
Staffing and routing in a two-tier call centre

Sameer Hasija*, Edieal J. Pinker and
Robert A. Shumsky

Simon School, University of Rochester,
Rochester 14627, NY, USA

Fax: 585 273 1140 E-mail: hasijas1@simon.rochester.edu

E-mail: pinker@simon.rochester.edu

E-mail: shumsky@simon.rochester.edu

*Corresponding author

Abstract: This paper studies service systems with gatekeepers who diagnose a customer problem and then either refer the customer to an expert or attempt treatment. We determine the staffing levels and referral rates that minimise the sum of staffing, customer waiting, and mistreatment costs. We also compare the optimal gatekeeper system (a two-tier system) with a system staffed with only experts (a direct-access system). When waiting costs are high, a direct-access system is preferred unless the gatekeepers have a high skill level. We also show that an easily computed referral rate from a deterministic system closely approximates the optimal referral rate.

Keywords: staffing; routing in queueing systems; call centres; gatekeeper systems.

Reference to this paper should be made as follows: Hasija, S., Pinker, E.J. and Shumsky, R.A. (2005) 'Staffing and routing in a two-tier call centre', *Int. J. Operational Research*, Vol. 1, Nos. 1/2, pp.8–29.

Biographical notes: Sameer Hasija is currently a PhD candidate in Operations Management at the Simon Business School, University of Rochester. He has a BTech (Major: Naval Architecture and Ocean Engineering, Minor: Industrial Engineering) from the Indian Institute of Technology at Madras and a MS (Management Science Methods) from the University of Rochester. His research interests include efficiency and flexibility issues in service systems, with a focus on Call-Center and Health Care Management.

Edieal J. Pinker is an Associate Professor of Computers and Information Systems at the Simon School of Business, University of Rochester. He conducts research on the use of contingent workforces, cross-training, and experience-based learning in service sector environments as it applies to work and workflow design. He also studies the use of online auctions in electronic commerce and the issues faced by legacy firms trying to transition into electronic commerce. Pinker has consulted for the United States Postal Service, the financial services industry, and the auto industry. His work has been published in *Management Science*, *Manufacturing and Service Operations Management*, the *European Journal of Operational Research*, *IIE Transactions* and the *Communications of the Association of Computing Machinery*. He serves on the editorial boards of *M&SOM*, *POMS*, *Decisions Sciences* and *IJOR*. Professor Pinker earned his MS and PhD in Operations Research from the Massachusetts Institute of Technology.

Robert A. Shumsky is an Associate Professor of Operations Management at the Simon School of Business, University of Rochester. Professor Shumsky has research and teaching interests in the modelling and control of service systems. Current research focuses on the dynamic use of flexible capacity, the use of incentives for operational control of service systems, and the application of revenue management under competition. His research has been published in *Management Science*, *Operations Research*, *Manufacturing & Service Operations Management (M&SOM)*, and *Air Traffic Control Quarterly*. He is an Associate Editor for *Management Science* and *Operations Research*, serves on the editorial review boards of *M&SOM* and the *Journal of Revenue and Pricing Management*, and is a Senior Editor for *Production and Operations Management*. He has also conducted research on the USA air traffic management system and studied transportation operations for the Massachusetts Port Authority and the Federal Aviation Administration. Professor Shumsky earned his MS and PhD in Operations Research from the Massachusetts Institute of Technology.

1 Introduction

In this paper we consider the problem of capacity planning and call routing in a two-tier service system in which the first tier consists of gatekeepers who diagnose the customer's problem, may solve the problem, or may refer the customer to an expert in the second tier. A typical example: call centres for health-care services are often staffed by certified nurses who diagnose the problem and provide advice. If justified by the nature and severity of the call, a nurse may refer a call to a specialist (Bernett 2003). Shumsky and Pinker (2003) provide additional examples of service systems with this two-tier architecture, but with the health-care motivation in mind, we refer to the resolution of a customer's problem as a 'successful treatment'.

Call centres must balance the relatively low cost of less-skilled gatekeepers with the benefits of the expert's ability to handle difficult calls. A manager of such a call centre must determine the staffing levels of gatekeepers and experts as well as the optimal referral rate to minimise total costs. These costs may include the cost of staffing, the 'cost' of customer waiting time, and mistreatment costs when a customer must see an expert for successful treatment despite spending time receiving (unsuccessful) treatment from the gatekeeper, as well.

This paper focusses on a staffing and referral strategy that is based only on call difficulty and steady state queue lengths and not on real-time queue lengths (the policies are static, not dynamic). We use a square-root staffing rule to approximate the optimal staffing for both tiers, given any particular referral rate. This rule is asymptotically optimal as system size increases. We then use the staffing approximation to determine the optimal referral rate and to compare the gatekeeper (or 'two-tier') system with a direct-access (or one-tier) system in which customers do not encounter a gatekeeper but instead immediately see an expert. Our main findings are:

- The choice of system (gatekeeper or direct access) has a complex relationship with the customer's waiting time cost. In particular, as the waiting cost increases, it may or may not be optimal to choose a direct-access system, depending on the other parameters such as the gatekeeper's skill level.
- If it is optimal to choose the direct-access system when the unit cost of a customer's waiting time is low (or in a deterministic system in which waiting costs are not assessed), then it is optimal to choose the direct-access system when the cost of the customers' waiting time is higher. However, this does not imply that a high waiting cost always leads to a one-tier system. As in point 1, it is possible to prefer a gatekeeper system, no matter how high the waiting cost per unit time.
- When two-tier systems are preferred, a simple, deterministic model can be used to choose the referral rate. The optimal referral rate converges to this 'deterministic' referral rate as the size of the system grows.
- If we construct an optimal staffing plan, given the deterministic referral rate, then we will only see a small increase in cost over the cost of the globally optimal system that uses the optimal referral rate.

In Section 2, we review the related literature. In Section 3 we introduce our model and in Section 4 we describe an approximation that generates asymptotically optimal staffing levels and referral rates as the system size increases. In Section 5 we use the approximation to characterise the behaviour of the referral rate as a variety of parameters change. In that section we also use numerical experiments to test the accuracy of the approximation and to confirm the four observations described above. Section 6 contains a discussion of our results and describes additional directions for research.

2 Literature review

This paper is related to the literature on capacity planning and routing in queuing networks. Halfin and Whitt (1981) establish heavy-traffic stochastic limits for multi-server queues in which the number of servers is allowed to increase along with the traffic intensity but the steady-state probability that all servers are busy is held fixed. Whitt (1992) shows that by using the square-root staffing principle, discussed below in Section 4, one can generate the same limiting regime as in Halfin and Whitt (1981). Borst et al. (2004) use a similar framework to approximate the waiting time distribution of an $M/M/N$ queue and demonstrate the asymptotic optimality of the square root staffing principle, given a cost function involving both waiting and staffing costs. We apply their approximation to a two-tier queuing network. We use the square-root staffing rule to find the number of servers for each level as a function of the routing strategy. We then determine the total cost of operating such a system and minimise that cost with the static routing strategy as the decision variable.

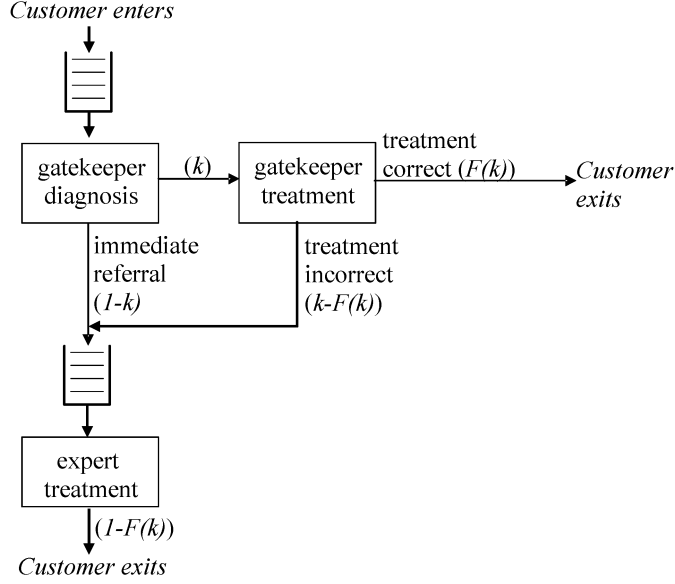
There exists a substantial literature on optimal routing strategies in call centres with cross-trained servers ('skill-based routing'). For example, Örmeci (2004) and Chevalier et al. (2004) study loss models with specialised and fully flexible servers. Wallace and Whitt (2004) examine systems with an arbitrary cross-training pattern (e.g., each server may be cross-trained in any subset of six skills). They use heuristics and simulation to find the minimum number of cross-trained servers needed to satisfy performance goals for each customer type. However, these models of skill-based routing differentiate calls by type, and not by difficulty level; a server is either sufficiently skilled to handle a call or is not. In our model, gatekeepers have some probability of success with a particular call. As in our model, de Véricourt and Zhou (2004) assume that each server has a different call resolution probability (p), and they also assume that each server may have a different service rate (μ). They identify the routing policy of calls to servers (a ' $p\mu$ rule') that minimises the total time a call spends in the system, including re-calls. While they assume that the staffing level is given – one server of each type – our model considers both the staffing and routing problem for large systems. The structure of our service system is also quite different. We assume that there are two pools of servers: the expert pool has a resolution probability equal to one and the gatekeeper pool attempts to treat calls or passes them along the expert pool.

Our model is closest to Shumsky and Pinker (2003). They determine the optimal routing strategy for a deterministic system and then formulate a principal/agent model to determine the impact of performance-based incentives on the gatekeeper's behaviour. Their model does not incorporate queuing effects, for they assume that the firm maintains a level of staffing sufficient to satisfy exogenous waiting time goals. Here we model queuing effects but we do not consider the incentive problem. We show how the cost-minimising referral rate varies with changes in parameters related to queuing, i.e., arrival rates and service rates. We also show that as the arrival rate increases, the optimal referral rate converges to the optimal referral rate for the deterministic case.

3 The queueing network model

In this section we describe an open queueing network model of a service centre with gatekeepers. The 'network' is essentially two queues in series: n_g gatekeepers and n_e experts, with staffing costs c_g and c_e per unit time, respectively (see Figure 1). Customers (or 'calls') arrive to the gatekeepers according to a Poisson process with rate λ . To the gatekeepers, the calls vary in difficulty and complexity, and we represent the difficulty of each call with a random draw from a uniform distribution, $U[0, 1]$. This random variable represents the call's percentile in a ranking of calls by treatment complexity. Given that a call has complexity x , the probability that the customer can be treated successfully by the gatekeeper is $f(x)$. As in Shumsky and Pinker (2003), we will refer to $f(x)$ as the 'treatment function'. Because complexity increases with x , we assume that $f'(x) \leq 0$.

Figure 1 Customer flows



With each new call, a gatekeeper spends time diagnosing the problem and determining the complexity (the value of x). The gatekeeper may then either send the call directly to the expert pool or attempt to solve the problem. If the gatekeeper successfully solves, or ‘treats’, the problem, the call leaves the system. If the gatekeeper attempts to treat and the treatment fails, we assess a cost m due to the inconvenience to the customer, and the call is sent to the expert pool. Once a call has reached an expert, it is served and leaves the system.

Both server pools have unlimited waiting space, and there is a cost w for each unit of time spent waiting. The time required for an expert to treat a call averages $1/\mu$. The time for a gatekeeper to diagnose a call averages $1/\mu_d$, while the average time to diagnose *and* treat is $1/\mu_t > 1/\mu_d$. If the gatekeeper follows a static policy and treats a proportion k of calls, then the gatekeeper’s service rate is,

$$\bar{\mu}(k) = \left(\frac{1-k}{\mu_d} + \frac{k}{\mu_t} \right)^{-1}.$$

We assume that service times are distributed as independent, exponential random variables, even when the gatekeeper only diagnoses some calls, and combines diagnosis with treatment in other calls. Given these assumptions, the gatekeeper and expert pools can each be modelled as $M/M/N$ queueing systems (see Gross and Harris, 1985, Section 4.1), where the arrival rate to the expert pool is the sum of the rate of calls untreated by the gatekeeper and the rate of calls mistreated by the gatekeeper. We will discuss additional implications of the exponential service-time assumption in Section 5.3.

Our objective is to minimise the sum of staffing, waiting, and mistreatment costs. Given the complexity of a call, we must decide whether the gatekeeper should treat the call or refer it immediately to an expert. Suppose that the staffing is fixed at (n_g, n_e) and the gatekeeper treats all calls in S , where S is a (possibly non-contiguous) subset of $[0, 1]$.

Let k be the proportion of the range $[0, 1]$ covered by S . If the gatekeeper replaces the set S with the set $[0, k]$, we know that

- The gatekeeper's service rate does not change because the proportion k does not change.
- The rate of untreated calls does not change.
- The rate of mistreated calls and the waiting time stays the same or decrease because $f'(x) \leq 0$.

Therefore, given any staffing configuration and treatment set S with proportion k , the waiting, staffing, and mistreatment costs will not increase if the gatekeeper instead treats calls in $[0, k]$. This argument indicates that the optimal treatment set S takes the form $[0, k]$, and we will refer to k as the 'treatment threshold'. Given treatment threshold k , the gatekeeper refers a proportion $1 - k$ of calls without attempting treatment. The expected fraction of calls treated successfully by a gatekeeper is $F(k) = \int_0^k f(x)dx$, the fraction mistreated is $k - F(k)$, and the fraction of calls seen by the expert pool is $1 - F(k)$.

We now develop the objective function for our problem. The decision variables are k , the proportion of calls treated by the gatekeeper, and the staffing levels n_g and n_e . Let $q(n, \lambda, \mu)$ be the expected wait for an $M/M/N$ queueing system with n servers, arrival rate λ , and service rate μ . The total cost per unit time is:

$$C_2(n_g, n_e, k) = w\lambda [q(n_g, \lambda, \bar{\mu}) + (1 - F(k))q(n_e, (1 - F(k))\lambda\mu)] + c_g n_g + c_e n_e + m\lambda(k - F(k)). \quad (2)$$

The subscript 2 indicates that this is a cost function for a two-tier service system (as opposed to the one-tier direct-access system described below). The first term is the expected cost of waiting in front of the gatekeeper and expert pools, the second and third terms are the staffing costs, and the last term is the mistreatment cost. Therefore, we consider the following problem:

$$\min_{n_g, n_e, k} C_2(n_g, n_e, k) \quad (3)$$

subject to

$$n_g > \lambda / \bar{\mu} \quad (4)$$

$$n_e > \lambda(1 - F(k)) / \mu. \quad (5)$$

The constraints ensure that the gatekeeper and expert pools are stable.

In the following sections we will be comparing this two-tier system with an all-expert system in which customers do not see gatekeepers, but instead go directly for treatment in the expert server pool. This one-tier system is simply an $M/M/N$ system with n_e servers, arrival rate λ and service rate μ . Therefore, the total cost is,

$$C_1(n_e) = w\lambda q(n_e, \lambda, \mu) + c_e n_e. \quad (6)$$

While we can numerically find a routing policy and staffing levels that minimise C_1 and C_2 , in the next section we will use a square-root staffing heuristic that will

- 1 allow us to solve these problems quickly
- 2 enable us to characterise the effects of certain parameters on the optimal solution
- 3 allow for direct comparison between the one-tier and two-tier systems.

4 An approximation using a square-root staffing rule

In both the one and two-tier systems, each server pool is an $M/M/N$ queue with linear staffing and waiting-time costs. Borst et al. (2004) demonstrate that when the number of servers is adjusted to minimise total staffing and waiting costs, and when we allow $\lambda \rightarrow \infty$, the ratio of staffing and waiting costs is bounded. Such a system is described as being in the ‘rationalised regime’. Building on the work of Whitt, Borst et al. (2004) also show that for systems in the rationalised regime, a simple square-root staffing heuristic is asymptotically optimal as $\lambda \rightarrow \infty$. In this section we describe the heuristic and apply it to our system. In Section 4.1 we show how the heuristic can be used to generate both near-optimal staffing levels and an approximation of the total cost function for a single-server pool. Section 4.1 is primarily a summary of the work of Borst et al., and these results can be applied directly to the one-tier system. In Section 4.2 we apply these staffing results to the two-tier system, so that the routing and staffing problem reduces to a single-variable optimisation in the treatment threshold, k .

4.1 Approximation for an $M/M/N$ queue

Consider an $M/M/N$ queue with load $\rho = \lambda/\mu$, staffing cost c per unit time and waiting cost w per unit time. Borst et al. show that staffing n^h servers according to the following square-root rule is asymptotically optimal (the superscript ‘ h ’ refers to either ‘heuristic’ or ‘Halfin-Whitt’):

$$n^h = \rho + y^*(c, w)\sqrt{\rho}. \quad (7)$$

At least ρ agents are needed to guarantee stability and $y^*(c, w)\sqrt{\rho}$ is the safety staffing for protection against stochastic variability. The quantity $y^*(c, w)$ can be thought of as the optimal service level and is found by balancing the staffing and waiting costs. Specifically, $y^*(c, w)$ minimises the function

$$\alpha(y, c, w) = cy + \frac{w\pi(y)}{y}, \quad (8)$$

where

$$\pi(y) = \left[1 + \frac{y\Phi(y)}{\phi(y)} \right]^{-1} \quad (9)$$

and $\phi(y)$ and $\Phi(y)$ are the unit normal pdf and cdf, respectively. That is,

$$y^*(c, w) = \arg \min_{y>0} \alpha(y, c, w) \quad (10)$$

Because the function $\alpha(y, c, w)$ has a finite, unique, and positive minimum value, $y^*(c, w)$ can be found quickly by numerical methods.

The function $\pi(y)$ has an important interpretation that will be useful for constructing the approximate cost function. It is known as the Halfin–Whitt delay function, and it is an asymptotically exact approximation of the probability of delay, $Pr\{\text{wait} > 0\}$, for the $M/M/N$ queue. Let $D(\rho, c, w)$ be an approximation for the total cost per unit time of staffing and waiting under the rationalised regime, given load ρ , unit staffing cost c , and unit waiting cost w . Given that $\pi(y)$ is the approximation for the probability of delay under the rationalised regime,

$$\begin{aligned} D(\rho, c, w) &= cn^h + \frac{w\lambda\pi(y^*(c, w))}{n^h\mu - \lambda} \\ &= c\rho + \alpha(y^*(c, w), c, w)\sqrt{\rho}. \end{aligned} \quad (11)$$

The second expression follows by substitution for n^h and the definition of α .

Because the one-tier system is simply an $M/M/N$ queue, our approximation of the optimal total cost for this direct-access system is

$$\widehat{C}_1 = D(\lambda / \mu, c_e, w) \quad (12)$$

where λ is the arrival rate to the system, μ is the service rate of the experts, c_e is the cost of experts per unit time, and w is the waiting cost per unit time.

In numerical experiments, Borst et al. show that this staffing heuristic is remarkably robust, even for offered loads as low as 10. We present similar results in our numerical experiments (See Section 5.3), and we also find that using the approximate total cost function $D(\rho, c, w)$ allows us to find near-optimal solutions to the staffing and routing problem in the two-tier system with gatekeepers.

4.2 Approximation for a two-tier system

Given the size of the load to each server pool in the two-tier system, we use the square-root staffing heuristic to determine the optimal number of servers for that pool. In the two-tier system, the choice of the treatment threshold k determines the arrival and service rates of the gatekeeper and expert pools, and therefore determines the load for each pool. Specifically, the load for the gatekeeper pool, $\rho_g(k) = \lambda/\bar{\mu}(k)$ and the load for the expert pool, $\rho_e(k) = (1 - F(k))\lambda/\mu$. Therefore, for a given k , the number of servers in each pool is,

$$n_i^h = \rho_i(k) + y^*(c_i, w)\sqrt{\rho_i(k)}, \quad i = g, e, \quad (13)$$

and our approximation for the total cost of the two-tier system is

$$\widehat{C}_2(k) = D(\rho_g(k), c_g, w) + D(\rho_e(k), c_e, w) + m\lambda(k - F(k)). \quad (14)$$

Because the square-root staffing rule specifies the number of servers in each pool, k is the remaining decision variable, and our problem is to find the cost-minimising value of k

$$k^h = \arg \min_{0 \leq k \leq 1} \widehat{C}_2(k) \quad (15)$$

and to compare the optimal two-tier cost $\widehat{C}_2(k)$ with the one-tier cost, \widehat{C}_1 . While k^h , $\widehat{C}_2(k^h)$, and \widehat{C}_1 are approximations, in numerical experiments we will see that these approximations follow closely the optimal values derived from a more realistic model (This alternate model relaxes both the ‘asymptotic’ assumptions of the rationalised regime and the markovian assumptions of the original network model presented in Section 3).

For an arbitrary treatment function $f(k)$, $\widehat{C}_2(k)$ can take an arbitrary form, e.g., it need not be unimodal. In the next section we assume that the treatment function is linear. Working with these approximations, and with a linear treatment function, allows us to analytically characterise the behaviour of the (approximately) optimal treatment threshold and to quickly identify relative advantages of the one-tier and two-tier systems as the system parameters change.

5 Analysis and numerical experiments with a linear treatment function

Now assume that $f(k)$ belongs to a class of linear functions, $f(k) = b(1 - k)$, where $b \in [0,1]$. With this treatment function, gatekeepers have a positive chance to successfully treat all calls, although the probability approaches 0 for the most difficult calls. Parameter b is a measure of the gatekeeper’s skill: as b rises, the gatekeeper has a greater chance to handle all calls. For analytical tractability, we have chosen a functional form for f so that the vertical intercept and slope are both equal to b . A byproduct of this choice is that as b increases, the implied variance in call difficulty to the gatekeeper increases as well.

For brevity, throughout this section we use the following notation:

$$\rho = \lambda / \mu$$

$$\rho_i = \lambda / \mu_i$$

$$\rho_d = \lambda / \mu_d$$

$$y_i = y^*(c_i, w) \quad \text{for } i = g, e$$

$$\alpha_i = \alpha(y_i, c_i, w) \quad \text{for } i = g, e$$

The following Proposition states that the total cost functions $\widehat{C}_2(k^h)$ and \widehat{C}_1 are minimised with a single, optimal system design and treatment threshold. All proofs are in the appendix.

Proposition 1: *When minimising $\widehat{C}_2(k)$ to find k^h , and when comparing \widehat{C}_1 with $\widehat{C}_2(k^h)$, there are two possible outcomes:*

- 1 *a two-tier system with a unique treatment threshold k^h is optimal*
- 2 *a one-tier system is optimal.*

A comparison of $\widehat{C}_2(k^h)$ and \widehat{C}_1 also demonstrates that a two-tier system is favored when parameters c_g , m , and μ are low, and when c_e , μ_d and μ_t are high.

Before considering how k^h changes as the parameters change, it is convenient to introduce a simple, deterministic model and the deterministic treatment threshold, k^d .

5.1 The deterministic model

Consider a deterministic model of the two-tier system with no stochastic variability in the arrival or service rates, so that the capacity of the gatekeeper and expert pools are set equal to the load. Given the linear treatment function $f(k)$, the total cost of this system is

$$C_2^d(k) = c_g \lambda \left(\frac{1-k}{\mu_d} + \frac{k}{\mu_t} \right) + c_e \frac{\lambda(1-bk + bk^2/2)}{\mu} + m\lambda(k - bk + bk^2/2) \quad (16)$$

and the optimal treatment threshold is,

$$k^d = \left[1 - \frac{1}{b} \frac{m + c_g(1/\mu_t - 1/\mu_d)}{m + c_e(1/\mu)} \right]^+ \quad (17)$$

Note that k^d is equivalent to the optimal treatment threshold for the model in Shumsky and Pinker (2003), which also focuses on a deterministic gatekeeper system.

A one-tier deterministic model has total cost,

$$C_1^d = c_e \frac{\lambda}{\mu} \quad (18)$$

The quantity k^d will be useful in the following analysis and will also be useful for generating a simple ‘rule of thumb’ for the system design in the numerical experiments.

5.2 Analysis of the optimal treatment threshold

Here we examine how the optimal treatment threshold is affected by the system’s parameters. In this section, we limit our attention to cases where both $0 < k^h < 1$ and $0 < k^d < 1$. The proof of Proposition 1 demonstrated that when k^h is an interior solution,

$$\partial^2 \widehat{C}_2(k) / \partial k^2 > 0.$$

By using this fact, and applying the implicit function theorem to $\widehat{C}_2(k^h)$, we find:

$$\begin{aligned} \partial k^h / \partial c_g < 0, \quad \partial k^h / \partial c_e > 0, \quad \partial k^h / \partial m < 0, \quad \partial k^h / \partial \mu_t > 0, \quad \partial k^h / \partial \mu_d < 0, \\ \partial k^h / \partial \mu < 0 \quad \text{and} \quad \partial k^h / \partial b > 0. \end{aligned}$$

Therefore, for large values of c_g , m , μ_d , μ and small values of c_e , μ_t , b , it is optimal for gatekeepers to treat only the less difficult calls.

The impact of the arrival rate λ and the waiting cost w is more complex. First, we consider λ . We find that as λ increases, k^h can either fall or rise, and that it monotonically converges to k^d .

Proposition 2:

- 1 *If $k^h \geq k^d$ then $\partial k^h / \partial \lambda \leq 0$*
- 2 *If $k^h < k^d$ then $\partial k^h / \partial \lambda > 0$*
- 3 *$k^h \rightarrow k^d$ as $\lambda \rightarrow \infty$.*

Figures 2 and 3 show convergence from above and below, respectively. Convergence to k^d has an intuitive explanation: for very large λ , waiting costs are relatively small, compared to the sum of staffing and mistreatment costs. Therefore, for very large λ it is optimal to use the treatment threshold from the deterministic model, which only considers staffing and mistreatment costs.

Figure 2 Treatment threshold k vs. λ . Other parameters are $\mu_t = 0.75, \mu_d = 5, \mu = 1, c_g = 1, c_e = 4, m = 1, b = 1, w = 5$

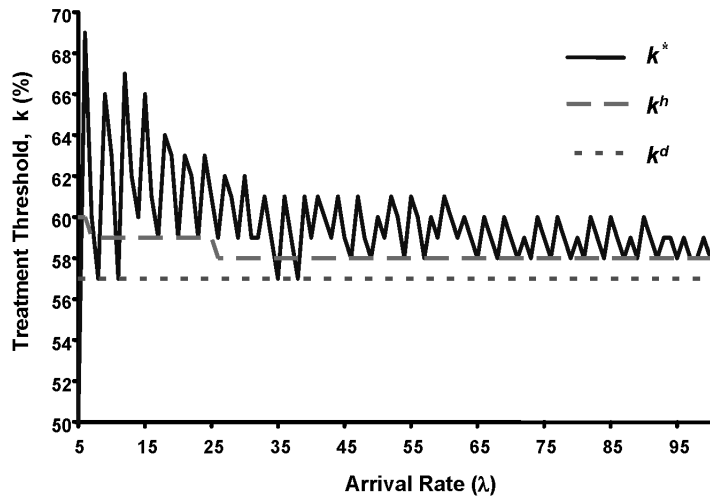
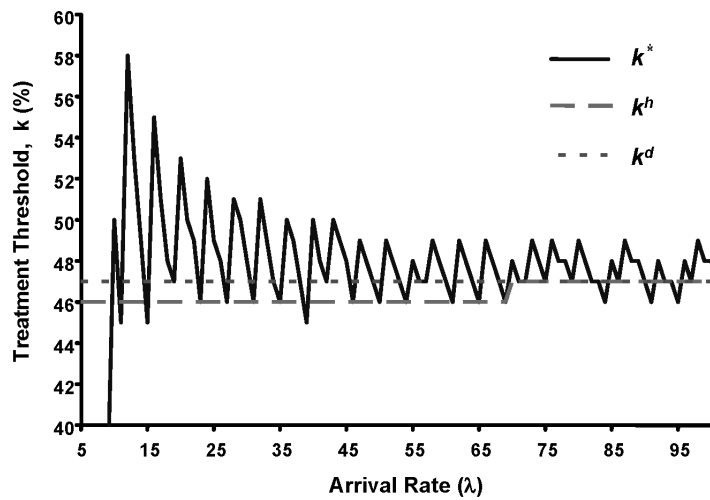


Figure 3 Treatment threshold k vs. λ . Other parameters are the same as for Figure 2 except $b = 0.8$ and $w = 0.5$



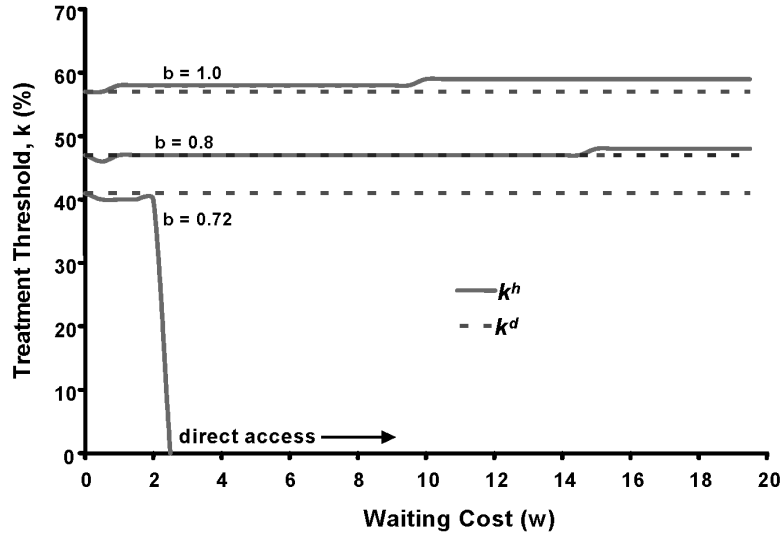
To understand the impact of w on the optimal treatment threshold, it is useful to examine the expression for the partial derivative of k^h with respect to w :

$$\frac{\partial k^h}{\partial w} = \left[\frac{\sqrt{\rho}b(1-k^h)}{2\sqrt{1-bk^h+b(k^h)^2/2}} \frac{\pi(y_e)}{y_e} - \frac{(\rho_t-\rho_d)}{2\sqrt{\rho_d+k^h(\rho_t-\rho_d)}} \frac{\pi(y_g)}{y_g} \right] \left[\frac{\partial^2 \widehat{C}_2(k^h)}{\partial (k^h)^2} \right]^{-1}. \quad (19)$$

Given that k^h is an interior solution, the denominator is positive. Therefore, the sign of the derivative depends upon the sign of the numerator. If the numerator is multiplied by w , then the first term is the marginal decline in the cost of waiting at the expert queue as k^h increases. The second term is the marginal cost of waiting at the gatekeeper queue as k^h increases. Therefore, if the marginal cost at the gatekeeper queue is lower, then k^h rises with w , shifting some of the workload to the gatekeepers and reducing the expert queue.

Expression 19 allows us to see how the parameters affect the relationship between w and k^h . For example, if b is high and the gatekeeper is skilled, the first term in the numerator dominates, and k^h rises as w rises. On the other hand, if $\rho_t - \rho_d$ is large, implying that treatment by the gatekeepers is slow, then the second term dominates and k^h falls as w rises. Figure 4 shows how k^h changes with w for three different gatekeeper skill levels (other parameters are the same as for Figure 2).

Figure 4 Treatment threshold k vs. w for $b = 0.72, 0.8,$ and 1



Intuitively, as w rises, queueing economies of scale become more important. These economies of scale imply that it is more efficient to have one large and one small server pool, rather than two pools that are closer in size. If the parameters give an advantage to the gatekeepers, then a rising w implies a rise in k^h , expanding the ranks of the gatekeepers and reducing the expert pool. If the parameters give an advantage to the experts, then rising w implies that pooling should occur on the expert level, dropping k^h and eventually producing a one-tier system. The following proposition shows that this effect is monotonic if k^h is above k^d .

Proposition 3: *If $k^h > k^d$, then $\partial k^h / \partial w > 0$.*

In the next section we will see numerical examples of these effects, and we will compare the one and two-tier systems under a variety of system parameters.

5.3 Numerical experiments

In this section we demonstrate the accuracy of the approximation described above, and investigate how the optimal design of the service centre is influenced by the parameter values. In particular, we see numerically how the optimal treatment threshold changes, and we compare one-tier and two-tier systems under a variety of scenarios. We also show that the treatment threshold derived from the deterministic model, k^d , is an excellent approximation for the optimal treatment threshold in stochastic systems, as long as it is optimal to use a two-tier, rather than a one-tier, system.

Recall that in the model introduced in Section 3, we assume that the gatekeeper's service time is exponentially distributed with mean $\bar{\mu}(k)$. However, the gatekeeper's actual service time is a mixture of time spent only diagnosing a customer and time spent both diagnosing and treating. Because these two types of services may have significantly different average times, a more accurate model would use a mixture of two exponential service times: a proportion k with mean $1/\mu_t$ and $1 - k$ with mean $1/\mu_d$. Given that the gatekeeper's service times follow such a distribution, we model the gatekeeper pool as an $M/H_2/N$ queue and the expert pool as a $G/M/N$ queue. The arrival process to the expert pool is difficult to characterise, and we use the approximation suggested by Adan (2004).

In this section, we compare our heuristic solution, using the square-root staffing rule, with the optimal solution determined by numerically solving a model based on the more general queueing systems described above. We use a software package that implements the $G/G/N$ approximations by Whitt (1993) to find the optimal combination of k^* , n_g^* , n_e^* that minimises the total staffing, waiting, and mistreatment cost. Henceforth we will call this solution the 'optimal' staffing and routing strategy and we will call the values k^h , n_g^h , n_e^h , determined by equations (7), (10), and (15) the 'heuristic' staffing and routing strategy.

We first verify the accuracy of the square-root staffing rule in the two-tier setting. The accuracy of this approximation will be driven by the sensitivity of the results to the assumption that the gatekeeper's service times are exponential and that arrivals to the experts are Poisson. Therefore, the larger the difference between μ_d and μ_t , the worse the performance of the heuristic. However, we find that even with extremely large differences (μ_d/μ_t as large as 100), the cost of a system operated according to the heuristic is within 1% of the optimal cost, as long as $\lambda > 20$. We also observe that the referral rates and staffing levels generated from the heuristic converge quickly to the optimal levels as λ increases. In this section, we will focus on more reasonable examples than $\mu_d/\mu_t = 100$; we set $\mu_t = 0.75$ and $\mu_d = 5$, while varying other parameters, such as the skill level b and the waiting cost w . For example, Figure 2 shows how k^* converges to k^h as the arrival rate increases, and how k^h converges to k^d from above, as implied by Proposition 2. Most of the variation of k^* around k^h is due to the integrality of n_g^* and n_e^* . Figure 3 shows a similar pattern, although with a lower value of b and w , and here k^h converges to k^d from below. In Figure 5, we see that using the heuristic does not significantly increase system

costs, given large λ , and in Figure 6 we see that the staffing levels n_g^* , n_e^* and n_g^h, n_e^h are nearly identical. (Figures 5 and 6 use the set of parameters that led to Figure 2.)

Figure 5 Percentage cost penalty for using heuristic solution rather than the optimal solution

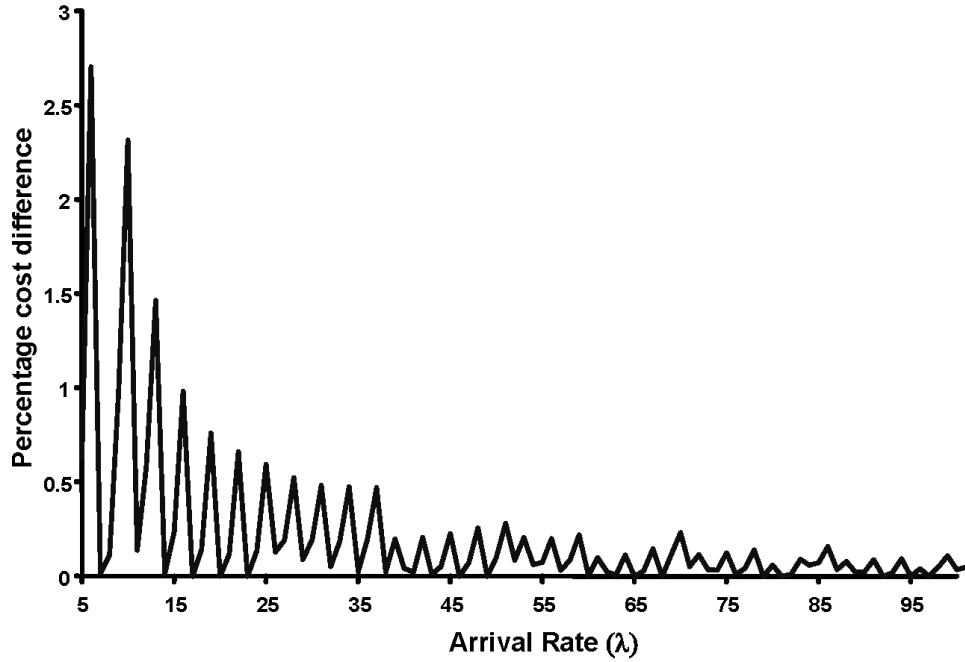
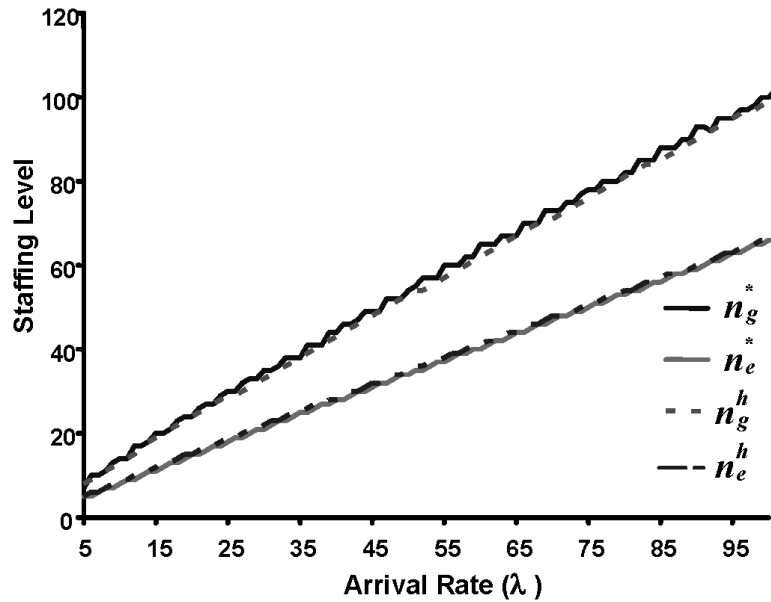


Figure 6 Staffing levels vs. λ



In all remaining figures, we do not show the optimal solution, but in every case the heuristic and optimal solutions are nearly identical, and the difference in total cost when using each is negligible. We will also be using as a baseline the parameter values for the system shown in Figures 2, 5 and 6. While we will only be presenting a subset of our experiments, we observed that the heuristic solution was nearly optimal over a wide range of parameter values for labor and waiting costs, service times, and gatekeeper skills.

Figure 7 plots the staffing levels of each server pool as a function of b , a measure of the gatekeepers' skills. Below a certain skill level a direct-access system is optimal and above that skill level a two-tier system is optimal. As the skill level continues to increase, the gatekeeper pool grows and the expert pool shrinks.

Figure 7 Staffing levels vs. b

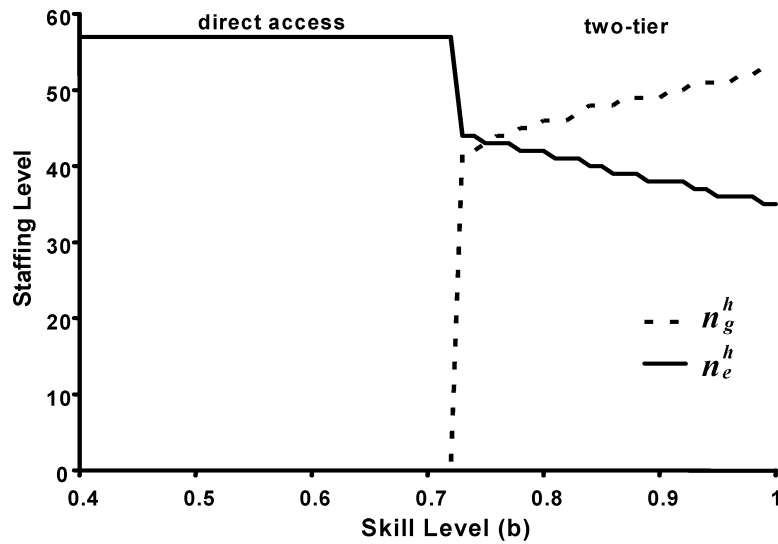


Figure 8 shows k^h and k^d as a function of the skill level. The steep, initial increase in each threshold represents the transition from one to two-tier systems; note that k^d rises at a lower value of b than k^h . We consistently observed this phenomenon in all numerical experiments we conducted. To understand why k^d should rise before k^h , it is useful to interpret k^d as the optimal treatment threshold when w is extremely low (a deterministic system essentially ignores waiting costs). To justify having gatekeepers in systems with high waiting costs, they must have higher skills than needed to justify gatekeepers in systems with lower waiting costs. In other words, if it is optimal to choose the direct-access system when w is very low (as in the deterministic system), then it is optimal to choose the direct-access system when the customer's cost of waiting is higher. This effect can be explained by the fact that a one-tier system offers benefits from pooling and that these benefits are more powerful when waiting costs are high. However, this does not imply that a high waiting cost always leads to a one-tier system. As in Proposition 3, it is possible to prefer a gatekeeper system, no matter how high the value of w (we saw an example of that in Figure 4).

Figure 8 Treatment threshold k vs. b

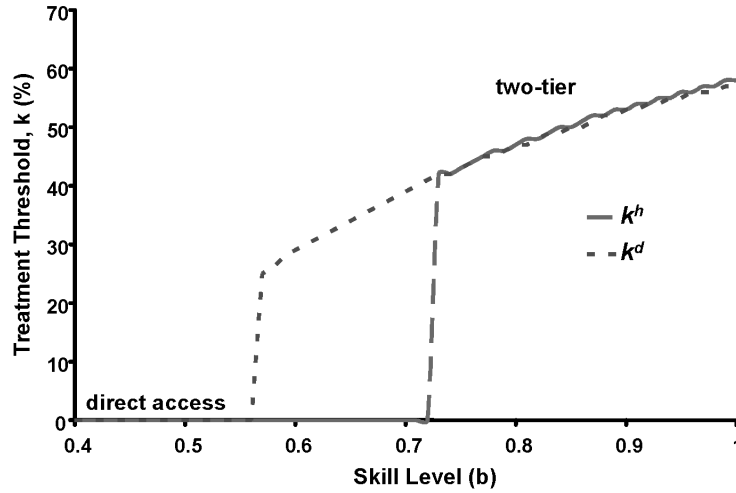
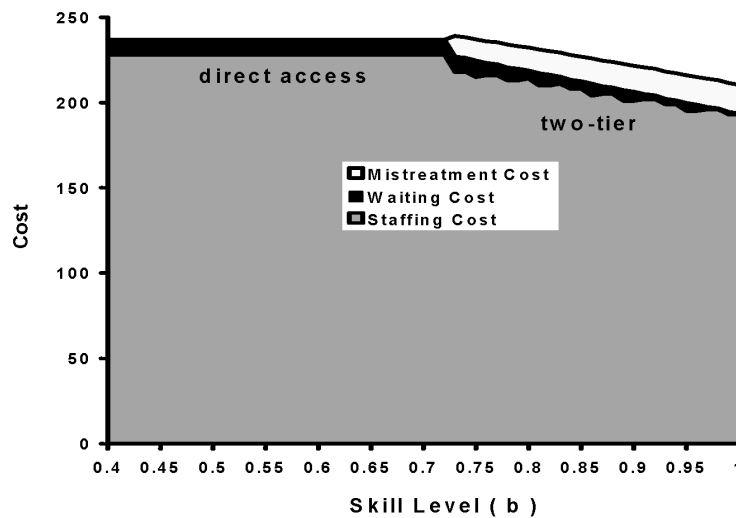


Figure 9 shows the contribution of mistreatment, waiting, and staffing costs to the total cost as a function of gatekeeper skill. This plot shows the ‘actual’ costs, calculated from the $G/G/N$ models, given the heuristic solution. The staffing cost decreases as b increases because we substitute gatekeepers for experts, as seen in Figure 7. An increase in b coupled with an increase in k^h also implies that a higher fraction of calls leave the system after successful treatment by the gatekeeper, thus reducing the queues and the waiting time in the system. It is somewhat counterintuitive that total mistreatment cost increases with b . On the one hand, increasing b reduces the probability that gatekeepers mistreat each call that they address. On the other hand, increasing k^h increases the number of calls treated by the gatekeeper, and therefore increases the mistreatment rate. In all of our experiments, we observed that the second effect dominates the first, so that rising b always increases total mistreatment costs.

Figure 9 Cost of staffing, waiting and mistreatment for the heuristic solution vs. b



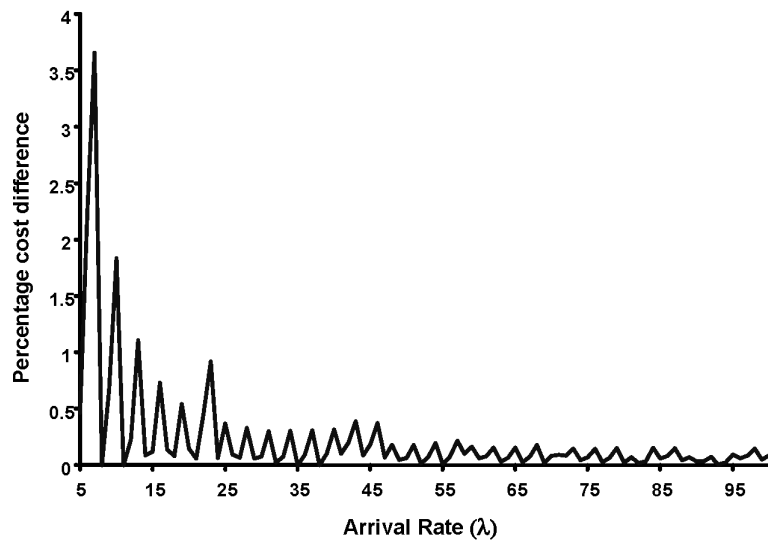
In Section 5, we saw that the response of the optimal treatment threshold to changes in w is complex. In Figure 4 we plot k^h and k^d for different values of b as a function of waiting cost. We see that for high values of b , k^h increases with waiting cost while for low values of b the optimal treatment threshold decreases until a direct-access system is preferable. As we discussed in Section 5, the optimal location to pool resources as w increases depends upon the skill level of the gatekeepers and the cost parameters.

In all of our experiments, we noticed that when a two-tier system is preferred to a direct-access system, k^h and k^d are remarkably close (see, for example, Figures 2–4, 8). Therefore using k^d as an estimate of the optimal treatment threshold does not increase total costs significantly (see Figure 10). However, the choice of a one or two-tier system should be based on a cost comparison that takes optimal staffing and waiting costs into account, as was seen in Figures 4 and 8. Therefore, we propose the following rule of thumb for choosing the optimal system:

- Calculate k^d using equation (17).
- Using k^d as the treatment threshold, use the square-root staffing rule to determine the number of gatekeepers and experts in a two-tier system. Given these staffing levels, calculate the total cost $\hat{C}_2(k^d)$. Also using the square-root staffing rule, determine the number of experts in the direct-access system and calculate the cost \hat{C}_1 .
- If $\hat{C}_2(k^d) < \hat{C}_1$, choose a two-tier system using k^d as the treatment threshold. Otherwise, choose a direct-access system.

This rule of thumb does not require managers to find k^* or k^h , both of which require significant computational effort compared to finding k^d .

Figure 10 Percentage cost penalty for using k^d and the square-root staffing rule rather than the optimal solution



6 Conclusions

In practice most call centres have multiple tiers, where the tiers are distinguished by their abilities to serve the customers. Differing abilities typically imply differing compensation rates and service rates as well. Managers must determine staffing at each tier in conjunction with routing rules to balance customer queueing delay costs, mistreatment costs and staffing costs. In this paper we have developed an approach that greatly simplifies this complex managerial problem. By drawing upon recent results showing the asymptotic optimality of square-root staffing rules for stand-alone queues, we have shown that the optimal design of a two-tier system can be reduced to determining an optimal routing rule. Further, we have shown that the easily computed routing rule from a deterministic system can be used whenever a two-tier system is preferred to a one-tier system. It is well known in the queueing literature that pooling resources can create economic benefits by reducing variability. In a two-tier system in which the second tier is staffed with higher skilled and more expensive servers, it is not clear how to take advantage of pooling. We find that when waiting costs are higher, gatekeepers need a higher skill level to be worthwhile. That is, pooling economies are achieved using the experts only. However, we also see that if the gatekeepers' skills are high enough, it is optimal to achieve pooling economies at the first-tier for even very high values of the waiting costs, w . So we see that depending on the combination of (b, w) we may seek pooling economies at different locations in the system. Much of our analysis was restricted to the case of a linear treatment function. Further research is necessary to test the validity of our results for more general treatment functions. Other possible areas for future research include extending the model and analysis to three or more tiers of servers, considering dynamic routing policies, and incorporating incentive systems for controlling gatekeeper referral behaviour into the model, as is done in Shumsky and Pinker (2003).

Acknowledgement

We would like to thank Harry Groenevelt for supplying us with his software package, 'QMacros,' and for patiently answering our questions about the software. QMacros includes an implementation of the $G/G/N$ approximation proposed by Whitt (1993), and we used the software for the numerical experiments in Section 5.

References

- Adan, I. (2004) *Teaching Note on Multi-Machine Systems*, available at <http://www.win.tue.nl/~iadan/sdp/h11.pdf>.
- Bennett, H.(2003) 'Healthcare call centers: a technology migration', *horizons, Perspectives in Healthcare Management and Information Technology*, September, pp.17–20.
- Borst, S., Mandelbaum, A. and Reiman, M.I. (2004) 'Dimensioning large call centers', *Operations Research*, Vol. 52, No. 1, pp.17–34.
- Chevalier, P., Shumsky, R.A. and Tabordon, N. (2004) *Routing and Staffing in Large Call Centers with Specialized and Fully Flexible Servers*, working paper, Simon School, University of Rochester, Rochester, NY.

- de Véricourt, F. and Zhou, Y.-P. (2004) *A Routing Problem for Call Centers with Customer Callbacks after Service Failure*, Working Paper, Fuqua School of Business, Duke University, Durham, North Carolina.
- Gross, D. and C.M. Harris (1985) *Fundamentals of Queueing Theory, Second Edition*, Wiley, New York.
- Halfin, S. and Whitt, W. (1981) ‘Heavy-traffic limits for queues with many exponential servers’, *Operations Research*, Vol. 29, No. 3, pp.567–587.
- Örmeci, E.L. (2004) ‘Dynamic admission control in a call center with one shared and two dedicated service facilities’, *IEEE Transactions on Automatic Control*, Vol. 49, No. 7, pp.1157–1161.
- Shumsky, R.A., Pinker, E.J. (2003) ‘Gatekeepers and referrals in service’, *Management Science*, Vol. 49, No. 7, pp.839–856.
- Whitt, W. (1992) ‘Understanding the efficiency of multi-server service systems’, *Management Science*, Vol. 38, No. 5, pp.708–723.
- Whitt, W. (1993) ‘Approximations for the GI/G/m queue’, *Production and Operations Management*, Vol. 2, No. 2, pp.114–161.
- Wallace, R.B. and Whitt, W. (2004) *Resource Pooling and Staffing in Call Centers with Skill-Based Routing*, Working Paper, Columbia University, <http://www.columbia.edu/~ww2040/pooling2.pdf>.

Appendix: Proofs

Proof of Proposition 1:

For the given treatment function,

$$\begin{aligned} \frac{\partial \widehat{C}_2(k)}{\partial k} = & c_g(\rho_t - \rho_d) + \frac{(\rho_t - \rho_d)\alpha_g}{2\sqrt{\rho_d + k(\rho_t - \rho_d)}} - \rho c_e b(1-k) \\ & - \frac{\sqrt{\rho b(1-k)}\alpha_e}{2\sqrt{1-bk + bk^2/2}} + m\lambda(1-b + bk). \end{aligned} \quad (20)$$

and,

$$\frac{\partial^2 \widehat{C}_2(k)}{\partial k^2} = -\frac{(\rho_t - \rho_d)^2 \alpha_g}{4[\rho_d + k(\rho_t - \rho_d)]^{3/2}} + \rho c_e b + \frac{\sqrt{\rho b(2-b)}\alpha_e}{4[1-bk + bk^2/2]^{3/2}} + m\lambda b. \quad (21)$$

The total cost functions have the following properties:

P1: $\left. \frac{\partial \widehat{C}_2(k)}{\partial k} \right|_{k=1} > 0$,

P2: $\frac{\partial^2 \widehat{C}_2(k)}{\partial k^2}$ is an increasing function in k ,

P3: The cost of staffing no gatekeepers is less than the cost of staffing gatekeepers who only do a diagnosis of the incoming calls. The cost of having no gatekeepers is given by,

$$\widehat{C}_1 = \rho c_e + \sqrt{\rho}\alpha_e. \quad (22)$$

Explore all possible cases. For each case, we see that either the direct-access system is optimal, or the two-tier system has a unique cost-minimising solution, k^h .

$$\text{I} \quad \frac{\partial^2 \widehat{C}_2(k)}{\partial k^2} \Big|_{k=0} > 0.$$

From P2 $\frac{\partial^2 \widehat{C}_2(k)}{\partial k^2} > 0$ for all values of k , $\Rightarrow \widehat{C}_2(k)$ is convex in the domain $k \in [0, 1]$.

From P1 observe that two subcases are possible.

- $\frac{\partial \widehat{C}_2(k)}{\partial k} \Big|_{k=0} > 0$. Here, $\frac{\partial \widehat{C}_2(k)}{\partial k} = 0$ has no root on the interval $[0, 1]$. In this case it is optimal for the centre to staff only experts.
- $\frac{\partial \widehat{C}_2(k)}{\partial k} \Big|_{k=0} < 0$. Here, $\frac{\partial \widehat{C}_2(k)}{\partial k} = 0$ has one root in the interval $[0, 1]$. k^h is the unique point which minimises $\widehat{C}_2(k)$. If $\widehat{C}_2(k^h) < \widehat{C}_1$, then staff generalists who treat incoming calls of difficulty level $< k^h$, else only staff specialists.

$$\text{II} \quad \frac{\partial^2 \widehat{C}_2(k)}{\partial k^2} \Big|_{k=0} < 0 \quad \text{and} \quad \frac{\partial^2 \widehat{C}_2(k)}{\partial k^2} \Big|_{k=1} > 0.$$

P2 $\Rightarrow \exists \hat{k} \in (0, 1)$ such that $\widehat{C}_2(k)$ is concave for $k < \hat{k}$ and is convex for $k > \hat{k}$. There will be four subcases here:

- No root for $\frac{\partial \widehat{C}_2(k)}{\partial k} = 0$ in the interval $[0, 1]$. It is optimal to staff only experts in this case.
- One root for $\frac{\partial \widehat{C}_2(k)}{\partial k} = 0$ in the interval $[0, 1]$. It is optimal to staff only experts in this case.
- $\frac{\partial \widehat{C}_2(k)}{\partial k} \Big|_{k=0} < 0$. This case will also have one root in $[0, 1]$ for $\frac{\partial \widehat{C}_2(k)}{\partial k} = 0$. Compare the total cost at that root (k^h) with \widehat{C}_1 to determine which system is optimal.
- $\frac{\partial \widehat{C}_2(k)}{\partial k} = 0$ has two roots in $[0, 1]$. This case is shown in Figures 4 and 9. Compare the total cost at the larger root (k^h) with \widehat{C}_1 .

III $\frac{\partial^2 \widehat{C}_2(k)}{\partial k^2} \Big|_{k=0} < 0$ and $\frac{\partial^2 \widehat{C}_2(k)}{\partial k^2} \Big|_{k=1} < 0$ Therefore, $\widehat{C}_2(k)$ is concave in the range $k \in [0, 1]$.

Here $\frac{\partial \widehat{C}_2(k)}{\partial k} = 0$ has no roots in $[0, 1]$ and it is optimal to staff the centre with only experts.

Proof of Proposition 2:

Using implicit differentiation with the first order condition $\partial \widehat{C}_2(k) / \partial k = 0$ produces

$$\partial k^h / \partial \lambda = -A / B \quad (23)$$

where

$$A = \left(\frac{1}{\mu_t} - \frac{1}{\mu_d} \right) c_g + \frac{((1/\mu_t) - (1/\mu_d)) \alpha_g}{2\sqrt{1/\mu_d + k^h} ((1/\mu_t) - (1/\mu_d))} \frac{1}{2\sqrt{\lambda}} - \frac{1}{\mu} c_e b(1 - k^h) - \frac{b(1 - k^h) \alpha_e}{2\sqrt{1 - bk^h + b(k^h)^2 / 2}} \frac{1}{\sqrt{\mu}} \frac{1}{2\sqrt{\lambda}} + m(1 - b + bk^h) \quad (24)$$

and

$$B = -\frac{(\rho_t - \rho_d)^2 \alpha_g}{4[\rho_d + k^h(\rho_t - \rho_d)]^{3/2}} + \rho c_e b + \frac{\sqrt{\rho} b(2 - b) \alpha_e}{4[1 - bk + b(k^h)^2 / 2]^{3/2}} + m\lambda b.$$

Substituting terms in A ,

$$2\lambda \frac{\partial k^h}{\partial \lambda} = -[c_g(\rho_t - \rho_d) - \rho c_e b(1 - k^h) + m\lambda(1 - b + bk^h)] / B. \quad (25)$$

The expression B is always positive. Therefore the sign of $\partial k^h / \partial \lambda$ will be the opposite of the sign of the numerator of the r.h.s of 25. The numerator is an increasing function of k^h and is positive at $k^h = 1$. If $k^d = 0$, then the numerator is non-negative for $k \geq 0$. Therefore, $k^h \geq k^d$ for any k^h , and $\partial k^h / \partial \lambda \leq 0$. If $k^d > 0$, then the numerator is non-negative for $k \geq k^d$. Therefore, if $k^h \geq k^d$ then $\partial k^h / \partial \lambda \leq 0$. If $k^h < k^d$, the numerator is negative, and $\partial k^h / \partial \lambda > 0$.

For the third statement in the proposition, note that as $\lambda \rightarrow \infty$, the solution to the first-order condition $\partial \widehat{C}_2(k) / \partial k = 0$ approaches k^d . Statements 1 and 2 of this proposition imply monotonic convergence.

Proof of Proposition 3:

From equation (19),

$$\text{Sign}\left(\frac{\partial k^h}{\partial w}\right) = \text{Sign}\left(\frac{\pi(y_e)}{y_e} \frac{\sqrt{\rho}b(1-k^h)}{2\sqrt{1-bk^h+b(k^h)^2/2}} - \frac{\pi(y_g)}{y_g} \frac{(\rho_t - \rho_d)}{2\sqrt{\rho_d + k^h(\rho_t - \rho_d)}}\right). \quad (26)$$

From Proposition 2 $k^h > k^d$ implies,

$$\alpha_e \frac{\sqrt{\rho}b(1-k^h)}{2\sqrt{1-bk^h+b(k^h)^2/2}} - \alpha_g \frac{(\rho_t - \rho_d)}{2\sqrt{\rho_d + k^h(\rho_t - \rho_d)}} > 0. \quad (27)$$

Therefore proving the following inequality completes the proof:

$$\frac{\pi(y_e)}{y_e} \frac{y_g}{\pi(y_g)} > \frac{\alpha_e}{\alpha_g}. \quad (28)$$

It can be verified that the above inequality is equivalent to proving:

$$\frac{c_g y_g^2}{\pi(y_g)} > \frac{c_e y_e^2}{\pi(y_e)}. \quad (29)$$

The first-order condition for α is,

$$c_i + \frac{w\pi'(y_i)}{y_i} - \frac{w\pi(y_i)}{y_i^2} = 0 \quad \text{for } i = g, e. \quad (30)$$

and therefore,

$$\frac{c_i y_i^2}{\pi(y_i)} = w \left(1 - \frac{y_i \pi'(y_i)}{\pi(y_i)}\right) \quad \text{for } i = g, e. \quad (31)$$

The r.h.s. of equation (31) is increasing in y_i . Further, it can be shown that $y_g > y_e$. This implies that inequalities given by 28 and 29 hold.