Managing the Outsourcing of Two-Level Service Processes:
Literature Review and Integration

Edieal Pinker
University of Rochester
pinker@simon.rochester.edu

Robert Shumsky
Dartmouth College
Robert.A.Shumsky@dartmouth.edu

Hsiao-Hui Lee
University of Rochester
Hsiao-Hui.lee@simon.rochester.edu

Sameer Hasija
INSEAD
sameer.hasija@insead.edu

Abstract
The outsourcing of service processes presents several complex management challenges. Services typically have variability in both the pattern of customer arrivals and service times, which make capacity planning challenging. Workers also have discretion over the workflow, and therefore each worker can affect the workload passed to other parts of the system. When a service process is outsourced, this discretion may be in the hands of a vendor who then makes choices that influence the customer experience. A firm must decide upon the design of the process, which parts of the process to outsource, and how to contract with the vendor to overcome issues related to information asymmetry. This paper addresses these questions within the context of a two-level service process where the first level serves as a gatekeeper for experts in the second level. We integrate the results from several papers in the literature to give a comprehensive perspective on how to approach service outsourcing.

1. Introduction

There are many service processes that are organized in two levels. The first level performs triage to determine which cases may be treated at the first level, and which require attention at the second level. Typically the first level is staffed by cheaper, lower-skilled workers while the second level is staffed by more expert workers who command higher wages. Examples of such systems include technical support centers, medical services, and the credit application divisions of banks. In all of these systems some customer cases can be successfully handled at the first level while others cannot. We refer to the first-level workers as gatekeepers because they control the flow of customer requests to workers at the second level, who we refer to as “experts”.

In this paper we will discuss the management of such two-level service processes. We choose to focus on such systems because a) they are common in practice, b) they are simple enough to be amenable to detailed analysis, and c) worker decisions affect the flow of work. Therefore within the context of such systems we can address operational questions such as how the process should be designed and staffed. We can also address economic questions, such as what incentives should be given to the workers and how the outsourcing of the process can be managed via contract design. Our goal is to improve our understanding of the interplay between these different factors in the context of service process outsourcing.

To this end, we summarize and synthesize results from a variety of papers. Each particular paper in this area describes some variant of these service systems. One paper analyzes one-level systems, while others analyze two-level systems. Some examine centrally controlled systems, while others examine outsourcing and contracting decisions. One paper describes a deterministic model while others include stochastic queueing phenomena. In this paper we gather together these results and integrate their answers to general questions, e.g., when should one or two-level systems be used? If the one-level or two-level system is outsourced, does an optimal contract exist, and what form might it take?

Figure 1: Schematic of Workflow

Despite the wide variety of potential applications, throughout this paper we will use language derived from the health care setting. We will 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. Figure 1 provides a schematic of the system. It shows that all customers are initially diagnosed by the gatekeeper, with a fraction $k$ receiving treatment from the gatekeeper as well. Of those receiving treatment from the gatekeeper, a fraction $F(k)$ are successfully treated and thus leave the system immediately without seeing an expert. A fraction $1-k$ of the customers are directly referred to the experts after diagnosis by the gatekeeper, while a fraction $k-F(k)$ are “mistreated” by the gatekeeper and ultimately end up at the expert. To simplify our analysis, we assume that experts can resolve all customer requests.

A manager of such a system has a number of decisions to make. A very fundamental decision is: should this process be operated as a two-level process at all? It is possible that a one-level system staffed only with experts is more efficient. For both one and two-level systems the manager must decide how many workers to staff. For a two-level system the manager must also decide which cases should be treated by the gatekeeper and which ones should be directly referred to the experts. A higher-level decision relates to outsourcing. Should the manager outsource all or some part of the process? Finally, from the manager’s perspective there are potentially two kinds of agents in this system: the workers, and the vendor to which work is outsourced. Both agents may have information about their actions and capabilities that are difficult for the manager to observe. This information asymmetry creates an agency problem that must be managed as well. Therefore the manager must construct an incentive system to ensure that the agents are acting in his interests.

As part of our analysis we will also investigate how these decisions are influenced by the presence of stochasticity in the system. Service processes often involve large amounts of uncertainty that is driven by the customer’s role in the process. Because services require co-production between a firm and its customers, randomness in the arrival process of customer requests cannot be managed through the use of inventory. Further, customer involvement makes each provision of the service a somewhat customized activity, which introduces uncertainty into the service time.

Our goal is to provide a set of contracting guidelines for managers who are outsourcing two-level service processes, by integrating results of studies of various aspects of service process management and outsourcing. We intend our analysis to also contribute more broadly to the understanding of what factors managers of general service processes need to consider when outsourcing, as well as to the understanding of the tools available for analyzing such systems. We draw upon several streams of research which we summarize in Section 2. In Section 3 we describe a deterministic model of the two-level system managed internally, which forms the basis of the paper. In Section 4 we consider the stochastic version of the system and illustrate the new issues raised when congestion and queueing become factors in the analysis. In Section 5 we describe models for outsourcing of both deterministic and stochastic systems and present optimal contracting schemes. In Section 6 we present results on outsourcing with information asymmetry about agent and vendor capabilities. We conclude in Section 7.

2. Summary of Related Literature

To address the issues raised above we draw upon several different streams of research: business process reengineering and design, principal-agent models, models of stochastic service systems, and outsourcing. One of the most common business process reengineering (BPR) techniques is to replace a process in which functional specialists perform sequential tasks with a system that triages (efficiently routes) customers among workers. In a typical triaging system, customers first interact with a generalist who determines if the customer requires the attention of a specialist. Case studies described in [14] indicate that this approach can have enormous efficiency benefits. In a series of papers using queueing analysis, [28], [29], and [30] investigate the factors that influence the benefits of triaging. [5] also uses queueing models to investigate the conditions under which various reengineering strategies, such as the use of parallel caseworkers rather than serial processing, improve process performance. All of these papers focus on how structural choices affect system performance. This stream of research, however, does not examine the interactions between the incentives and the behaviors of individual workers.

More precisely, in all of these papers, for a given system configuration the rate at which work is routed among work stations is determined exogenously. We believe that the ultimate success of any attempt to reengineer a system hinges on the degree to which the employees adhere to the process goals and protocols. In our triaging system, the gatekeepers’ decisions regarding when to refer a case to an expert are critical to the process performance. As in the classic principal-agent model, the gatekeeper’s preferences may differ from the principal’s, and there may be information asymmetry between the principal and the agent (in our case the gatekeeper—but not the firm— may see the
details of each customer's problem and the suitability of the gatekeeper's skills for that customer).

There is a rich literature in economics on situations characterized by such information asymmetry: the research on the principal-agent model as well as related work on contract (or, more generally, mechanism) design. This stream of research includes the dynamic principal-agent model in [17], and the multitask principal-agent models in [18] and [20]. These papers usually begin with extremely general models of the relationship between worker actions and output. A standard model is described in [17], in which the worker (agent) chooses an action $p$, and the impact of that action is fully described by its affect on the probabilities of outcomes $\theta_1, \ldots, \theta_n$. In this paper, however, we focus on models that describe the operational details of actions and outcomes. Therefore, our models are less general — they only apply to gatekeeper systems — but they provide deeper insights into the challenges of managing those systems, as well as clearer recommendations for operating decisions (staffing, routing) and economic decisions (contract design).

In particular, the setting here is distinguished from previous work in both agency theory and personnel economics in that accounting for how an agent allocates her time is important. For example, if a gatekeeper decides to treat a particular customer to obtain a reward, then the gatekeeper is sacrificing the ability to treat any other customer during that period of time (i.e., “time is money”). [31] have developed a model that takes these factors into account to derive insights about the limits of compensation schemes that are tied to time-intensive investments by the agents.

Within the economics literature, both [10] and [25] do deal explicitly with gatekeepers and referral processes. Again, the modeling of time in [31] is a distinguishing feature of the work. The model in [10] describes how low-skilled and high-skilled agents may design contracts to share an income-producing “opportunity” obtained by the low-skilled agent. [25] describes a service system in which a gatekeeper must decide whether to make a referral that is costly to the firm, but the gatekeeper does not have the option to offer any treatment herself. They focus on the inefficiencies created when a gatekeeper is allocated a budget that expires after a fixed time period. The distinction between [31] and these other studies, in terms of how the agent's time is modeled, is an important one. In [31] the decisions made by the gatekeeper have implications for how they spend their time, which affects utilization, which then affects staffing decisions. The staffing decision itself becomes even more important when one considers a stochastic model, where the manager must balance staffing costs and the implicit costs of customer waiting time. There has been some related work on incentives in stochastic service settings. [11] use a principal-agent framework to compare a service network with a common queue to a service network with separate queues. In their model, a central, coordinating agency makes routing decisions. In our case we are interested in how the agents themselves make make routing decisions. [22] also formulate a model in which agents direct customers to service centers. In their model, the agents (or gatekeepers) do not have the option to treat the customer, and each agent seeks to minimize her customers’ expected waiting times at the service centers. [12] examine how incentives in a call center can affect the quantity of service offered to different classes of customers, focusing on how incentives interact with the cost of congestion and the customer segmentation decision. [23] is closely related to [31] in that the principal designs incentives to induce agent effort that affects workflow. [23], however, is based on a manufacturing setting in which defective jobs may be routed back to the worker who caused the defect, to a rework specialist, or to another worker who may be working on new jobs. Therefore, in [23], one worker’s actions may influence both the flow of jobs as well as the content of jobs seen by other workers. While not explicitly modeling a principal-agent setting and incentives [19] also consider situations in which workers have discretion about how they allocate their time in white-collar work.

[15] and [21] extend the deterministic model of [31] to a stochastic setting. With stochastic customer arrivals and service times it is necessary to model the system as a queuing system and to take into account customer waiting costs and staffing levels. These works make use of asymptotic queuing approximations developed in [13] and the square root staffing rules developed in [4]. They show that essentially the same gatekeeper incentives developed for the deterministic gatekeeper system modeled in [31] can be applied effectively in the stochastic setting. These results give a foundation upon which to consider outsourcing all or part of the service process.

While considerable attention has been given to outsourcing contracts in manufacturing supply chains (see [6] and the references therein), the literature on outsourcing contracts for service supply chains is more limited. The work in this area has focused on call-center outsourcing because that environment is most amenable to analysis using tools from queueing theory. In addition, call centers have well-defined performance metrics. See [8] for a survey of call-center research.

[9] and [1] consider a client who can outsource some fraction of service calls to a vendor. [9] study the
centralized capacity decision and queuing control problem. [1] compare the equilibrium performance of service systems in which the client either outsources a steady stream of calls or outsources peak demand. These studies view outsourcing as a mechanism for handling variability in service demand. Essentially they are asking which portion of the workload should be done in-house versus outsourced. Our attention here is on which portion of the process should be outsourced. We are also interested in determining the incentive contract that will make the arrangement work best.

[2] 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 service supply chain coordination; we will also consider similar contracts. [26] 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 essentially ignores queueing effects. [16] also model a client that outsources a call-center to a vendor, and use a stochastic model that preserves queueing effects. They consider both the case in which the firm knows the service rate of the vendor and the case in which it does not. With full information, the client must design a contract that will encourage the vendor to staff sufficiently. With unknown service rates the firm must structure the contract to both encourage the vendor to work as fast as he can and to staff appropriately.

Studying quality, beyond waiting time, in service outsourcing is challenging because performance measurements are poorly defined. An exception is [3] who 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.

3. Deterministic Model

In this Section we summarize the analysis of the deterministic gatekeeper model from [31]. This model is the basic building block of the analysis in subsequent sections. We assume that each gatekeeper can rank-order customer jobs in terms of increasing complexity. Let \( k \in [0,1] \) be a fractile of call volume, ranked by treatment complexity, that is, \( k \) denotes a position in the ranking of calls such that \( k \times 100\% \) of calls are less complex. We define the gatekeeper’s treatment function as \( f(k) \) denoting the probability that the gatekeeper can successfully treat the problem during a treatment time with expected duration \( t \), given that the customer’s problem is at the \( k \)th fractile of difficulty (assuming, of course, that the gatekeeper makes an effort to provide treatment to the customer). In our model, the complexity ordering of tasks may vary from gatekeeper to gatekeeper (e.g., at a technical support desk, modem glitches may present difficulties for one gatekeeper while problems with e-mail may be difficult for another). However, here we assume that the skill level for each level of complexity, defined by \( f(k) \) is the same for all gatekeepers. We also assume to simplify the analysis that the expected treatment time \( t \) does not vary with call complexity \( k \).

From the definition of \( k \), it follows that \( 0 \leq k \leq 1 \), \( 0 \leq f(k) \leq 1 \), and \( f'(k) < 0 \) (the last inequality stems from the ordering of calls by treatment difficulty). Let \( F(k) \) be the expected fraction of all calls that are successfully treated by the gatekeeper, given that the gatekeeper chooses to treat all calls up to and including the \( k \)th fractile, so that \( F(k) = \int_0^k f(x)dx \) and \( 0 \leq F(k) \leq k \). We define the following cost parameters:

\[ C_g \] = gatekeeper wages ($/time period).
\[ C_e \] = expert wages ($/time period).
\[ C_m \] = expected cost of incorrect treatment by a gatekeeper ($/customer). This includes any “goodwill” cost as well as the expected cost of customer reentry.

We are assuming here that it is difficult for a manager to observe the difficulty of each case handled by a gatekeeper and therefore know the gatekeeper’s likelihood of successfully treating the case. Without knowing this likelihood the manager cannot say if the case should be treated by the gatekeeper or referred directly to an expert. This information asymmetry between the gatekeeper and the manager means that the manager would like to construct an incentive system that compels the gatekeeper to refer in a system-optimal way.

We assume that customers arrive according to a deterministic process and that the diagnosis and treatment times are deterministic. The gatekeeper’s parameters are: \( d \) for gatekeeper diagnosis time, \( t_g \) for gatekeeper treatment time and \( t_e \) for expert treatment time. The firm’s problem is to minimize the labor and mistreatment costs per case. Staffing levels are fully
4. Stochastic Models

Staffing any less would lead to an overloaded system determined by the gatekeeper treatment decision.

and staffing more would not add any value.

The firm's objective function is:

\[ \text{Min}_{n_g, n_e} C_g n_g + C_e n_e + C_w \left( W_g \lambda_g + W_e \lambda_e(k) \right) + C_m (k - F(k)) \lambda. \]  \hspace{1cm} (3)

Note that the waiting times at the gatekeepers and experts are dependent upon the treatment threshold \( k \) and the staffing levels. As shown in Figure 1, given \( k^* \) at the gatekeeper level, the service rate of the gatekeeper is determined, and the rate of flow from gatekeepers to experts is also determined. To derive expressions for the waiting time at each level that will enable us to determine optimal staffing levels, we make a few simplifying approximations.

First we assume, as in standard in such models, that the customer arrival process to the gatekeepers is Poisson. Second, we assume that the service processes at both the gatekeeper and expert levels are exponential. The gatekeepers' service rate is \( \mu_g(k) \) while the experts' is \( \mu_e \). For the gatekeepers this approximation is more significant because their service times are a mixture of diagnoses and treatments. Our final approximation is that the arrival process to the expert queue is Poisson as well.

The above approximations imply that for a fixed threshold \( k \) the two sub-systems (gatekeeper and expert) can be analyzed independently as M/M/N queueing systems for the purpose of determining the staffing levels that optimally balance waiting and staffing costs. In particular, it means that if we assume we are operating in the QED regime of [13] we can use square-root staffing rules as in [4] to determine optimal staffing. In addition, in the QED regime there exist closed-form expressions for expected waits.

Therefore, we have made two sets of approximations: that each queue is an M/M/N system, and that each operates in the QED regime. Experiments in [4], [13], [15], and [21] show that both of these approximations are extremely accurate. For example, experiments in [21] demonstrate that over a wide variety of parameters, the mean error for the total time in system is less than 1%, and for large systems (over 250 servers), the maximum error is 3.5% or less.

4.1. Centralized System

For a firm that has complete control over the staffing and treatment threshold the objective function is,

\[ \text{Min}_{n_g, n_e, k} C_g n_g + C_e n_e + C_w \left( W_g \lambda_g + W_e \lambda_e(k) \right) + C_m (k - F(k)) \lambda. \]  \hspace{1cm} (3)

In a system with stochastic arrivals and service times, new dynamics enter the analysis. First, stochastic service systems (queues) have congestion and thus we are concerned with customer waiting. Second, staffing decisions become important because of the inherent trade-off between staffing and waiting costs. Third, from a process design perspective, multiple levels introduce additional delays thus changing the calculus of the decision to have one or two levels. Fourth, economies of scale can be important in stochastic systems because they enable the pooling of variability. This will also influence the choice of one or two-level systems and how work is divided between the two levels. For the stochastic system we define the following additional notation:

\( \lambda \) customer arrival rate to the system \\
\( \lambda_e \) arrival rate to the experts (a function of \( k \)) \\
\( n_g, n_e \) the staffing at the gatekeeper and expert levels, respectively \\
\( W_g, W_e \) the expected waiting time at the gatekeeper and expert levels, respectively \\
\( C_w \) cost of waiting, per unit time

The above approximations imply that for a fixed threshold \( k \) the two sub-systems (gatekeeper and expert) can be analyzed independently as M/M/N queueing systems for the purpose of determining the staffing levels that optimally balance waiting and staffing costs. In particular, it means that if we assume we are operating in the QED regime of [13] we can use square-root staffing rules as in [4] to determine optimal staffing. In addition, in the QED regime there exist closed-form expressions for expected waits.

Therefore, we have made two sets of approximations: that each queue is an M/M/N system, and that each operates in the QED regime. Experiments in [4], [13], [15], and [21] show that both of these approximations are extremely accurate. For example, experiments in [21] demonstrate that over a wide variety of parameters, the mean error for the total time in system is less than 1%, and for large systems (over 250 servers), the maximum error is 3.5% or less.
In addition, the largest errors are not caused directly by the approximations, but rather by the rounding of staffing levels (the QED approximation and square-root staffing rule lead to a continuous staffing level, while the test simulation in [21] requires integer staffing levels). Likewise, experiments in [15] show that the error in the total cost of the system due to the approximation is consistently less than 1% and that staffing decisions made under the approximation are consistently identical to, or within 1 server, of the exact optimal solution.

4.2. Square-Root Staffing Approximation

The square root staffing rule follows [4]. Consider a single-queue system in the QED regime with an arrival rate of $\lambda$, a service rate of $\mu$, and $n_i$ servers. The optimal staffing level is:

$$n_i = \rho_i + \beta_i \sqrt{\rho_i},$$

where $\beta_i$ can be seen as the standard excess capacity in order to manage the system variability and $\rho_i = \lambda/\mu_i$. We can always find a $\beta_i$ that meets the desired service requirement. In this paper, we quantify the service requirement by the waiting time, and the goal is to minimize the expected total cost $C_{ini} + C_iW_i$, in which $C_w$ is the waiting cost per unit time, $C_i$ is the unit cost per staff in system $i$, and $W_i$ is the mean waiting time in system $i$. For an M/M/N queue, the expected waiting time $W_i$ can be represented as a function of $\beta_i$, $\lambda$, and $\mu_i$ ([13]):

$$W_i = \frac{\alpha(\beta_i)}{\sqrt{\lambda \rho_i}},$$

where $\alpha(\beta) = [\beta + \beta^2 \Phi(\beta)/\phi(\beta)]^{-1}$ and $\Phi(\beta)$ and $\phi(\beta)$ are the CDF and PDF of a standard normal distribution. Therefore $\beta_i$ can be written as:

$$\beta_i = \arg \min \hat{C}(\beta_i),$$

where

$$\hat{C}(\beta_i) = C_i \beta_i + C_i \alpha(\beta_i).$$

For a centralized decision maker using square root staffing rules optimally managing the two-level service system is reduced to a one variable problem: choosing the gatekeeper treatment threshold $k$.

4.3. Optimal Treatment Threshold

The optimal treatment threshold is difficult to calculate for a stochastic system. However [15] have found that for a linear treatment function the deterministic treatment threshold is very close to optimal. The cases in which it is significantly differs from the optimal are those cases in which it is optimal to operate a one-level system of experts anyway. [21] show that for a large class of treatment functions the optimal treatment threshold asymptotically approaches the deterministic one for large systems. In particular the optimal threshold is described by:

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

Therefore, $k^*$ is calculated recursively, and because $f$ is continuous, from the fixed-point theorem we know that a solution $k^*$ exists.

We also find that functions $\Theta_g$ and $\Theta_e$ asymptotically approach 0 as the system size increases. Taken together these results suggest that uncertainty in arrival and service processes are not important for determining the optimal treatment threshold for a two-level service process and thus the guidelines in [31] are robust. It also means that the incentive scheme of pay per diagnosis and per solved case (see Section 5.3) can be used in a variety of environments and can be derived just using information about average service times, treatment functions, and pay rates for experts and gatekeepers.

4.4. One-Level vs. Two-Level Systems

Analysis and numerical experiments in [15] and [21] describe the environments in which one-level or two-level systems are preferred. In the health care setting, a two-level system is sometimes called a direct access system because patients bypass gatekeepers such as nurses or primary care physicians and instead move directly to specialists. Patients often prefer such systems, so it is important to determine how direct access systems compare with gatekeeper systems in terms of overall cost.

It is not surprising that we find that one-level systems tend to be preferred when expert staffing costs ($C_e$) and gatekeeper productivity ($\mu_g$) are low. Likewise, one-level systems are favored when gatekeeper staffing costs ($C_g$), mistreatment costs ($C_m$) and expert productivity ($\mu_e$) are high. Another important driver of the system preference is the skill level of the gatekeepers. When the function $f(k)$ is relatively high (close to 1) over a wide range of $k$, the two-level system is preferred.

While the effects of these parameters on the system choice are straightforward, the impact of the cost of waiting, $C_w$, is more subtle. We identify instances in which a two-level system is preferred when waiting costs are low, and where a one-level system becomes more appealing as waiting costs rise. The transition from a two-level to a one-level system as waiting costs rise is due to the queueing economies of scale and lower waiting times that may be achieved in a one-level system. However, we also find combinations of parameters for which rising waiting costs do not lead to the superiority of a one-level system. In particular, if the gatekeepers have sufficiently high skills ($f(k)$ sufficiently high), then as
$c_r$ rises, the optimal treatment threshold rises, and a one-level system is never preferred. In that case, the economies of scale are achieved at the gatekeeper level, rather than in an all-expert system.

Finally, note that these results comparing one-level and two-level systems apply whether or not the system is outsourced.

5. Outsourcing the Service System

When outsourcing a service process a firm faces challenges that are different than when outsourcing a manufacturing process. Figure 3a depicts a schematic of traditional manufacturing outsourcing in which a client firm contracts with a vendor to provide some intermediate production input. In this setting the customer pays the firm and interacts directly with the vendor. The firm pays the vendor and can perform quality control before any product reaches the customer. From the customer’s perspective the vendor is invisible. In the service setting (see Figure 3b) the situation is quite different. Consider the case of a computer maker who outsources technical support. In this case the customer pays the client firm but interacts directly with the vendor when receiving technical support. The firm wants the customer-service experience to meet some standard but does not have direct control over it. Also, monitoring the vendor’s staffing, procedures, and interactions with customers can be very costly.

![Figure 3a: Manufacturing](image1)

![Figure 3b: Service](image2)

5.1. Outsourcing a Deterministic System

If the firm outsources the process to a vendor it has three options: outsource both levels (or a one-level system if that is optimal), outsource the expert, or outsource the gatekeeper. We find that for a deterministic system, in all three cases it is straightforward to design an optimal (first-best) contract. If the firm outsources both levels then it may appear as if the firm loses visibility about the gatekeeper’s treatment decisions, and therefore cannot enforce a contract that demands optimal referral rates. However, in a deterministic system, a particular treatment threshold uniquely determines the average total system time of a customer request. Therefore the client firm can contract on the system time. If the firm outsources the expert only, then it can pay the vendor for its staffing costs, the vendor does not make any decision besides staffing levels and it is easily observable if the vendor has sufficient staff. If the firm outsources the gatekeeper then it can pay the vendor the same as if they were paying its own workers to make the correct referral decisions.

5.2. Outsourcing the Experts or a One-Level Stochastic System

When outsourcing only the expert level the firm’s objective function is,

$$\begin{align*}
\min_{n_e, k, T_e} C_g n_e + C_w [W_g \hat{\lambda} + W_e \hat{\lambda}_c(k)] \\
+ C_m (k - F(k)) \hat{\lambda} + T_e,
\end{align*}$$

where $T_e$ is the payment received by the vendor under the outsourcing agreement. $T_e$ of course, may depend upon system performance, and the specific structure of $T_e$ will be discussed below. Note that because the firm operates the gatekeeper level it controls the flow of customers to the vendor.

If the firm knows the service rate of the vendor’s experts then this situation is equivalent to the situation analyzed in [16], and the same contracts can be used to achieve the first-best outcome. An example contract structure described in [16] would pay the vendor per call handled and extract a penalty on the average time customers spend in the system. To enforce such a contract the firm would only have to monitor when customer requests arrived to the vendor and when they left, without observing the inner workings of the vendor’s operations.

As we have discussed earlier, in some cases it may be more efficient to operate the system as a one-level process in which all the workers are experts. In this case there is only one decision for a vendor to make, choosing the number of staff. Outsourcing this system is structurally equivalent to outsourcing only the experts. Therefore the same contract structure is optimal for the firm.

5.3. Outsourcing the Gatekeeper in a Stochastic System

When outsourcing only the gatekeeper level the firm’s objective function is,

$$\begin{align*}
\min_{n_g, k, T_g} C_g n_g + C_w [W_g \hat{\lambda} + W_e \hat{\lambda}_c(k)] \\
+ C_m (k - F(k)) \hat{\lambda} + T_g,
\end{align*}$$

where $T_g$ is the payment received by the vendor under the outsourcing agreement. $T_g$ may depend upon system performance, such as the rate of service and the
Table 1: Optimal outsourcing contracts, assuming no information asymmetry

<table>
<thead>
<tr>
<th>Type of Outsourcing</th>
<th>Vendor’s internal decisions</th>
<th>Terms of each optimal contract</th>
<th>Deterministic System</th>
<th>Stochastic System</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-Level Process or Expert Only</td>
<td>Expert staffing</td>
<td>Pay-per-service</td>
<td>Pay-per-service, system-time penalty</td>
<td></td>
</tr>
<tr>
<td>Gatekeeper Only</td>
<td>Gatekeeper staffing, treatment threshold</td>
<td>Pay-per-service, Pay-per-successful treatment</td>
<td>Pay-per-service, Pay-per-successful treatment, system-time penalty</td>
<td></td>
</tr>
<tr>
<td>Both Gatekeeper and Expert To One Vendor</td>
<td>Gatekeeper and expert staffing, treatment threshold</td>
<td>Pay-per-service, payment contingent on System Time</td>
<td>No optimal contract</td>
<td></td>
</tr>
</tbody>
</table>

Table 1 shows the optimal outsourcing contracts for different types of outsourcing, considering no information asymmetry. Each contract type specifies the vendor’s internal decisions and the terms of the optimal contract, including whether it is deterministic or stochastic.

rate of successful treatment. In this case the vendor chooses how many gatekeepers to staff, \( n_g \), as well as the treatment threshold, \( k \). As a result the vendor will not only determine the customer waiting time at the gatekeeper level but also both the flow of customers to the expert level and the mistreatment rate. The client firm needs to construct an incentive scheme (generating \( T_g \)) that will induce the vendor to choose the system-optimal staffing level and treatment threshold. In other words, the client firm is interested in both the capacity investment of the vendor and how it does its work.

[21] show that it is possible to construct a system-optimal contract that only requires observing how many customers enter the gatekeeper level, how much time they spend at the gatekeeper level, and whether the customer’s problem has been solved when the customer leaves the gatekeepers. The contract scheme draws upon the results of [31] and [16]. The analysis in [21] describes a specific contract that pays per customer, pays per successful treatment, and levies a system time penalty. These contract terms incentivize the vendor to choose both the correct treatment threshold and the correct staffing level.

5.4. Outsourcing Both Levels in a Stochastic System

When the firm outsources the entire process to one vendor, its objective function is,

\[
\min_{T_b} C_w \left[ W_e \lambda + W_r \lambda_e (k) \right] + C_w (k - F(k)) \lambda + T_b,
\]

where \( T_b \) is the payment received by the vendor under the outsourcing agreement. In this case the firm is incurring all the costs related to the service experience i.e. the waiting and mistreatment costs, while not directly controlling any of the operational decisions that determine them. Without any special effort, the firm can only observe the number of customers that enter and exit the process, and their total system time.

[21] show that it is impossible to construct a contract, using just this information, which will lead to a system optimal outcome. By outsourcing the entire process the firm loses knowledge about the inner workings of how the process is performed. It has thus turned the process into a “black box” and this creates contracting inefficiency. Note, however, that if the client outsources the gatekeepers to one vendor, outsources the experts to another vendor, and observes the flow of jobs between then, then the analyses described in Sections 5.2 and 5.3 apply, and optimal contracts are possible. We summarize our analysis of outsourcing in Table 1.

6. Outsourcing with Information Asymmetry on Capabilities

Most service centers have a significant amount of variability in the capabilities of customer service agents (for one example, see the study of the referral rates of primary care physicians by [7]). In the setting of this paper there are two kinds of capabilities that workers/vendors have that may be imprecisely known to the client firm. First the client may not know what rate the vendor’s workers are capable of. This implies that the vendor may not necessarily have an incentive to work as fast or as efficiently as possible. Second, with regard to gatekeepers, the client firm may not know the treatment abilities of the vendor’s workers. This implies that the firm may not know what the correct treatment threshold is for a set of gatekeepers.

The economics literature can give us guidance on how to manage in such situations with information asymmetry. The standard approach is to view this as a screening problem, see, for example, [27] and [24]. The uncertainty in worker or vendor capabilities can be interpreted this way: there is a spectrum of vendor types, and the firm needs to screen the vendors by encouraging the best possible performance from all types. A possible mechanism is to offer a variety, or menu, of contracts that lead vendors to self-select and thereby reveal their types. We use this approach for our problem as well, and we show that by offering variations on pay-per-case, pay-per-time, pay-per-solve and SLA contracts, the client can overcome the
vendor’s relative information advantage and achieve optimal results.

6.1. Service rate screening

[16] analyze how to structure contracts when the firm does not know the service rate capabilities of the vendor and is outsourcing a one-level process. They have shown that when the vendor’s service rate is known, then the client can incentivize the vendor to staff correctly by offering a contract with a combination of pay-per-case (PPC) and waiting time penalties, or a combination of pay-per-time (PPT) and waiting-time penalties. Similarly, replacing the waiting time penalties with a service level agreement (SLA) also works well. When the vendor’s service rate is unknown, for example if the vendor is either fast (high rate) or slow (low rate), the client would like the offered contract to push the vendor to staff correctly, to be sufficient for a slow vendor to participate, and yet induce a fast vendor to work as fast as they can and not slow down. It can be shown that offering a choice between a PPC-based contract and a PPT-based contract will achieve just that. Slow vendors will select the PPT contract while staffing in a system-optimal way, and the fast vendors will choose the PPC contract and also staff in a system-optimal way.

6.2. Gatekeeper treatment skill screening

If the firm did not know the skills of the gatekeeper it would be equivalent to not knowing what the treatment function looked like. As a result the firm would not know what treatment threshold to select and what incentives to offer the gatekeeper or vendor to achieve that threshold. Again screening contracts can be of use in this situation.

[31] consider the case in which the pool of gatekeepers is made up of “low” and “high” skill workers. The treatment function of the high skill gatekeepers is greater than or equal to that of the low skill for every complexity level. They find that there exists a combination of payment for diagnosis and payment for solved case that will induce screening. That is each type of gatekeeper will find it optimal to choose the treatment threshold that the client firm would want them to use given their abilities.

It remains an open question, for future research, as to how to perform screening when both service rate and treatment skill information asymmetry exist in an outsourcing setting. Given that a hierarchical approach to managing such a system, i.e. setting the treatment threshold and then using square-root staffing rules to set staffing levels, work well; it seems likely that the construction of screening contracts can be similarly decomposed.

7. Conclusions

Typically firms outsource because they want to reduce costs. They want to take advantage of the fact that some vendors have lower labor costs and thus can deliver the service more efficiently. There are clearly costs to outsourcing and vendors require a profit margin. If, taking these into account, the firm can still reduce its costs it may be a viable option. One of the potential costs of outsourcing is contracting inefficiency.

Our analysis in this paper shows that for a two-level process contracting inefficiency should not be a concern if the firm is only outsourcing a single part of the process, or a one-level version of the process, to a particular vendor. In these cases, if the vendor has lower labor costs the firm can construct that contracts that incentivize the vendor to behave in a system optimal way. How the costs of operating the system are shared by the client and vendor becomes primarily a function of the negotiating strength of each party.

When the client outsource the entire two-level process to a single vendor, it is not possible to contract efficiently. However, it may still be worthwhile to outsource anyway if the cost advantages offered by the vendor are sufficient. Quantifying the contracting inefficiency cost is an area for future research.

We have based our analysis on a two-level gatekeeper process because it occurs often in practice and because it requires decisions about process design as well as involves worker discretion on workflow. Another area for further research is to explore the applicability of our analysis to other more general process structures.

8. References


