\text{A3})$$

Recall the definition of $V(k) = -\partial^2 T_e(k)/\partial k^2$. By Condition 2 $V(k) < 0$. This inequality, combined with Condition 1, $\lambda > \widetilde{\lambda}_c$, ensures that the sum of Equations (A1), (A2), and (A3) is positive. Therefore, the client's cost function is strictly convex. ■

Proposition 1 Proof. The contract implies that the client pays the transfer payment $\lambda_e[P^e - Q^e(T_e(k)/(1 - F(k)) + W_e)]$ to the vendor and the vendor staffs the expert subsystem with a unit waiting cost Q^e . By applying the square root rule to both subsystems, the client's cost can be formulated as a function of k , Q^e , and P^e ,

$$\begin{aligned} \pi_c^e(k, Q^e, P^e) &= C_m\lambda(k - F(k)) + C_g\rho_g[1 + 2\Theta_g(k, C_w)] \\ &\quad + (1 - F(k))\lambda(P^e - Q^e T_e(k)/(1 - F(k))) + (C_w - Q^e)\alpha(\beta_e^*(Q^e))\sqrt{\rho_e}. \end{aligned}$$

The value of Q^e in this contract ensures that the vendor staffs the expert subsystem at the optimal level if the client sets the threshold to k^* . It also ensures that the client does not bear the waiting cost at the vendor, e.g., $(C_w - Q^e) \alpha(\beta_e^*(Q^e)) \sqrt{\rho_e} = 0$ in the client's cost function. By substituting (Q^e, P^e) into the client's cost function,

$$\pi_c^e(k) = C_m \lambda (k - F(k)) + C_g \rho_g [1 + 2\Theta_g(k, C_w)] + C_e \rho_e [1 + 2\Theta_e(k, C_w)].$$

Because the lowest cost the client can achieve is $\pi_c(k^*)$, the client chooses the optimal threshold k^* and the optimal staffing level for the gatekeeper subsystem $n_g^*(k^*, C_w)$. ■

Lemma 3 Proof. The second derivative of the profit function $\pi_v^g(k)$ is

$$\frac{\partial^2 \pi_v^g(k)}{\partial k^2} = \lambda [f'(k) R^g - Q^g t'_g - C_g t'_g] + \frac{C_g \eta_g(Q^g) \sqrt{\mu_g}}{4} [t_g^2 \mu_g - 2t'_g] \sqrt{\lambda}.$$

$\lambda > \tilde{\lambda}_g$, implies that the right-hand side is positive, and hence $\pi_v^g(k)$ is a strictly concave. ■

Proposition 2 Proof. The contract implies that the client pays the transfer payment $\lambda[P^g + R^g F(k) - Q^g(1/\mu_d + T_g + W_g)]$, and the vendor staffs the gatekeeper subsystem, given the waiting cost Q^g . Because the threshold k^g is determined by solving $\partial \pi_v^g(k)/\partial k = 0$ and the expert's staffing is determined by the client using the square root staffing rule, the client's cost is a function of the contract terms,

$$\begin{aligned} \pi_c^g(Q^g, P^g, R^g) &= C_m \lambda (k^g - F(k^g)) + (P^g + F(k^g) R^g) \lambda - Q^g \lambda (1/\mu_d + k^g T_g) \\ &\quad + (C_w - Q^g) \alpha(\beta_g^*(Q^g)) \sqrt{\rho_g(k^g)} + C_e \rho_e(k^g) [1 + 2\Theta_e(k^g, C_w)]. \end{aligned}$$

First, the value of Q^g ensures that the threshold is set to the optimal level. Furthermore, because $Q^g = C_w$, the waiting cost that the client bears, $(C_w - Q^g) \alpha(\beta_g^*(Q^g)) \sqrt{\rho_g}$, becomes 0. Specifically, by using the system-time penalty, the client shifts the waiting cost to the vendor and force him to staff optimally. Second, for $k^g = k^*$, we require $r^g = r^*$. Specifically, $R^g = t_g [C_w + C_g (1 + \Theta_g(k^*, C_w))] / r^*$, in which Q^g has been replaced by C_w . Because $k^g = k^*$, the expert's staffing level for the client is optimal as well. As a result, this contract coordinates the system. The contract also needs to satisfy the vendor's reservation level, i.e., $\pi_v^g(k^*) = 0$. Thus, to ensure that the client extracts all the vendor's profit while the vendor is still willing to accept the contract, we use

$$P^g = C_g \rho_g [1 + 2\Theta_g(k^*, C_w)] / \lambda - F(k^*) R^g + C_w (1/\mu_d + T_g).$$

■

Lemma 4 Proof. We first decompose the profit function $\pi_v^b(k)$ into three parts: $\widehat{R}^b(k) - \widehat{C}_g^b(k) - \widehat{C}_e^b(k)$, in which $\widehat{R}^b(k) = P^b\lambda - Q^b\lambda[1/\mu_d + T_e + T_g]$, $\widehat{C}_g^b(k) = C_g\rho_g(1 + 2\Theta_g(k, Q^b))$, and $\widehat{C}_e^b(k) = C_e\rho_e(1 + 2\Theta_e(k, Q^b))$. (A1) is replaced by $Q^b(-V(k) + t'_g)$ and the waiting cost C_w in Equation (A2) and (A3) is replaced by Q^b . As a result, if the conditions are satisfied, $\pi_v^b(k)$ is strictly concave. ■

Lemma 5 Proof. We abbreviate $\Theta_e(k^b, Q^b)$ as Θ_e and $\Theta_g(k^b, Q^b)$ as Θ_g . By taking the derivative of Equation (7) we find that $\partial k^b/\partial Q^b$ satisfies,

$$A_v^b \frac{\partial k^b}{\partial Q^b} = t_g - t_e f(k^b) + \frac{t_g \alpha(\beta_g^*(Q^b))}{2\sqrt{\lambda/\mu_g}} - \frac{t_e f(k^b) \alpha(\beta_e^*(Q^b))}{2\sqrt{\lambda T_e}}, \quad (\text{A4})$$

in which $A_v^b \lambda$ is the second derivative of $\pi_v^b(k)$. Under Conditions 1 and 2, we know that $\pi_v^b(k)$ is concave, and hence $A_v^b < 0$. Condition 3 guarantees that $t_g - t_e f(k^b) \geq 0$ and that the difference between the last two terms in the right-hand side of Equation (A4) is positive. Because $A_v^b < 0$ and the right-hand side of Equation (A4) is positive, $\partial k^b/\partial Q^b < 0$. ■

Proposition 3 Proof. Let $k^b(Q^b)$ be the vendor's optimal threshold, given Q^b , and let $n_g^*(k, Q^b)$ and $n_e^*(k, Q^b)$ be the vendor's optimal gatekeeper's and expert's staffing levels, given threshold k and penalty Q^b . We want to show that $k^b(Q^b) = k^*$, $n_g^*(k^b(Q^b), Q^b) = n_g^*(k^*, C_w)$, and $n_e^*(k^b(Q^b), Q^b) = n_e^*(k^*, C_w)$ if and only if Q^b and the parameters of the model satisfy the conditions above.

First, note that k^* is unique due to the strict convexity of the objective function (see Lemma 2). Given the workflow implied by k^* , both $n_g^*(k^*, C_w)$, and $n_e^*(k^*, C_w)$ are unique (Borst et al., 2004). If $Q^b = [r^* C_e t_e (1 + \Theta_e(k^*, C_w)) - C_g t_g (1 + \Theta_g(k^*, C_w))] / [t_g - r^* t_e]$, then $r^b = r^*$. and so $k^b(Q^b) = k^*$. Given k^* , if $Q^b = C_w$, the vendor applies the square root staffing so that for $i \in \{g, e\}$, $n_i^*(k^b(Q^b), Q^b) = n_i^*(k^*, C_w)$.

To show necessity: If k^* is optimal for the vendor then $r^b = r^*$, and therefore

$$Q^b = [r^* C_e t_e (1 + \Theta_e(k^*, Q^b)) - C_g t_g (1 + \Theta_g(k^*, Q^b))] / [t_g - r^* t_e].$$

By the vendor's square root staffing rule, $n_i^*(k^*, Q^b) = \rho_i(k^*) + \beta_i^*(Q^b) \sqrt{\rho_i(k^*)}$. Recall that the optimal centralized staffing solution is,

$$\text{If } n_i^*(k^*, Q^b) = n_i^*(k^*, C_w) = \rho_i(k^*) + \beta_i^*(C_w) \sqrt{\rho_i(k^*)}.$$

Then $\beta_i^*(Q^b) = \beta_i^*(C_w)$ and by Lemma 1, $Q^b = C_w$. ■

Corollary 2 Proof. We abbreviate $\Theta_e(k^*, C_w)$ as Θ_e and $\Theta_g(k^*, C_w)$ as Θ_g . After some algebra, $C_m(1 - f(k^*)) = C_w(t_g - t_e f(k^*))$. Taking the derivative with respect to C_w ,

$$\frac{dC_m}{dC_w}(1 - f(k^*)) = A^* \frac{\partial k^*}{\partial C_w} + (t_g - t_e f(k^*)),$$

in which $A^* = [C_w t_e - C_m](-f'(k^*)) + C_w(t'_g - t'_e f(k^*))$. Next, suppose that given Conditions 1 and 2 are satisfied, $\partial k^*/\partial C_w < 0$. Also from Condition 2, $t_g - t_e f(k^*) \geq 0$, because $f(k) \leq 1$, for $0 \leq k \leq 1$. After replacing C_m with $C_w(t_g - t_e f(k^*)) / (1 - f(k^*))$,

$$A^* = C_w [(t_e - t_g)(-f'(k^*)) + (t'_g - t'_e f(k^*))(1 - f(k^*))] / (1 - f(k^*)).$$

From Condition 3, $A^* \leq 0$. Therefore, given the four conditions, $dC_m/dC_w \geq 0$.

Now we prove that $\partial k^*/\partial C_w < 0$. From Equation (5), $r^*(C_m + C_e t_e(1 + \Theta_e)) = C_m + C_g t_g(1 + \Theta_g)$. Taking derivatives of both sides with respect to C_w , we have

$$C_m \frac{dr^*}{dC_w} + C_e \frac{d(t_e r^*(1 + \Theta_e))}{dC_w} = C_g \frac{d(t_g(1 + \Theta_g))}{dC_w}.$$

$$C_m dr^*/dC_w = C_m f'(k^*) \partial k^*/\partial C_w.$$

$$C_e \frac{d(t_e r^*(1 + \Theta_e))}{dC_w} = C_e \frac{dk^*}{dC_w} \left[V(k^*)(1 + \Theta_e) + t_e f(k^*) \frac{\partial \Theta_e}{\partial k^*} \right] + C_e t_e f(k^*) \frac{\partial \Theta_e}{\partial C_w}.$$

Because $\Theta_e = \eta_e(C_w) / (2\sqrt{\rho_e})$, $\rho_e = \lambda T_e$ and $\eta_i(C_w) = \beta_e^*(C_w) + (C_w/C_e) \alpha(\beta_e^*(C_w))$, we have

$$\frac{\partial \Theta_e}{\partial k^*} = \Theta_e \frac{t_e f(k^*)}{2T_e} \text{ and } \frac{\partial \Theta_e}{\partial C_w} = \frac{1}{C_e} \frac{\alpha(\beta_e^*(C_w))}{2\sqrt{\lambda T_e}},$$

in which the third equality uses Lemma 1. Similarly,

$$C_g \frac{d(t_g(1 + \Theta_g))}{dC_w} = C_g \left[t'_g(1 + \Theta_g) + t_g \frac{\partial \Theta_g}{\partial k^*} \right] \frac{\partial k^*}{\partial C_w} + C_g t_g r^* \frac{\partial \Theta_g}{\partial C_w},$$

in which

$$\frac{\partial \Theta_g}{\partial k^*} = -\Theta_g \frac{t_g}{2/\mu_g} \text{ and } \frac{\partial \Theta_g}{\partial C_w} = \frac{1}{C_g} \frac{\alpha(\beta_g^*(C_w))}{2\sqrt{\lambda/\mu_g}}.$$

Finally, by substituting the three terms back, we have

$$\begin{aligned} & \frac{dk^*}{dC_w} \left[(C_m f'(k^*) + C_e(1 + \Theta_e)V(k^*)) + C_e \Theta_e f^2(k^*) t_e^2 / 2T_e - C_g t'_g(1 + \Theta_g) + \frac{C_g t_g^2 \Theta_g}{2/\mu_g} \right] \\ &= t_g \frac{\alpha(\beta_g^*(C_w))}{2\sqrt{\lambda/\mu_g}} - f(k^*) t_e \frac{\alpha(\beta_e^*(C_w))}{2\sqrt{\lambda T_e}} \text{ or, } \frac{dk^*}{dC_w} A_c = t_g \frac{\alpha(\beta_g^*(C_w))}{2\sqrt{\lambda/\mu_g}} - f(k^*) t_e \frac{\alpha(\beta_e^*(C_w))}{2\sqrt{\lambda T_e}}, \end{aligned}$$

in which $-A_c \lambda$ is the second derivative of $\pi_c(k)$, and hence $A_c < 0$ if Conditions 1 and 2 are satisfied. Moreover, Condition 3 implies that the right-hand side of the equation is positive. As a result, $\partial k^*/\partial C_w$ is negative and k^* is strictly decreasing with respect to C_w . ■

Proposition 4 Proof. We abbreviate $\beta_e^*(Q^b)$ as β_e^* , $\beta_g^*(Q^b)$ as β_g^* , $\Theta_e(k^b, Q^b)$ as Θ_e and $\Theta_g(k^b, Q^b)$ as Θ_g . By Lemma 5 Q^b and k^b have a one-to-one mapping.

Next, the first and second derivatives of $\pi_c^b(Q^b(k^b))$ are

$$\frac{\partial \pi_c^b(Q^b(k^b))}{\partial k^b} = C_m \lambda (1 - f(k^b)) + C_g t_g \left(1 + \Psi_g(k^b)\right) \lambda - C_e t_e f(k^b) \left(1 + \Psi_e(k^b)\right) \lambda, \text{ and}$$

$$\begin{aligned} \frac{\partial^2 \pi_c^b(Q^b(k^b))}{\partial (k^b)^2} &= -C_m \lambda f'(k^b) + C_g t_g' \left(1 + \Psi_g(k^b)\right) \lambda - C_e V(k^b) \left(1 + \Psi_e(k^b)\right) \lambda \\ &\quad - \Psi_g C_g t_g^2 \mu_g \lambda / 2 - C_e \Psi_e t_e^2 f^2(k^b) \lambda / (2T_e), \end{aligned}$$

in which $\Psi_i(k) = [\beta_i^*(Q^b) + (C_w/C_i)\alpha(\beta_i^*(Q^b))] / [2\sqrt{\rho_i}]$. As in Lemma 2, $\lambda > \tilde{\lambda}_*^b$, implies $\pi_c^b(Q^b(k^b))$ is strictly convex.

Finally, the optimal k^b that the client prefers can be obtained by setting the first derivative of $\pi_c^b(Q^b(k^b))$ with respect to k^b to zero. As a result, $k^b = k_*^b$, in which $k_*^b = f^{-1}(r_*^b)$, and

$$r_*^b = \frac{C_m + C_g t_g \left(1 + \Psi_g(k_*^b)\right)}{C_m + C_e t_e \left(1 + \Psi_e(k_*^b)\right)}.$$

Therefore, the client offers a contract $(Q^b, P^b) = (Q_*^b, P_*^b)$, in which

$$\begin{aligned} Q_*^b &= \left[C_e t_e \left[1 + \Theta_e(k_*^b, Q_*^b)\right] r_*^b - C_g t_g \left[1 + \Theta_g(k_*^b, Q_*^b)\right] \right] / [t_g - r_*^b t_e], \text{ and} \\ P_*^b &= Q_*^b t(k_*^b) - \left[C_e \rho_e \left[1 + 2\Theta_e(k_*^b, Q_*^b)\right] + C_g \rho_g \left[1 + 2\Theta_g(k_*^b, Q_*^b)\right] \right] / \lambda. \end{aligned}$$

■

Proposition 5 Proof. The optimal system-time penalty Q_*^b can be obtained by setting the first derivative of the cost function $\pi_c^b(Q^b)$ equal to zero, given that the unit waiting cost is C_w . By the implicit theorem,

$$\frac{dQ_*^b}{dC_w} = - \frac{\partial^2 \pi_c^b(Q^b) / \partial Q^b \partial C_w \Big|_{Q^b=Q_*^b}}{\partial^2 \pi_c^b(Q^b) / \partial (Q^b)^2 \Big|_{Q^b=Q_*^b}}.$$

Because $\pi_c^b(Q^b)$ is convex in Q^b , the sign of dQ_*^b/dC_w is the sign of $-\partial^2 \pi_c^b(Q^b) / \partial Q^b \partial C_w \Big|_{Q^b=Q_*^b}$,

$$\begin{aligned} - \frac{\partial^2 \pi_c^b(Q^b)}{\partial Q^b \partial C_w} \Big|_{Q^b=Q_*^b} &= \frac{C_g}{Q_*^b} \frac{\partial \beta_g^*(Q^b)}{\partial Q^b} \Big|_{Q^b=Q_*^b} \sqrt{\rho_g(k_*^b)} + \frac{C_e}{Q_*^b} \frac{\partial \beta_e^*(Q^b)}{\partial Q^b} \Big|_{Q^b=Q_*^b} \sqrt{\rho_e(k_*^b)} \\ &\quad - \frac{\alpha(\beta_g^*(Q_*^b))}{2\sqrt{\rho_g(k_*^b)}} (\lambda t_g) \frac{\partial k^b}{\partial Q^b} \Big|_{Q^b=Q_*^b} + \frac{\alpha(\beta_e^*(Q_*^b))}{2\sqrt{\rho_e(k_*^b)}} (\lambda t_e f(k_*^b)) \frac{\partial k^b}{\partial Q^b} \Big|_{Q^b=Q_*^b}. \end{aligned}$$

By using Lemma 1 and Lemma 5, we see that $-\partial^2 \pi_c^b(Q^b) / \partial Q^b \partial C_w \Big|_{Q^b=Q_*^b} > 0$ and thus $dQ_*^b/dC_w > 0$. The second part of this proposition can be proved by using the chain rule and Lemma 5. ■

Proposition 6 Proof. The proof is similar to Proposition 5, except that

$$-\frac{\partial^2 \pi_c^b(Q^b)}{\partial Q^b \partial C_m} \Big|_{Q^b=Q_*^b} = -\frac{\partial k^b}{\partial Q^b} \left(1 - f(k^b)\right) > 0.$$

As a result, $dQ_*^b/dC_m > 0$. Furthermore, because $\partial k^b/\partial Q^b < 0$ (from Lemma 5), we also have $\partial k_*^b/\partial C_m < 0$ by the chain rule. ■

Lemma 6 Proof. To simplify the notation, we suppress the dependence of $P^1(Q)$ and $P^b(Q)$ on Q in the proof. First, a contract (Q, P) in which $P < \min(P^1, P^b)$ is not feasible because it violates the vendor's IR constraints regardless of his process choice. Next, for a given Q , the staffing and treatment threshold decisions (for a two-level process) are known; the value of P will not affect the vendor's decision variables. Therefore, for a given Q , the client will reduce P but still keep it feasible for the vendor, i.e., $P \in [\min(P^1, P^b), \max(P^1, P^b)]$.

Next, we can divide the contract space into three regimes: (1) $P^1 < P^b$, (2) $P^1 > P^b$, or (3) $P^1 = P^b$. In the first regime, the vendor will choose the one-level process over two. If the vendor chooses a two-level process, he earns zero profit if $P = P^b$, or earns a negative profit if $P < P^b$. However, if he deviates to a one-level process, he earns a positive profit if $P^1 < P \leq P^b$ and earns a zero profit if $P = P^1$. Anticipating the vendor's choice, the client will not offer any extra pay-per-service payment to the vendor. She will only offer $P = P^1$ in this case. In the second regime, the vendor will choose a two-level process, and hence the client offers $P = P^b$ to the vendor. In the third regime, the only contract the client can offer is $P = P^1 = P^b$. ■

Proposition 7 Proof. The proof refers to Figure B1, illustrating the contract space with only two intersections of the vendor's two rationality constraints plotted. It is not difficult to prove that $P^1(Q)$ and $P^b(Q)$ are strictly increasing and strictly concave with Q . We divide the contract space into two regimes, (1) $P^b(Q) < P^1(Q)$ (areas I and III in Figure B1) and (2) $P^1(Q) < P^b(Q)$ (area II). When the client offers contracts on the intersections (where $P^1(Q) = P^b(Q)$), the vendor is indifferent between a one-level and a two-level system.

When the client prefers S_b : If the optimal contract with process monitoring, (Q_*^b, P_*^b) , is in regime 1, or on one of the intersections (i.e., $P_*^b \leq P^1(Q_*^b)$), the vendor will conform to the client's choice. However, if (Q_*^b, P_*^b) is in regime 2, the vendor will deviate to a one-level process. Therefore to induce the vendor to choose a two-level process, the client must offer contracts in areas I or III. By Lemma 6 the contracts in regime 1 satisfy $P = P^b(Q)$ and the client's cost function $\pi_c^b(Q, P^b(Q))$ is equivalent to the cost function $\pi_c^b(Q^b)$. Following the proof of Proposition 4, we know that $\pi_c^b(Q, P^b(Q))$ is strictly convex with Q , and hence the further the contract is away from (Q_*^b, P_*^b) , the higher the client's cost (the line marked as $\pi_c^b(Q, P^b)$ above the contract space

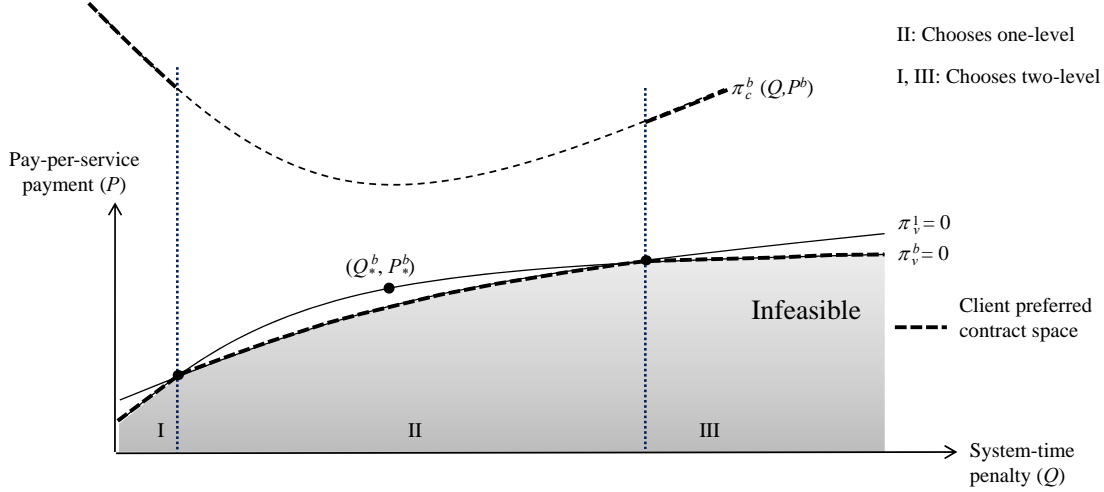


Figure B 1: The contract space and the corresponding IR constraints when the vendor chooses a one-level or a two-level process, i.e., the lines with π_v^1 and π_v^b on the right.

in Figure B1). For Q_*^b , we can find the two adjacent intersections such that $Q_L < Q_*^b < Q_{L+1}$. The convexity of $\pi_c^b(Q, P^b(Q))$ implies that for any $Q > Q_{L+1}$, the client's cost is no lower than $\pi_c^b(Q_{L+1}, P_{L+1})$, and for any $Q < Q_L$, the client's cost is no lower than $\pi_c^b(Q_L, P_L)$. Therefore, the client will only need to choose one from the two nearest intersections as the optimal and incentive compatible contract, i.e., (Q_L, P_L) or (Q_{L+1}, P_{L+1}) , depending on which one yields a smaller client's cost.

When the client prefers a one-level system: If (Q^1, P^1) is in regime 2 or on one of the intersections, then the vendor will conform to the client's choice. If (Q^1, P^1) is in regime 1, then the client will choose one of the contracts in regime 2 that satisfy $P = P^1(Q)$. Given $P = P^1(Q)$, after some simple algebra, the client's cost function when the vendor chooses a one-level system is

$$\pi_c^1(Q, P^1(Q)) = C_e \rho_e(0) \left(1 + (\beta^*(Q) + (C_w/C_e) \alpha(\beta^*(Q))) / (2\sqrt{\rho_e(0)}) \right).$$

Because the first derivative of $\pi_c^1(Q, P^1(Q))$ is

$$\frac{d\beta^*(Q)}{dQ} \left(1 - \frac{C_w}{Q} \right) \frac{C_e \sqrt{\rho_e(0)}}{2},$$

we know that when $Q < C_w = Q^1$, $\pi_c^1(Q, P^1)$ decreases with Q , while when $Q > C_w = Q^1$, $\pi_c^1(Q, P^1)$ increases with Q , in which Lemma 1 shows that $d\beta^*(Q)/dQ > 0$. Following the same logic as for Q^b , we know that the client will only choose one of the two nearest intersections. ■