Gatekeepers and Referrals in Services

Robert A. Shumsky • Edieal J. Pinker
William E. Simon Graduate School of Business Administration, University of Rochester, Rochester, New York 14627
shumsky@simon.rochester.edu • pinker@simon.rochester.edu

This paper examines services in which customers encounter a gatekeeper who makes an initial diagnosis of the customer’s problem and then may refer the customer to a specialist. The gatekeeper may also attempt to solve the problem, but the probability of treatment success decreases as the problem’s complexity increases. Given the costs of treatment by the gatekeeper and the specialist, we find the firm’s optimal referral rate from a particular gatekeeper to the specialists. We then consider the principal-agent problem that arises when the gatekeeper, but not the firm, observes the gatekeeper’s treatment ability as well as the complexity of each customer’s problem. We examine the relative benefits of compensation systems designed to overcome the effects of this information asymmetry and show that bonuses based solely on referral rates do not always ensure first-best system performance and that an appropriate bonus based on customer volume may be necessary as well. We also consider the value of such output-based contracts when gatekeepers are heterogeneous in ability, so that two gatekeeper types face different probabilities of treatment success when given the same problem. We show that the firm may achieve first-best performance by either offering two contracts that separate the gatekeeper types or by offering a single contract that coordinates the treatment decisions of both gatekeepers.

(Stochastic Model Applications; Personnel; Principal-Agent Problems; Routing/Triage)

1. Introduction

Many service systems are arranged with a front line of gatekeepers who refer jobs to a stable of experts. In a health maintenance organization (HMO), primary care physicians (PCPs) serve as gatekeepers to medical specialists. In the telephone call center of a mutual fund, customer-support representatives are gatekeepers to skilled financial advisers and traders. In the help center of an office technology manufacturer, generalist agents screen telephone calls for highly trained, highly paid technical experts.

This paper considers both the routing of customers through such systems and the impact of referral-based incentives on how gatekeepers perform that routing. We first examine the referral decision: When should a gatekeeper attempt to help (or “treat”) a customer and when should the customer be passed along to a specialist? Given the relative costs of gatekeepers and specialists, the cost of mistreatment by the gatekeeper, and a model of the gatekeepers’ effectiveness it is possible to identify a referral policy that is optimal for the system as a whole. However, implementing any particular policy can be difficult. Gatekeepers may prefer to refer customers or may prefer to attempt treatment. Gatekeepers may be heterogeneous in their abilities, and the data needed to enforce a particular policy may be prohibitively expensive to collect. Therefore, gatekeepers have attributes and make decisions that may not be visible to managers, and this leads naturally to a principal-agent model with information asymmetry between the gatekeeper and the employer. We use such a model to evaluate incentive systems that are based on referral rates.

The following examples will serve as motivation for our modeling choices:

• The help desk at the University of Cincinnati provides support for computer users. Callers first talk to a Level 1 technician who performs some diagnosis
and, if possible, tells a customer what to do (in this case we say the gatekeeper “treats” the customer’s problem). If this technician cannot solve the problem, the customer is referred to a Level 2 specialist. Because these specialists are paid at least twice as much as a technician, and because specialists often have other duties (such as computer network support), management asks those on Level 1 to “protect” specialists by referring a customer only when all other measures have failed. However, Level 1 technicians can be too aggressive when trying to protect the specialists, wasting their own time, the time of the customer, and perhaps even the time of the specialist for “sometimes the solution a Level 1 tech proposes can ultimately complicate the work for the Level 2” (Tahmincioglu 2001).

- A trade journal for human resource professionals reports on a survey of compensation systems for inbound call centers, finding a wide variety of bonus systems and incentives. One bonus system ties payments to the number of calls answered (regardless of outcome), while another offers a bonus for the resolution of customer complaints, rewarding gatekeepers when they minimize referrals to specialists (Thaler-Carter 1999). In an interview with the authors of this paper, a manager of a help center for high-technology office products described how the latter system is used in his own facility and gave it the nickname “Paying for Solves.”

- A study of PCPs within a single managed care organization in Rochester, New York, examined the rate of referrals, defined as the probability that the patient of a particular doctor would be referred to a specialist during a one-year period. The observed probability varied widely, with a range from 0.01 to 0.69. Physician attributes such as length of tenure and specialty (family physician or internist) had a significant effect on referral rates, even when controlling for patient case mix. For the physicians in this study, no direct financial incentives, or disincentives, were linked to referrals (Franks et al. 2000).

To summarize our results, we identify environments in which an incentive system that offers a flat wage, a bonus for customer volume, and a bonus for successfully treating a customer (“Paying for Solves”) is sufficient to achieve system performance that is equivalent to a system with no information asymmetry—i.e., it achieves “first best.” Such a contract can achieve first best whether there is just one type of gatekeeper or two types with different levels of skill. Given two types of gatekeepers, the firm may also offer two contracts that lead the gatekeepers to self-select. Given either one or two gatekeeper types, we also find that offering only a bonus for treatment, and not for volume, may not be sufficient for first best. For example, when specialists are extremely expensive the firm may prefer that a gatekeeper attempt treatment on a high proportion of customers and may therefore offer a large bonus for “solves.” But if treatment times are long, the gatekeeper will refer difficult patients to specialists because those patients are unlikely to produce a reward.

In §2, we review the related literature. In §3, we describe a generic model of a gatekeeper system that operates in a manner similar to the University of Cincinnati help desk described above. Given relevant costs and a characterization of gatekeeper effectiveness, we find the optimal referral rate from gatekeepers to specialists. In §4, we formulate a principal-agent model to determine the impact of performance-based incentives on gatekeeper behavior, incentives similar to those described in the trade journal, above. The model proposed in §4 assumes that gatekeepers are homogeneous in their abilities. In §5, we see that allowing heterogeneous gatekeepers seems to significantly complicate the problem but that many of the results derived for homogeneous gatekeepers continue to hold. Section 6 contains a discussion of our results and suggestions for further work.

2. Related Literature

This paper is related to four broad areas of economics and operations management: principal-agent models, personnel economics, health-care economics, and business process reengineering. In our model, the gatekeepers serve as agents, who are delegated the choice to treat or refer customers. 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). In some respects, our model is similar to the dynamic principal-agent model in Holmstrom and Milgrom (1987), and the multitask principal-agent models in Holmstrom and Milgrom (1991) and Lal and Srinivasan (1993).

In these traditional models, increased effort by the agent is likely to be beneficial to the firm but is costly to the agent. However, in our paper, the agent faces a decision that is, in some ways, more general than the standard “effort-level” decision. Here, the gatekeeper may or may not prefer to put in “effort” (treat customers), and the firm’s profit is not monotone in the treatment rate. A gatekeeper who provides too little or too much treatment may significantly reduce the firm’s profits. In other ways, however, this model is more specific (or explicit) than the principal-agent models described above. For example, the firm’s cost function is derived from a network model, and this allows us to gain insight into a particular process, the routing of customers through a referral system. This is similar in spirit to Fuloria and Zenios (1998) and Plambeck and Zenios (2000), in which an agent controls the transition probabilities of a system that is evolving over time.

Pollack and Zeckhauser (1996) also describe 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. Gunes and Aksin (2002) 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. Gilbert and Weng (1998) 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 model, the agents themselves make routing decisions. Lee and Cohen (1985) 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.

In the area of personnel economics, Lazear (1986) describes conditions under which a firm chooses a particular compensation method—salary vs. piece rate—while Fama (1991) focuses on the choice of hourly payoff vs. salary. Lazear (1986) and Fama (1991) study how compensation methods play a role in screening high-skilled and low-skilled workers. In this paper, we study the effects of incentives on the decisions of high and low-skilled agents while they work for the firm as well as the role of the compensation method in screening workers.

Another feature of our model that distinguishes it from previous work in both agency theory and personnel economics is its specification of the decision process and reward function. In our specification, the gatekeeper’s problem is dynamic and the passage of time is explicitly modeled. 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 (in our model, “time is money”). By modeling the gatekeeper’s cost function in this way, we derive the gatekeeper’s opportunity cost of time as well as interesting insights about the limits of compensation schemes that are tied to time-intensive investments by the agents. This is also one aspect of our model that distinguishes it from the model of Garicano and Santos (2001), who describe how low-skilled and high-skilled agents may design contracts to share an income-producing “opportunity” obtained by the low-skilled agent.

The PCP is an important application of the concept of a gatekeeper, and the impact of gatekeepers on health care and the effect of financial incentives to influence their behavior has attracted much public attention. Ferris et al. (2001) challenge the importance of gatekeepers in health care in a study of a large, multispecialty group practice before and after the elimination of a system that required patients to receive authorization from a PCP before being
allowed to visit a specialist. They find that eliminating gatekeeping led to almost no changes in the utilization of specialists. Lawrence (2001) points out the difficulty in generalizing from the study by Ferris et al. (2001) because it reports on an environment in which PCPs and specialists often collaborate, “Financial incentives are shared by the group as a whole” (p. 1342), and the PCP is not really a gatekeeper but rather a trusted partner for the patient in navigating the health-care system. In our paper, we model a situation in which the customer has no preexisting relationship of this kind with the gatekeeper and therefore would probably seek a specialist more frequently if not screened by a gatekeeper first. Gaynor et al. (2001) empirically demonstrate the influence of financial incentives to PCPs on the costs of an HMO and provide evidence that, in health care, incentives for gatekeepers do influence resource utilization by patients.

Another related research area is business process reengineering (BPR). One of the most common BPR techniques is to replace a process in which functional specialists perform sequential tasks by a triaging system. In a triaging system, customers first interact with a generalist who determines if the customer requires the attention of a specialist. This approach has been known to have enormous efficiency benefits (Hammer and Champy 1993). In a series of papers, Seidmann and Sundararajan (1997a, b, c) investigate the factors that influence the benefits of triaging using queueing analysis. Buzacott (1996) uses queueing models to investigate the conditions under which various reengineering strategies improve process performance. In all of these papers, the rate at which work is routed from generalists to specialists is an exogenous parameter. We believe that the ultimate success of any such reengineering strategy hinges on the degree to which the employees adhere to the process goals and protocols. In a triaging system, the generalists’ decisions regarding when to refer a case to a specialist are critical to the process performance.

3. The Model and the First-Best Solution

We consider a stylized model of a two-level service system in which the first level serves as a gatekeeper for the second level. Figure 1 displays the flow of customers through the system. A customer arrives to a gatekeeper who diagnoses the problem and then may choose to treat the patient. If the gatekeeper treats the customer and the treatment is successful then the customer leaves the system. If the gatekeeper chooses not to treat the customer, or if the gatekeeper treats and the treatment fails, then the customer visits a specialist. We assume that a customer visiting a specialist will (eventually) be treated successfully and will then leave the system.

Figure 2 displays this process from the gatekeeper’s point of view. The end of service with one customer may be followed by some period of idle time. Then, when a customer arrives, diagnosis begins immediately. This may be followed by treatment for the customer, or a wait for the next customer. Define $d$ as the expected time between the end of service of one call and the end of diagnosis of the next call (the gatekeeper wait plus diagnosis duration) and define $t$ as the expected duration of treatment.

Before describing the model’s cost structure, we discuss a few of the assumptions embedded in the model. First, there is no explicit queueing, for we do not directly model the effects of customer congestion on the system. Instead, we focus on the average flows between gatekeepers and specialists, and we assume that the firm maintains a level of staffing sufficient to
satisfy exogenous service (waiting-time) goals. When planning staffing needs, a firm must first determine referral rates and gatekeeper treatment workloads, as we do in this paper, and then a higher-level staffing model uses this information. Changes in the exogenous service goals also influence the parameters of the model in this paper. For example, if the service center maintains a high gatekeeper staffing level to reduce queueing, then gatekeeper utilization is low, the gatekeeper’s average wait between calls is long, and \( d \) is large.

This simplification does reduce the model’s generality. We assume that gatekeepers do not make real-time referral decisions based on the lengths of queues in front of gatekeepers or specialists. Nor do we capture the economies of scale inherent in queueing systems. However, this simpler structure will allow us to focus on the long-term impact of incentives on gatekeeper behavior.

Second, the model assumes a strict separation between the gatekeeper’s diagnosis and treatment steps. From a modeling perspective this is not essential, for it is not difficult to adapt the model so that some parts of diagnosis and treatment are merged. However, this sequential model captures the essence of the motivating applications from health care and call centers. This separation between diagnosis and treatment was also proposed as part of a more general service model in Apte et al. (1997).

Third, we assume that all gatekeepers have the same diagnostic capabilities and that all specialists are homogeneous in terms of cost. That is, the expected cost of a customer who travels from a gatekeeper to a specialist does not depend on the identity of the gatekeeper or specialist. In this section and the next section, we will also assume that all gatekeepers have the same treatment capabilities (defined by the “treatment function,” below). However, this assumption will be relaxed in §5.

Finally, the model seems to assume that a customer who receives incorrect treatment by the gatekeeper is sent directly to a specialist. However, in many actual systems a customer may leave after treatment by the gatekeeper and may reenter the system after discovering that the treatment is incorrect. In fact, the model captures this sequence of events as long as the customer, upon reentry, is sent to a specialist (as would often be the case).

The cost of this inconvenience to the customer and the firm is captured by the parameter \( m \). The following parameters describe the primary costs to the firm:

\[
\begin{align*}
\text{\( w \)} &= \text{gatekeeper wages ($/time \text{ period}$). This parameter is a decision variable in the model.} \\
\text{\( s \)} &= \text{expected cost of treatment by a specialist ($/\text{customer}$). This is an exogenous parameter.} \\
\text{\( m \)} &= \text{expected cost of incorrect treatment by a gatekeeper ($/\text{customer}$). This includes any “goodwill” cost as well as the expected cost of customer reentry and is an exogenous parameter.}
\end{align*}
\]

Now we define the “treatment function” that describes the gatekeeper’s treatment capabilities. To specify the domain of the function, 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. Let \( f(k) \) denote 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 kth 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 relax this assumption in §5. Throughout the paper, we also assume that the expected treatment time t does not vary with call complexity k. In fact, all results in §§3 and 4 hold if t rises as k rises. To keep the notation simpler here, we assume that t does not depend on 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 kth fractile, so that \( F(k) = \int_0^k f(\tau) \, d\tau \) and \( 0 \leq F(k) \leq k \).

We assume that the firm and gatekeeper are both risk neutral, and that the firm is minimizing costs while the gatekeeper is maximizing income per unit time. The gatekeeper’s objective includes a wage (in \$/hour), the time-average value of bonuses (\$ of bonuses/time to achieve those bonuses), and the value of the gatekeeper’s intrinsic preferences, averaged over time. We assume that the treatment and diagnosis times are small relative to the gatekeeper’s planning horizon. Therefore, the gatekeeper’s decisions are made very frequently over a finite planning horizon and an average reward criterion is appropriate (Puterman 1994).

The gatekeeper also has an intrinsic preference to either spend or not to spend time treating customers. This preference is captured by the parameter \( h \) (in \$/time unit), which may be greater than 0 if each time unit of treatment is valuable to the gatekeeper, or may be less than zero if the gatekeeper wishes to avoid spending time on treatment. Therefore, if the gatekeeper is paid a wage \( w \) and decides to treat a proportion \( k \) of customers, then a simple application of renewal theory finds that as the gatekeeper’s time horizon increases the gatekeeper’s time-average income approaches \( w + htk/(d+kt) \) with probability 1 (we will use the phrase time-average income to describe a quantity that is actually time-average equivalent income, for it includes a nonmonetary component, \( h \)). Let \( \tilde{w} \) be the gatekeeper’s reservation income level. For technical reasons, we assume that \( h \) is small enough so that \( \tilde{w} > h(t/(d+t)) \). This implies that the expected intrinsic value from treating all patients is not sufficient to satisfy the gatekeeper’s reservation income.

Now we describe the firm’s problem. Here we assume that the firm pays a wage, \( w \), and that all relevant information is visible, at no cost to the firm. That is, the firm has perfect knowledge of the gatekeeper’s treatment function \( f(k) \), the rank in difficulty \( k \) of each customer’s problem, and the gatekeeper’s decision to treat or not treat a customer. Therefore, the firm can enforce a contract with the gatekeeper that requires the gatekeeper to treat all customers with rank up to \( k \), and to diagnose and immediately refer customers with rank greater than \( k \). In the following problem, the firm must determine the wage \( w \) and threshold \( k \) to minimize the expected cost of service per customer because the number of cases arriving per unit of time is independent of how they are served:

\[
\min_{w,k} \left[ w(d+tk) + (1 - F(k))s + (k - F(k))m \right] \quad (1)
\]

subject to \( w + \frac{htk}{d+tk} \geq \tilde{w}, \quad 0 \leq k \leq 1 \). (2)

The first term in the firm’s objective function is the average cost in wages per call, the second term is the expected cost per call for treatment by a specialist, and the third term is the expected cost of mistreatment. The left-hand side of (2) is the value per unit time earned by the gatekeeper. This participation constraint ensures that the gatekeeper earns her reservation wage.
Because the objective function is monotonic in the wage \(w\), the firm will choose \(w\) so that the participation constraint is satisfied at equality. Therefore, \(w + htk / (d + tk) = \hat{w}\), the reservation income. Substituting this equality into the objective function and rearranging terms produces the minimization problem

\[
\min C(k) = \min_k [(\hat{w}d + s) + (m + (\hat{w} - h)t)k - (m + s)F(k)]. \tag{3}
\]

The solution to this problem is the first-best treatment rate (proofs for all lemmas and propositions are in the Appendix).

**Lemma 1.** The optimal treatment rate \(k^*\) (minimizing (3)) is given by

\[
k^* = \begin{cases} 
0 & \text{if } Z > f(0), \\
\frac{f^{-1}(Z)}{f(0) - f(1)} & \text{if } f(0) < Z < f(1), \\
1 & \text{if } Z < f(1),
\end{cases} \tag{4}
\]

where \(Z = (m + (\hat{w} - h)t)/(m + s)\).

The ratio \(Z\) and the optimal value \(k^*\) have many natural properties. As the cost of mistreatment \(m\) rises, \(Z\) approaches 1 and \(k^*\) falls, for the firm would want fewer gatekeeper treatments and thus fewer mistreatments. As the cost of a specialist rises, \(k^*\) rises, for the firm wishes to reduce the number of referrals to the specialist. Finally, as the marginal wage \(\hat{w} - h\) rises, \(k^*\) falls, because the gatekeeper becomes relatively more expensive.

### 4. The Principal-Agent Problem with Homogeneous Agents

In practice, firms find it difficult to monitor the treatment decisions of its gatekeepers directly. Managers of call centers sometimes listen in on the calls of customer service agents, or audit taped conversations, but they usually monitor only a small sample of all calls handled by an agent. HMOs and other health-care organizations ask doctors and nurses to complete and submit treatment plans, but composing detailed plans is costly to the agent and auditing the plans for accuracy is costly for management. Even if the firm can determine whether treatment has been offered to a particular patient, it is even more difficult to determine whether treatment should have been offered. Each gatekeeper may have a clearer understanding than the firm of the level of difficulty (rank \(k\)) that each customer presents, i.e., whether the customer is a good fit for the particular talents of the gatekeeper. Therefore, it often makes sense for the firm to delegate the treat/refer decision to the gatekeeper. However, a gatekeeper acting independently may make referral decisions that are dramatically suboptimal for the system as a whole. For example, if \(h < 0\) and there are no other incentives, the gatekeeper prefers not to treat and all patients will be referred immediately after diagnosis. Therefore, when monitoring is difficult, it may be necessary for the firm to consider incentives based on observable measures, such as the number of customers seen or referred.

One might think of the individual treatment decisions as an input to the service production process, similar in some ways to the effort variable that is common in the economics and marketing literature. Rather than monitoring inputs, some firms find that it makes economic sense to monitor and reward output. In the gatekeeper’s case, there are two outputs that may be measured by the firm without examining the content of individual interactions between the gatekeeper and the customer: The number of customers seen by the gatekeeper, and the number of customers seen by the gatekeeper and ultimately referred to a specialist (or, equivalently, successfully treated and thus not referred). Therefore, we assume that the firm offers an incentive (or “bonus” or “reward”) \(r_d\) for each customer diagnosed and a reward \(r_s\) for each customer successfully treated (or “solved” or “not referred”).

The number of diagnoses conducted by a gatekeeper can be tracked by counting the number of customers they handle, because all customers require a diagnosis. Successful treatment can be tracked by seeing if customers return with unresolved problems or are referred. For example, technical support call centers usually establish a “trouble ticket” with a unique identifier for each new problem a customer presents. The trouble ticket contains a log of the staff-customer interactions related to the problem. An automated search through the database of trouble tickets would
be able to identify instances, for each gatekeeper, in which a problem was first diagnosed by that gatekeeper, was not seen by a specialist, and the customer did not return to seek help with the same problem. Such an instance would represent a successful treatment\footnote{As mentioned earlier, we assume that a customer who received an incorrect treatment by the gatekeeper will eventually return and be seen by a specialist.} and could be identified without analyzing the details of the interactions. In terms that are often used in agency theory, these outputs are easily observable, verifiable, and a contract based on these outputs is enforceable. Input measures (e.g., treat/not treat) may not meet all three criteria.

The general structure and timing of the problem is identical to the standard principal-agent problem with hidden information. First, the firm offers a long-term contract to the gatekeeper that specifies the wage and a bonus plan, and the gatekeeper decides whether to accept the contract. If she accepts the contract, then she implements a referral strategy that maximizes her own time-average income, and the gatekeeper will accept the contract only if that income exceeds $\hat{w}$.

The firm’s problem is

$$\min_{w, r_d, r_s} \left[ w(d + tk) + r_d + r_s F(k) + (1 - F(k))s + (k - F(k))m \right]$$

subject to

$$w + \frac{r_d + htk + r_s F(k)}{d + tk} \geq \hat{w},$$

$$k = \arg \left\{ \max_{k'} w + \frac{r_d + htk' + r_s F(k')} {d + tk'} \right\},$$

$$0 \leq k \leq 1, \quad r_s \geq 0.$$ (7) (8)

This problem is similar to (1) and (2) above, although now the objective function includes the expected bonus payment per call. Constraint (6), the participation constraint, ensures that the time-average value of wages and bonuses is at least as large as the reservation wage $\hat{w}$. The incentive compatibility constraint, (7), ensures that the gatekeeper chooses a treatment threshold that maximizes her own time-average income. The constraint $r_s \geq 0$ is necessary to make sure that the gatekeeper makes an effort to provide treatment at all. If there were a penalty for correct treatment, $r_s < 0$, then the gatekeeper might have an incentive to spend time providing worthless, or incorrect, treatment. Note, however, that we do not place any nonnegativity restriction on either $w$ or $r_d$. If $w < 0$, then the gatekeeper pays a fee to work (a “franchise fee”), while if $r_d < 0$, there is a fee for each customer routed to the gatekeeper by the firm.

Because the firm adjusts the wage $w$ so that the participation constraint (6) is satisfied at equality, the objective function can be rewritten as

$$\min_{w, r_d, r_s} \left\{ (\hat{w} d + s) + (m + (\hat{w} - h)t)k - (m + s)F(k) \right\},$$

subject to the constraints (6)–(8). This objective function is identical to the objective function for the centralized case (1). The only difference here is that we optimize over the contract parameters, the wage $w$ and the bonuses $r_d$ and $r_s$. While the contract parameters do not appear in the objective function, they determine the value of $k$ through the incentive compatibility constraint. Therefore, first best can be achieved if and only if there exists values of $r_d$ and $r_s$ that produce the optimal $k^*$ from (7). We ask: Is $r_d$, or $r_s$, or both sufficient as a “lever” to guide the gatekeeper towards system-optimal behavior? To find an answer, we examine the gatekeeper’s problem in more detail.

The gatekeeper maximizes the following function subject to the constraint on $k$:

$$\theta(k) = \frac{r_d + htk + r_s F(k)}{d + tk}, \quad 0 \leq k \leq 1.$$ (10)

Let $k^*$ denote the gatekeeper’s response to incentives $r_d$ and $r_s$. The first-order condition $\theta(k^*) = 0$ is equivalent to

$$\frac{(r_d - hd)t}{r_s} = (d + tk^*)f(k^*) - tF(k^*),$$

and this can be rewritten as

$$\frac{r_d + htk^* + r_s F(k^*)}{d + tk^*} = \frac{r_s f(k^*)}{t} + h.$$ (12)

This condition states that the gatekeeper chooses a treatment threshold $k^*$ such that the gatekeeper’s current time-average income (the left-hand side) is equal
to the time-average income obtained by attempting treatment of a customer with difficulty \( k^g \). The first term on the right-hand side is the marginal expected reward for a “solve” per time unit spent on the treatment, and \( h \) is the intrinsic reward for treatment per unit time.

The following lemma will help us to characterize the gatekeeper’s response.

**Lemma 2.** Let \( k^g \) be an interior \( (0 < k^g < 1) \) critical point that satisfies \( \theta' (k^g) = 0 \). Then \( k^g \) is a unique, global maximum and is the gatekeeper’s optimal response to incentives \( r_d \) and \( r_s \).

From the condition \( \theta' (k^g) = 0 \) we find

\[
\begin{align*}
\text{(i)} & \quad \text{sign} \left( \frac{\partial k^g}{\partial r_s} \right) = \text{sign} \left( \frac{r_d}{d} - h \right), \\
\text{(ii)} & \quad \frac{\partial k^g}{\partial r_d} < 0, \\
\text{(iii)} & \quad \frac{\partial k^g}{\partial r_s} < 0, \\
\text{(iv)} & \quad \text{sign} \left( \frac{\partial k^g}{\partial d} \right) = \text{sign} \left( \frac{r_s f(k)}{t} + h \right),
\end{align*}
\]

where \( \text{sign}() \) indicates the sign of the argument. Result (i) is a bit counterintuitive, for we might expect an increase in \( r_s \) to always lead to an increase in treatment rates. If \( r_d/d > h \) our expectation is confirmed. In this case, the intrinsic reward for time spent on treatment, \( h \), is less than the incentive for diagnosis per unit time, \( r_d/d \). Therefore, the reward, \( r_s \) balances the gatekeeper’s tendency to diagnose, an increase in \( r_s \) pulls the treatment rate up, and \( \partial k^g/\partial r_s > 0 \). However, if \( h > r_d/d \), the gatekeeper already prefers to treat (regardless of the size of \( r_s \)). In this case, an increase in the reward for successful treatment (\( r_d \)) serves to pull the treatment rate down. As the reward rises, the gatekeeper only retains those customers who are most likely to result in a reward—those with a low value of \( k \).

In (ii), increasing \( r_d \) pulls gatekeepers away from treatment and towards diagnosis, and in (iii) an increase in treatment time makes treatment more costly to the gatekeeper and therefore treatment rates fall as \( t \) rises. In (iv), an increase in diagnosis time or the expected idle time between customers implies a larger opportunity cost for beginning work with a new customer, and therefore the gatekeeper would increase the proportion of customers treated as long as the incremental reward for treatment, per unit time, is positive.

Now we return to our original question: Is \( r_d, r_s \), or both, sufficient as a “lever” to guide the gatekeepers to system-optimal referral rate \( k^g \)?

**Proposition 1.** Suppose that under the first-best contract, the treatment threshold is \( k^*, 0 < k^* < 1 \). Then there exists an infinite number of incentive pairs \((r^*_s, r^*_d)\) with \( r^*_s > 0 \) that are sufficient for achieving first best. Incentive \( r^*_s > 0 \) is always necessary for achieving first best, and for certain parameter values, it may also be necessary to offer an appropriate \( r^*_d \neq 0 \).

Therefore, it may not be possible for the firm to implement first best when incentives are based only on referral rates, as they sometimes are in practice. For example, if \( h = 0 \), expression (11) indicates that the gatekeeper’s response is determined by the ratio \( r_d/r_s \). In general, the bonus \( r_s \) encourages treatment while \( r_d \) provides balance by encouraging quick referrals whenever appropriate. The following example also illustrates why the firm may need to offer both incentives to achieve system-optimal behavior.

**Example 1.** Let \( f(k) = 1 - k \), \( h = 1 \) (the gatekeeper prefers to treat), and \( d = 1 \) (time is normalized so that diagnosis takes 1 time unit). Given these parameters, we can solve the first-order condition \( \theta' (k^g) = 0 \) to find

\[
k^g = \frac{1}{t} \left( \sqrt{1 + 2t^2 + \frac{2t^2}{r_s} (1 - r_d)} - 1 \right).
\]

In Figures 3 and 4, we plot the gatekeepers response \( k^g \) as a function of \( r_s \) when \( r_d = 2 \) and \( r_d = 0.5 \), respectively. We can see from these graphs that it is possible
that if treatment incentive package that only includes the flat wage and the may be costly for the firm to rely on a compensa-
tion to achieve first best (as in Figure 3). In this case, it is necessary for first-best. If the first-best threshold \( k^* \) is lower than \( k^*_f \), a small volume incentive, or \( r_d^* = 0 \), is sufficient for first best.

In general, the firm can solve its problem by first calculating the optimal threshold \( k^* \) and then offering each gatekeeper a contract \((w^*, r_d^*, r_s^*)\) that satisfies the participation constraint and induces the gatekeeper to set treatment threshold \( k^* \) through (11) or (12). Therefore, the firm has one degree of freedom when deciding upon the contract parameters. For example, the firm may choose any \( r_s^* > 0 \). Then, the appropriate \( r_d^* \) is determined by

\[
r_d^* + h tk^* + r_s^* F(k^*) \quad \frac{d + tk^*}{d + tk^*} = \frac{r^*_s f(k^*)}{t} + h. \tag{15}
\]

Finally, \( w^* \) is determined by satisfying the participation constraint at equality. In fact,

\[
w^* = (\hat{w} - h) - \frac{r^*_s f(k^*)}{t}. \tag{16}
\]

This equation shows how the contract embodies a trade-off between a guaranteed wage and the bonus income. Given that the gatekeeper is risk neutral, any contract \((w^*, r_d^*, r_s^*)\) that satisfies (15) and (16) achieves first best.

However, if the gatekeeper were not risk neutral, raising \( r_s^* \) and lowering \( w^* \) would increase the income risk to the gatekeeper because the duration and outcome of each treatment is uncertain. If the gatekeepers were risk averse, the firm must consider the cost of the gatekeeper’s risk premium when deciding upon the optimal contract.

5. Heterogeneous Agents

Most service centers have a significant amount of variability in the skills of customer service agents (for one example, see the study of PCPs by Franks et al. 2000). Here, we consider a general service center in which there are two types of gatekeepers, low-skilled and high-skilled, with associated treatment functions \( f_i(k) \) and \( f_h(k) \), \( f_h(k) \geq f_i(k) \quad \forall k \in [0, 1] \). As in the last section, \( f_i^*(k) < 0 \) and \( F(k) = \int_0^k f_i(\tau) d\tau, \quad i = l, h \). We also assume that \( f_l(0) > f_i(0) \), so that \( F_l(k) > F_i(k) \quad \forall k \in [0, 1] \). Because of their superior treatment abilities, high-skilled gatekeepers demand a higher reservation wage than low-skilled gatekeepers: \( \hat{w}_h > \hat{w}_l \).
To focus on the effects of heterogeneity in the gatekeeper’s skills, we assume that the gatekeepers of either type have no intrinsic preference for or against treatment, so that \( h = 0 \) (we discuss the model with \( h \neq 0 \) at the end of this section).

We define \( p \) as the proportion of current workers who are type \( l \), and assume that the firm knows the value of \( p \). We show that the optimal treatment thresholds depend upon the value of \( p \), and we assume that \( p \) is stable over time so that the firm need not adjust thresholds and renegotiate labor contracts during each gatekeeper’s tenure. This assumption is appropriate for large service systems with stable labor supplies. For example, \( p \) might vary significantly over time in small medical groups and in call centers with just a few employees, but \( p \) is unlikely to change rapidly in a large network of health-care providers and in the many call centers with a hundred or more customer service representatives.

The firm would like to choose treatment thresholds \( k_{i} \) and \( k_{h} \) that minimize the expected system cost per customer. Let \( k_{i} \) and \( k_{h} \) denote optimal solutions to the homogeneous gatekeeper problem (Equation (1)) with \( F(k) \) replaced by \( F_{l}(k) \) and \( F_{h}(k) \), respectively. While \( k_{i} \) and \( k_{h} \) are cost-minimizing thresholds when there is a single type of gatekeeper, they are not optimal when two types of gatekeepers work in parallel. This is because the chosen thresholds influence both the expected cost as a customer passes through the system and the service rate of each type of gatekeeper. For example, if \( k_{i} < k_{h} \) then \( d + tk_{i} < d + tk_{h} \), low-skilled gatekeepers serve customers at a higher rate than high-skilled gatekeepers, and therefore the probability that a randomly selected arriving customer will encounter a low-skilled gatekeeper is larger than \( p \), the proportion of low-skilled gatekeepers.

We see that two groups of gatekeepers with different values of \( k \) may have different service rates, and we incorporate this complication into our model, below. However, there is another potential complication: The two groups may also have different values of \( d \). In the previous section, we assumed that \( d \), the expected sum of server idle and diagnosis times, is the same for all gatekeepers. We are now confronted with a queuing system with two types of gatekeepers (or servers), and each type has a different mean service time. Even if we assume that the diagnosis time is the same across server types, in such a system the two types of servers may have different expected idle times between calls.

To test the reasonableness of the assumption that low- and high-skilled gatekeepers see similar values of \( d \), we conducted simulation experiments in which the difference in service time between the two types of servers is exaggerated. The simulated service center is staffed by low-skilled gatekeepers who perform only diagnosis and high-skilled gatekeepers who diagnose and then always attempt to treat. We assume that the time to perform the diagnosis is the same for each server type but that treatment is significantly longer than diagnosis (see Shumsky and Pinker 2002 for a more detailed description of the model, its parameters, and the results). We find that for a relatively small system (ten gatekeepers of each type) and a utilization of 0.7, the difference between the values of \( d \) for low- and high-skilled gatekeepers is approximately 3%. As the system size increases or the utilization increases the difference in the values of \( d \) vanish. In general, we find that the assumption of equal \( d \)’s is reasonable as long as the diagnosis times are similar and the firm does not use a routing scheme that distinguishes between gatekeeper types.

In §5.1, we examine the first-best treatment thresholds \( k_{i} \) and \( k_{h} \) and numerically compare this solution with \( k_{i} \) and \( k_{h} \) as well as with an optimal, single threshold that is applied to both gatekeepers. In §5.2, we consider the adverse selection problem in which the firm does not observe the “type” of each gatekeeper at the time of hiring.

### 5.1. The First-Best Solution with Two Gatekeeper Types

Let \( q(k_{i}, k_{h}) \) denote the proportion of customers seen by a low-skilled gatekeeper. Therefore,

\[
q(k_{i}, k_{h}) = \frac{p(1/(d + k_{i}))}{p(1/(d + k_{i})) + (1 - p)(1/(d + k_{h}))}, \quad (17)
\]

Let \( C_{i}(k) \) denote the expected system cost of a customer who is handled by a gatekeeper of type \( i \), \( i = l, h \). The participation constraint may be met by
paying wages $\hat{w}_i$ and $\hat{w}_h$ to each gatekeeper type, and the expected cost to provide service to a single call is

$$C_i(k) = (\hat{w}_i d + s) + (m + \hat{w}_i t)k - (m + s)f_i(k),$$

$$i = l, h. \quad (18)$$

Because $q$ is the probability that a particular call will be routed to a low-skilled gatekeeper, the firm’s problem is

$$\min_{k_l, k_h} \Pi(k_l, k_h)$$

$$= \min_{k_l, k_h} [q(k_l, k_h)C_l(k_l) + (1 - q(k_l, k_h))C_h(k_h)]. \quad (19)$$

The first-order conditions for the firm’s profit function are

$$\frac{\partial \Pi}{\partial k_l} \bigg|_{(k_l^*, k_h^*)} = q(k_l^*, k_h^*) \frac{\partial C_l(k_l^*)}{\partial k_l} + (C_l(k_l^*) - C_h(k_h^*)) \frac{\partial q(k_l^*, k_h^*)}{\partial k_l} = 0, \quad (20)$$

$$\frac{\partial \Pi}{\partial k_h} \bigg|_{(k_l^*, k_h^*)} = (1 - q(k_l^*, k_h^*)) \frac{\partial C_h(k_h^*)}{\partial k_h} + (C_l(k_l^*) - C_h(k_h^*)) \frac{\partial q(k_l^*, k_h^*)}{\partial k_h} = 0. \quad (21)$$

Depending upon the parameter values $\Pi(k_l, k_h)$ may be either concave or convex in the region $0 \leq k_l \leq 1$, $0 \leq k_h \leq 1$. However, we can show that if there is an interior critical point then that point is a local minimum.

**Lemma 3.** Let $(k_l^*, k_h^*)$ be an interior solution of $\Pi(k_l, k_h)$ that satisfies (20) and (21). The solution $(k_l^*, k_h^*)$ is local minimum of $\Pi(k_l, k_h)$.

Let $k_l^*$ and $k_h^*$ be treatment thresholds that minimize system costs, given only low-skilled or high-skilled gatekeepers. In the following example, we see that the solution $(k_l^*, k_h^*)$ may be significantly different from $(k_l^*, k_h^*)$.

**Example 2.** Let $f_l(k) = t k + \frac{1}{2}(1 - k)$, $f_h(k) = m + \frac{1}{2}(1 - k)$, $h = 0$, $d = 1$, $t = 2$, $w_l = 1$, $w_h = 1.2$, $m = 10$, and $s = 10$. Figure 5 displays the optimal treatment thresholds $(k_l^*, k_h^*)$, the individual solutions $(k_l^*, k_h^*)$, and an optimal single-threshold $k^*$ (this is the optimal threshold if we require both gatekeeper types to use the same $k$). We see how these quantities vary as the proportion of low-skilled gatekeepers, $p$, changes.

Note that as $p$ rises $k_h^*$ declines. This is because the marginal value of additional customer flow to the high-skilled gatekeepers rises as the high-skilled gatekeepers become scarce. In general, a change in the value of $p$ can have a significant effect on the optimal solution.

5.2. The Principal-Agent Problem with Two Gatekeeper Types

As in §4, we assume that it is prohibitively costly for the firm to monitor each treatment decision and enforce contracts based on specified treatment thresholds. Again, the firm is faced with a principal-agent problem with hidden information. But now, we also assume that each gatekeeper knows her own skill level when applying for the job but that the gatekeeper’s skill level is hidden from the firm. The firm offers a contract or a set of contracts, and the gatekeeper either chooses the most valuable contract or refuses all contracts to accept an alternative with a guaranteed wage $\hat{w}_i$, $i = l, h$. If the gatekeeper accepts a contract, she chooses a treatment threshold that maximizes her long-run average income.

This is an example of adverse selection. In particular, this is a screening problem, for it is often optimal for the firm to offer contracts that “screen” the gatekeepers: Given these contracts, a gatekeeper reveals
her type when choosing from among the contracts. Rothschild and Stiglitz (1976) presented an influential model of this screening problem. In their framework, insurance companies compete for customers by offering policies with a variety of premiums, deductibles, etc. While our model captures the important element of customer heterogeneity—our gatekeeper types correspond to insurance customers with a variety of risk profiles—our model does not describe the competition among firms for labor, and therefore we assume that parameters $\tilde{w}_i$, $\tilde{w}_h$, and $p$ are exogenous. Our model is a hybrid of the monopolistic screening model described in Mas-Colell et al. (1995, p. 500) and a principal agent model with hidden information, for after the gatekeeper is hired the firm discovers the gatekeeper’s type but cannot monitor each interaction between gatekeeper and customer. We will show that the first-best result can be achieved by either (i) offering a single contract that simultaneously coordinates the treatment thresholds of both gatekeepers, or by (ii) offering two contracts that screen the gatekeepers and produce the optimal treatment thresholds from each of them.

For convenience, we describe the problem under the assumption that the firm offers two contracts. We will show that no more than two contracts are needed to achieve first best, and the single-contract formulation is a special case with two identical contracts. The two contracts are specified by two vectors $\mathbf{w} = (w^i, r^i_j, r^h_j)$, where the contracts with $i = l$ and $i = h$ are designed for the low-skilled and high-skilled gatekeepers, respectively.

To describe the firm’s problem, let

$$D_i(k_i) = w^i(d + tk_i) + r^i_d F_i(k_i) + (1 - F_i(k_i))s + (k_i - F_i(k_i))m \quad \text{for } i = l, h.$$ (22)

The firm’s problem is

$$\min_{\mathbf{w}, \mathbf{w}'} [q(k_i, k_h)D_i(k_i) + (1 - q(k_i, k_h))D_h(k_h)]$$ (23)

subject to

$$w^i + \frac{r^i_d + r^i_d F_i(k_i)}{d + tk_i} \geq \tilde{w}_i, \quad i = l, h,$$ (24)

$$\frac{r^i_d + r^i_d F_i(k_i)}{d + tk_i} = \frac{r^i_s f_i(k_i)}{t}, \quad i = l, h,$$ (25)

$$w^i + \frac{r^i_d + r^i_d F_i(k_i)}{d + tk_i} \geq \tilde{w}_i, \quad i = l, h,$$ (26)

$$0 \leq k_i \leq 1, \quad 0 \leq k_h \leq 1, \quad r^i_s, r^h_s \geq 0.$$ (27)

Constraints (24) are the participation constraints and constraints (25) are the gatekeepers’ first-order conditions, as expressed in (12), with $h = 0$. These are equivalent to the incentive compatibility constraint (7) from the original single-gatekeeper problem. Constraint (26) ensures that each type of gatekeeper chooses the contract designed for her type. These are sometimes called the truth-telling or self-selection constraints.

By replacing $w^i$ in each cost function $D_i(k_i)$ with the expression implied by the tight participation constraints, the firm’s problem can be rewritten as

$$\min_{\mathbf{w}, \mathbf{w}'} [q(k_i, k_h)C_i(k_i) + (1 - q(k_i, k_h))C_h(k_h)],$$ (28)

subject to the constraints (24)–(27). As in §4, this objective function is identical to the objective function for the first-best problem, and we can again reduce the problem to finding optimal contracts that guide the gatekeepers to choose the first-best treatment thresholds $(k^*_l, k^*_h)$.

We now use the procedure described at the end of §4 to construct a contract $(w^i, r^i_s, r^i_d)$ for the gatekeeper type $i, i = l, h$, that ensures the first-best solution and satisfies constraints (24)–(27). Given any $r^i_s > 0$, the corresponding value of $r^i_d$ is determined by the appropriate gatekeeper first-order condition,

$$\frac{r^i_d + r^i_d F_i(k^*_i)}{d + tk^*_i} = \frac{r^i_s f_i(k^*_i)}{t}, \quad i = l, h.$$ (29)

Then, by combining (29) with the participation constraint satisfied at equality, we find the worker’s wage

$$w^i = \tilde{w}_i - \frac{r^i_s f_i(k^*_i)}{t}, \quad i = l, h.$$ (30)

This equation describes an indifference curve (actually, a straight line) for the type-$i$ worker. In fact, any $(w^i, r^i_s)$ pair that satisfies (30) provides the gatekeeper with an expected time-average income of $\tilde{w}_i$ while ensuring that the gatekeeper chooses the first-best treatment threshold.
The following lemma will be useful for deriving the optimal contracts.

**Lemma 4.** There exists a bonus ratio \( R^* = (r_d/r_s)^* \) which induces \( k^*_1 \) from the low-skilled gatekeeper and \( k^*_h \) from the high-skilled gatekeeper.

From the discussion after Proposition 1, we know that the threshold chosen by each gatekeeper is determined by the ratio \( r_d/r_s \), and that this ratio induces the gatekeeper to balance time spent on diagnosis with time spent on treatment. When there are two gatekeeper types, the optimal thresholds \( k^*_1 \) and \( k^*_h \) can be quite different, but Lemma 4 indicates that the balance between the reward for volume and the reward for “solves” should be the same.

Now consider the indifference curves of the two gatekeeper types. If the curves satisfy the single-crossing property then there are two screening contracts that force the gatekeepers to self-select, i.e., satisfy constraint (26) (see Mas-Colell 1995, p. 453).

**Lemma 5.** For \( r_i > 0 \), the indifference curves \( w^i = \tilde{w}_i - r^i f_i(k^*_i)/t \), \( i = l, h, \) satisfy the single-crossing property: They intersect only once, and the indifference curve of the high-skilled gatekeeper has the steeper slope.

Given Lemmas 4 and 5, it follows that the firm can achieve first best by offering one contract, or by screening the gatekeepers with two contracts. To describe the contracts, define \( (w^i, r^i) \) as the point of intersection of the two indifference curves,

\[
\frac{r^i}{s} = \frac{t(\tilde{w}_h - \tilde{w}_l)}{f_s(k^*_l) - f_s(k^*_h)} \quad \text{and} \quad w^s = \tilde{w}_l - r^s f_s(k^*_l)/t. \tag{31}
\]

Let \( R^* = (r_d/r_s)^* \) be the optimal bonus ratio that induces \( k^*_1 \) \( (k^*_h) \) from the low-skilled (high-skilled) gatekeepers. Now consider the following sets of contracts.

**Single Contract.** The firm offers one contract \( w^i = (w^l, r^l, r^h) \), where \( w^s \) and \( r^s \) are determined by (31) and \( r^d = R^* r^s \).

**Screening Contracts.** The firm offers two contracts:

1. \( w^l = (w^l, r^l, r^h) \) where \( 0 < r^l < r^h \) and \( w^l \) satisfies (30), and
2. \( w^h = (w^h, r^h, r^h) \) where \( r^s < r^h \) and \( w^h \) satisfies (30).

**Proposition 2.** The firm can achieve first best with either the single contract \( w^i \) or with the screening contracts \( w^l \) and \( w^h \). The two screening contracts make sense: High-skilled gatekeepers gravitate towards \( w^h \), a contract with relatively large bonuses for solves, while low-skilled gatekeepers accept \( w^l \), a contract with lower bonuses and a larger flat wage. While the existence of these two contracts is not unexpected, it is surprising that the first-best result can also be achieved with a single contract.

In models with more than two types of gatekeepers, it is possible that Lemma 4 still holds (the results of numerical experiments with three and four types of gatekeepers are consistent with Lemma 4, although we have not been able to find a general proof). However, given more than two gatekeeper types, we know that Proposition 2 does not necessarily hold. For example, given three gatekeeper types, it may be impossible to find a set of contracts that are incentive-compatible and satisfy the participation constraints at equality. Given certain parameters for the three treatment functions, it may be necessary to pay one group of gatekeepers a bit extra to satisfy the participation constraints of the other two groups. (The screening model described in Mas-Collell et al. 1995 has a similar property.) When gatekeepers may be ordered along a continuum of types the first-best solution may be formulated as a continuum of treatment thresholds. But again, there is no guarantee that Proposition 2 holds and that the first-best solution to the principal-agent problem can be achieved when there is a continuum of types.

In practice, most firms set a limited number of treatment thresholds and offer a limited number of contracts because each additional threshold to monitor or contract to process adds complexity to the firm’s processes and costs to the firm. Therefore, the benefits of distinguishing among numerous gatekeeper types may not be worth the price. If gatekeepers can be sorted into two groups, and if the firm can define aggregate treatment functions that are reasonable approximations of the abilities of the members of gatekeepers.

---

2 When low-skilled and high-skilled gatekeepers have equivalent preferences for treatment, \( h \neq 0 \), Proposition 2 holds. However, if the two types of gatekeepers have different preferences for treatment, there exist situations in which their indifference curves do not cross, and screening contracts for the gatekeepers do not exist.
each group, then the contracts described here should achieve results that are close to first best.

6. Discussion and Extensions

In this paper, we have examined the tension between the gatekeeper’s role as a provider of treatment and her role in authorizing referrals to specialists, and we have described the problem this poses to managers who cannot assume that a gatekeeper will make system-optimal referral decisions. We have shown how two incentives, one based on customer volume and the other based on referral rates, serve to balance each other as they influence the gatekeeper’s decisions. When offered the proper combination of these incentives, gatekeepers will choose system-optimal referral rates. Both types of incentives are seen in practice (see Thaler-Carter 1999 for examples in call centers and Magnus 1999 for examples in health care). Of course, there are many other incentive systems in use, but many of these may be linked to the output-based incentives described here. We have also shown, as a caution to managers, that in general one incentive may not be sufficient to elicit the gatekeeper behavior the firm wants. When gatekeepers are heterogeneous, we have seen that it is possible to devise contracts that screen two types of employees and lead them to make referral decisions that are optimal for the firm.

Many firms seek to supplement or even replace these incentive systems by attempting to maintain direct control over gatekeeper behavior. In call centers, managers monitor a sample of gatekeeper calls and perform post-hoc examinations of gatekeeper referral decisions. Some HMOs require physicians to apply for preauthorization of referrals and some implement periodic utilization reviews. In terms of our model, such firms attempt to make gatekeeper employment contingent on the maintenance of a particular treatment threshold $k$. Of course, the net benefit of such a contract depends upon the cost to observe $k$ and the difficulty in determining the best value of $k$ for each gatekeeper. The backlash against gatekeeper systems in health care described in Lawrence (2001) can in part be blamed on controls such as preauthorization that were seen as burdensome by both physicians and patients.

In fact, many have questioned the need for gatekeeping itself. Suggested alternatives include increased use of teamwork (e.g., Lawrence 2001) and other methods for increasing the level of information-sharing between specialists and gatekeepers. Firms can capture information gathered by specialists and make it available to gatekeepers via training and access to a “knowledge base” that describes each treatment path. This database can also be used by the firm to construct treatment templates that guide gatekeepers as they treat customers. However, information-sharing can help reduce the number of referrals, but it does not eliminate the need for specialization. Specialization can be particularly valuable in volatile service environments such as high technology and health care in which new knowledge is continuously created. Team-based systems can also be helpful, but they reallocate the agency problem: How should the team performance be measured, and how should the team be compensated? In addition, such systems lose the economies of scale provided by gatekeeper systems in which customers arrive to a large pool of generalists and then may be referred to specialists.

We have identified several possible extensions to the paper. First, at the end of §5, we explained that there is no assurance that a first-best solution to the principal-agent problem can be achieved when there is more than two gatekeeper types. However, it may be possible to characterize the space of treatment functions and cost parameters that would permit first-best solutions, given an arbitrary number of gatekeeper types. Another potentially important extension to this model is to incorporate the role of gatekeeper learning in the routing decision. The model with two gatekeeper skill levels examined in §5 may be interpreted as a service center with two levels of gatekeeper experience (i.e., in steady-state a proportion $p$ of gatekeepers have little experience and $1 - p$ have significantly more experience). However, if the gatekeepers “learn by doing,” over the long run the level of experience itself may be affected by the gatekeeper’s treatment threshold. A firm may prefer that a gatekeeper tackle difficult problems so that the gatekeeper learns to handle that problem in the future. Pinker and Shumsky (2000) examine the effects of agent learning in a service center with
specialized and flexible workers, but that paper did not examine the role of incentives.

The traditional principal-agent literature models worker effort in a stylized way. We have developed a framework that shows explicitly how a worker’s choices affect the time available for other activities. There are many opportunities for applying this framework. For example, in sales force management the literature on incentives and sales uses highly stylized models to link “effort” to sales. One might imagine a more detailed model that distinguishes between easy and difficult leads analogous to our simple and complex customers. Difficult leads require more time commitment from the salesperson than easy leads, while the pursuit of each type of lead has different revenue implications for the firm. Additional applications are possible in healthcare or other professional services such as consulting or legal work.

Acknowledgments

The authors thank seminar participants at the Kellogg Operations Workshop, Northwestern University; the Center for Operations Research and Economics, Université Catholique de Louvain; the Weatherhead School of Management, Case Western Reserve University; and the University of Washington Business School. They particularly appreciated helpful suggestions from Marshall Freimer, Martin Lariviere, Yong-Pin Zhou, the editor, associate editor, and two anonymous referees.

Appendix: Proofs of Lemmas and Propositions

Proof of Lemma 1. We know that $C(k) > 0$ because $f(k) \leq k \leq 1$ and, by assumption, $\hat{w} > \frac{ht}{(d + t)}$. In addition, $C'(k) = (m + (\hat{w} - h)t) - (m + s)f'(k)$ and $C''(k) = -(m + s)f''(k)$. Because $f'(k) < 0$, $C''(k) > 0$, and the function $C(k)$ is strictly concave. Setting $C'(k) = 0$, $f'(k^*) = \frac{m + (\hat{w} - h)t}{m + s}$. (A1)

The lemma follows from this expression and the boundary conditions $0 \leq k \leq 1$. □

Proof of Lemma 2. Let $\theta'(k^*) = 0$. Taking the second derivative of $\theta(k)$ and employing equality (11) shows that $\theta''(k^*) = \frac{r_if'(k^*)}{(d + tk^*)} < 0$, where the inequality follows from the fact that $f(k)$ is strictly decreasing in $k$. Therefore, $\theta(k)$ is strictly concave at $k^*$. Because $\theta(k)$ is strictly concave at any interior critical point, $\theta(k)$ is strictly concave at $k^*$.

Proof of Proposition 1. Consider the gatekeeper’s first-order condition $\theta'(k) = 0$, which can be rewritten as (11). Define the left-hand side of (11), $(r_\ell - h_d)t/r_\ell = L(r_\ell, r_\ell^*)$ and define the right-hand side, $(d + tk)f(k) - \frac{tF(k)}{1 + t} = R(k)$. We want to determine if there exist parameters $(r_\ell^*, r_\ell^*)$ such that $L(r_\ell^*, r_\ell^*) = R(k)$ for any $0 < k^* < 1$. Clearly, a necessary condition is that $r_\ell^* > 0$, for if $r_\ell^* = 0$, $L(r_\ell^*, r_\ell^*) = -\infty$ or $\infty$, and the gatekeeper would set $k^* = 0$ or $k^* = 1$. The equality $L(r_\ell^*, r_\ell^*) = R(k^*)$ is equivalent to $r_\ell^* = \frac{(r_\ell^* - h_d)t}{R(k^*)}$. (A2)

Note that $R(k^*)$ can be either positive or negative. If we set $r_\ell^* = 0$ and if $h$ and $R(k^*)$ have the same sign, then $r_\ell^* < 0$, contradicting our assumption that $r_\ell^* > 0$. Therefore, $r_\ell^* \neq 0$ may be necessary to implement first best. Finally, if we allow any $r_\ell^* > 0$ and any $r_\ell^* \neq 0$, then Equation (A2) specifies an infinite number of pairs $(r_\ell^*, r_\ell^*)$ that satisfy the gatekeeper’s optimality condition and thus produces the first-best outcome. □

Proof of Lemma 3. To demonstrate that $II(k_i, k_s)$ is strictly convex at the critical point, we show that at $(k_i^*, k_s^*)$ the Hessian of the objective function $II$ is positive definite. The technique is straightforward but tedious, and details can be found in Shumsky and Pinker (2002). □

Proof of Lemma 4. To simplify the notation, let $q = q(k_i^*, k_s^*)$, $q = \frac{\partial q}{\partial k_i}$, $q'_i = \frac{\partial^2 q}{\partial k_i^2}$, $C_i = C_i(k_i^*)$, $C_s = C_s(k_s^*)$, $C_i^* = \frac{\partial C_i}{\partial k_i}$, $C_s^* = \frac{\partial C_s}{\partial k_s}$, $s_i = s_i(k_i^*)$, $s_s = s_s(k_s^*)$, and $F_i = F_i(k_i^*)$. The following identity, which follows from the first-order conditions (20) and (21), will be useful:

$pf_i + (1 - p)f_s = \frac{m(\hat{w}_i + (1 - p)\hat{w}_s)t}{m + s}$. (A3)

Subtracting (21) from (20),

$(C_s - C_i)(q_i^* - q_i) = qC_i - (1 - q)C_s$. (A4)

Using the expressions for $C_i$, $C_s$, etc., and rearranging terms in (A4), we find

$f_i = \frac{G[m + (p\hat{w}_i + (1 - p)\hat{w}_s)t - (m + s)(pf_i + (1 - p)f_s)]}{H} + (d + t_k)\hat{f}_i - (d + t_k)\hat{f}_s$. (A5)

where

$G = d^2(2p - 1) - 2d((1 - p)k_i^* - pk_i^*)^2 - f_i((1 - p)(k_i^*)^2 - p(k_i^*)^2),

H = -p(1 - p)(m + s)((d + k_s^*) + (d + k_s^*)).$

By using (A3) and rearranging terms, (A5) reduces to

$(d + t_k)\hat{f}_i - tF_i = (d + t_k)\hat{f}_s - tF_s$. (A6)

Now, choose a value of $R^* = (r_\ell^*/r_\ell^*)^*$ so that $k^*$ is induced from the low-skilled gatekeeper. From the first-order conditions (25) and Expression (A6),

$R^* = \frac{1}{T}\{(d + t_k)\hat{f}_i - tF_i\} = \frac{1}{T}\{(d + t_k)\hat{f}_s - tF_s\}$.

Therefore, $R^*$ also induces the optimal threshold $k^*$ from the high-skilled gatekeeper. □
Proof of Lemma 5. By assumption, $\hat{w}_l > \hat{w}_i$. Because the slopes of the two lines are $-f_i(k_l)$ and $-f_i(k_i)$, we must show that $f_i(k_l) > f_i(k_i)$. First, assume that $k_1 < k_l$. Because $f_i(k) \geq f_i(k), \forall k \in [0, 1]$, and the treatment function of the low-skilled gatekeeper is decreasing, $f_i(k_l) > f_i(k_i)$. Now assume that $k_1 > k_l$. We rewrite (A6) as

$$f_i(k_l) - f_i(k_i) = \frac{t}{d + k_i^0} \left( \int_0^{r_i^0} (f_i(\tau) - f_i(\tau)) d\tau + \frac{r_i^0}{k_l^0} f_i(k_l^0) \right) \geq \frac{t}{d + k_l^0} \left( \int_0^{r_i^0} (f_i(\tau) - f_i(\tau)) d\tau + \frac{r_i^0}{k_l^0} f_i(k_l^0) \right) \geq \frac{t}{d + k_l^0} (f_i(k_l^0) - f_i(k_i^0))^0$$

where the first inequality follows from the fact that $f_i(\cdot)$ is decreasing. □

Proof of Proposition 2. For $w^*$, we know from Lemma 5 that a suitable pair $(w^*, r_i^*)$ must exist with $r_i^* > 0$. From the definition of $w^*$, we know that the contract induces the optimal thresholds $k_l^*$ and $k_i^*$, the participation constraints are satisfied at equality, and the self-selection constraints are satisfied (also at equality). Therefore, the firm achieves first best.

Now consider $w^l$ and $w^h$. The construction of the contracts ensure that if the low-skilled gatekeeper chooses contract $w^l$ and if the high-skilled gatekeeper chooses contract $w^h$, then the participation constraints (24) are satisfied at equality, while setting $r_i = R_i r_i^*, i = l, h$, ensure that the gatekeepers choose the first-best treatment thresholds. We now show that the self-selection constraint (26) for low-skilled gatekeepers holds. The proof for high-skilled gatekeepers is similar.

Let $k_l^0$ be the low-skilled gatekeeper’s income-maximizing response to the contract $w^l$, so that $k_l^0$ satisfies

$$r_i^0 + r_i^0 f_i(k_l^0) \left( \frac{d + k_l^0}{d + k_l^0} \right) = r_i^0 f_i(k_l^0) \frac{d + k_l^0}{t}.$$ 

From Lemma 5, if $r_i^* > r_i^*$, $\hat{w}_i - r_i^0 f_i(k_l^0)/t > w^h$. Now let $k_i$ be the gatekeeper’s income-maximizing response to $w^h$. Therefore,

$$w^h + \frac{r_i^0 + r_i^0 f_i(k_l^0)}{d + k_l^0} = \frac{w^h + r_i^0 f_i(k_l^0)}{t}$$

$$> \frac{w^h + \frac{r_i^0 f_i(k_l^0)}{d + k_l^0}}{t}$$

$$= \frac{w^h + \frac{r_i^0 + r_i^0 f_i(k_l^0)}{d + k_l^0}}{t} \\forall k \in [0, 1].$$ □

References


Fama, E. F. 1991. Time, salary, and incentive payoffs in labor contracts. J. Labor Econ. 9(1) 25–44.


SHUMSKY AND PINKER
Gatekeepers and Referrals in Services


Thaler-Carter, R. E. 1999. Why sit and answer the phone all day? *HR Magazine* 44(3) 98.