

The Efficiency-Quality Trade-Off of Cross-Trained Workers

Edieal J. Pinker • Robert A. Shumsky

William E. Simon Graduate School of Business Administration, University of Rochester, Rochester, New York 14627
pinker@simon.rochester.edu

Does cross-training workers allow a firm to achieve economies of scale when there is variability in the content of work, or does it create a workforce that performs many tasks with consistent mediocrity? To address this question we integrate a model of a stochastic service system with models for tenure- and experience-based service quality. When examined in isolation, the service system model confirms a well-known “rule of thumb” from the queueing literature: Flexible or cross-trained servers provide more throughput with fewer workers than specialized servers. However, in the integrated model these economies of scale are tempered by a loss in quality. Given multiple tasks, flexible workers may not gain sufficient experience to provide high-quality service to any one customer, and what is gained in efficiency is lost in quality.

Through a series of numerical experiments we find that low utilization in an all-specialist system can also reduce quality, and therefore the optimal staff mix combines flexible and specialized workers. We also investigate when the performance of the system is sensitive to the staffing configuration choice. For small systems with high learning rates, the optimal staff mix provides significant benefits over either extreme case (a completely specialized or completely flexible workforce). If the system is small and the rate of learning is slow, flexible servers are preferred. For large systems with high learning rates, the model leans toward specialized servers. In a final set of experiments, the model analyzes the design options for an actual call center.

(Queues: Approximations, Service Quality, Learning Curves, Crosstraining, Worker Turnover; Personnel)

1. Introduction

1.1. Problem and Context

A true story: A young couple is expecting their first child, and the mother-to-be begins labor. They quickly start driving to the hospital where their midwife is waiting for them. Traffic is bad, and the labor is progressing much more rapidly than their birthing class led them to expect. Along the highway they see a sign for an emergency clinic. Relieved, they pull over and run into the admitting room. “My wife’s having a

baby!” the husband says to the sole doctor present that night. “A baby? I haven’t delivered one since I was a resident, 16 years ago.” Fortunately this story ended happily for all involved.

Why was there no obstetrician at the clinic that night? Because it would be prohibitively expensive to staff such a clinic with specialists. It is a well-known fact from the study of queues that, all things being equal, staffing flexible servers is more efficient than using specialists when customers are heterogeneous in the skills they require. This is the case with medical

services, the Department of Motor Vehicles, call centers in the financial services, and computer technical support lines, among many other service facilities.

Unfortunately, all things are rarely equal. Many things can be unequal between specialists and generalists in a service setting. Either may be more expensive than the other, and the cost difference depends upon the setting. A medical specialist typically receives higher compensation than a general practitioner. On the other hand, a technical support person conversant in several operating systems may receive a higher salary than one with a narrow skill set. Specialists may be faster service providers than generalists who constantly switch between tasks of different types.

While cost and speed of service are the traditional concerns for researchers who model service systems as networks of queues, our story illustrates that the skill of a server and the quality of service delivered can also be unequal for generalists and specialists. Much of this difference in skill can be attributed to differences in experience. The problem with the emergency room doctor in the story was not that he was improperly trained, but that he almost never encountered patients in labor. The same situation may hold for a customer service representative who answers a wide range of customer calls. The representative may serve some subgroups of customers poorly because she does not accrue enough experience with their specific needs.

In this paper we study the trade-off between the cost efficiency provided by cross-trained or generalist workers and the experience-based quality provided by specialists. To accomplish this we integrate a model of a queueing system that includes multiple types of servers, a model of an individual worker's career path, and a model of experience-based learning. This integrated model provides managers and analysts with a framework for investigating how the design of a service system can affect service quality. In particular, the model links decisions about staffing policies and worker specialization with the learning curves of individual workers, system costs, and service quality.

1.2. Previous Literature

While the literature on the individual subjects of service processing systems, learning curves, and the modeling of turnover and career paths is quite large and

spans many disciplines, we believe that our work is the first to integrate the three. Some researchers have studied the control problem of how to hire, fire and promote workers to maintain appropriate staff levels when career paths are stochastic (Grinold and Stanford 1974, Gaimon and Thompson 1984). Grinold (1976), Gerchak et al. (1990), and Bordoloi et al. (1998) consider variations of this control problem when learning is present. This is, to our knowledge, the existing work that comes closest to addressing the problems we consider here. We cite the relevant work from the three areas mentioned above as we develop our model.

1.3. Paper Overview

Our goal in this paper is to develop a relatively general staffing model that captures the inherent trade-off between efficiency and quality that cross-trained or generalist workers offer. We consider a firm that serves two types of customers, *A* and *B*, each requiring a server with a distinct set of skills. The firm must decide how many employees specializing in each type of customer to hire and how many cross-trained workers capable of serving both *A* and *B* customers to hire. According to a predefined routing protocol, customers are served by either the appropriate specialist or a generalist. In our model, a customer who finds that all servers of the appropriate type are busy leaves without service.

The model defines the output of the system as the revenue derived from served customers, where revenue varies with the quality of service. The quality of service depends on the cumulative experience of the server and is independent of the length of time a customer spends in the system. The cumulative experience of the server is, in turn, determined by both the length of employment of that server and the fraction of the work experience spent developing the skills needed for this particular customer.

We define the system performance as the gross profit to the firm: The monetary value of the output less the labor and variable technology costs. The two components of this performance measure suggest a trade-off. Flexible servers should be more efficient and may reduce costs. However, by our measure flexible servers may also offer lower quality, for they may have developed a large set of skills but have mastered none.

Numerical experiments with the model confirm this insight. In addition, the model provides us with a useful tool to evaluate when one side of this trade-off dominates the other, and it enables us to determine when an optimal mix of specialists and flexible servers provides significant benefits beyond the best of the extreme solutions, a completely specialized or flexible workforce.

In the next section we formulate our model by integrating a service process model with an employee tenure model and a model of experience-based learning. Section 3 describes the approximation method used to generate the results of our numerical experiments. Section 4 describes these experiments and provides insights into the factors that determine the optimal mix of specialists and cross-trained workers. The section ends with a case study in which the model is applied to data from an actual call center. Section 5 summarizes these results and discusses possible extensions of the model.

2. Model Formulation

In this section we develop our model from three distinct components: A service process model, an employee tenure model, and an experience-based learning model of service quality. The service process model determines whether a customer is served and, if so, who provides the service. The service process model together with the tenure and learning models determine the quality of the service the customers receive and its value to the firm. In this section we describe each of the models separately and integrate them in an expression for the average quality of service experienced by a customer. This expression is incorporated into an objective function that calculates the profit for the firm.

2.1. The Service Process Model

We consider a system that serves two types of customers, A and B . There are N_A specialists who may serve only A customers, N_B specialists for B customers, and N_F flexible servers who can provide service to both types of customers. Arrivals occur according to Poisson processes with rates λ_A and λ_B . Service times are independent and exponentially distributed with rates

μ_A and μ_B for the specialists and rate μ_F for the flexible servers.¹

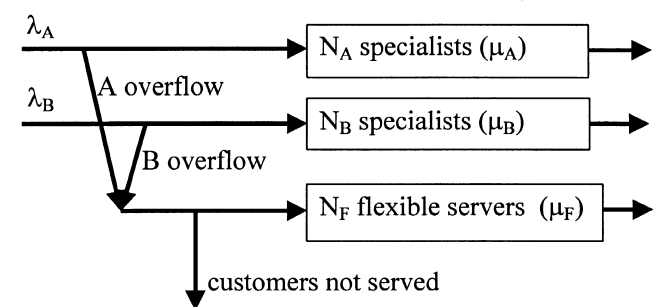
In this paper we focus on quality of service as a function of server experience, rather than the more traditional focus on waiting time or time in system. Therefore, we assume that service standards, such as the average waiting time or fraction of customers served, are determined exogenously and appear as constraints in the model. There are many service standards both in the literature and in practice, and we choose to model a loss system with the service standard defined by the fraction of customers served. We find that this structure captures the key relationships between service standards, labor costs, and server utilization while being more amenable to analysis than systems with queueing.

Figure 1 shows the routing of customers through our model. If the appropriate specialist is available, arrivals of either type immediately enter service. If all suitable specialists are busy, the customer “overflows” to a flexible server. If no flexible server is available, the customer leaves without service. Within each group of servers, arrivals are routed so that work is shared equitably among the servers.

For systems with no waiting space and static routing schemes between server groups, this customer routing scheme achieves the highest server utilization. For example, a system that sends customers to flexible servers first will have a lower throughput than our

¹The service rate for flexible servers may be an arbitrary function of μ_A and μ_B , for example a weighted average where the weights are volume driven. To make the analysis tractable we must assume that service times for flexible servers are distributed as an exponential random variable and are independent of the mix of customers in the system.

Figure 1 Routing of Customers through the Service System



specialist-first system, given identical staffing levels. While other routing schemes are possible, including dynamic ones, the one we have chosen occurs frequently in practice. In the numerical results section we discuss the impact of some alternative customer routings.

The quality model described in the next section uses two sets of statistics from this model: Throughput and utilization. Let R_{AF} represent the throughput of type A customers through flexible servers; R_{BF} , R_{AA} , and R_{BB} have similar interpretations. The total throughput $R = R_{AA} + R_{BB} + R_{AF} + R_{BF}$. Utilization is represented by ρ_{AF} , ρ_{BF} , ρ_{AA} , and ρ_{BB} , and each of these is calculated easily from the appropriate throughput via Little's Law. For example, the proportion of time flexible servers spend with A -customers is

$$\rho_{AF} = R_{AF}/(\mu_F N_F). \quad (1)$$

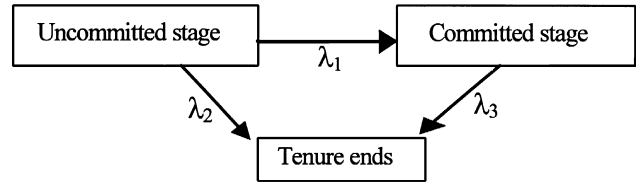
The throughput statistics of the specialists are derived easily from Erlang's loss formula (Kleinrock 1975). The behavior of the shared, flexible servers is more difficult to analyze because the arrival process is not Poisson but is instead characterized by bursts of arrivals that are sent when a group of specialists is fully occupied. While it is possible to solve for the exact throughputs of the flexible servers by numerical techniques for systems with small numbers of servers, it is impractical for large systems. In §3 we describe very accurate approximation methods to generate the throughput and server utilization statistics.

2.2. Tenure Process

The tenure of an employee is defined as the length of time from first starting a job to leaving the job for any reason, such as quitting, firing, promotion etc. We model the tenure of an individual employee as a random variable drawn from a mixed exponential probability distribution. The distribution function is derived from a career path model with a network structure that is described below.

The underlying assumption of the mixed-exponential model is that a worker's career is divided into stages (see Figure 2) that can be modeled as states of a continuous time Markov chain. The tendency to end employment (by being fired or quitting) varies from stage to stage. The time a worker stays in a stage

Figure 2 The Tenure Process



before leaving is exponentially distributed. There is a certain probability p that the worker progresses to the next stage and a probability $1 - p$ that the tenure comes to an end. We assume a two-stage model in which a worker begins in an uncommitted or probationary stage and either leaves or advances into a more committed stage in which the rate of departure is lower than the rate of departure from the first stage. In such a setting the tenure process is defined by three parameters: λ_1 , λ_2 , and λ_3 . The parameter λ_1 is the rate at which workers move from the first stage to the second stage; λ_2 is the rate at which they end their employment in the first stage; λ_3 is the rate at which workers in the second stage end their employment; and $\lambda_2 > \lambda_3$. For a worker in the uncommitted stage, $p = \lambda_1/(\lambda_1 + \lambda_2)$.

If x is the total tenure of a worker, then the probability a worker who began at time 0 remains in the system at time t is

$$\Pr\{x > t\} = G_x(t) = \frac{\lambda_2 - \lambda_3}{\lambda_1 + \lambda_2 - \lambda_3} e^{-(\lambda_1 + \lambda_2)t} + \frac{\lambda_1}{\lambda_1 + \lambda_2 - \lambda_3} e^{-\lambda_3 t}, \quad (2)$$

and the expected length of tenure is:

$$E(x) = \frac{1}{\lambda_1 + \lambda_2} \left(1 + \frac{\lambda_1}{\lambda_3} \right). \quad (3)$$

Given this model of the tenure process, we are interested in the experience of an arriving customer and therefore find the distribution of the time a randomly selected server has been on the job. An arriving customer enters into a renewal process of hired employees, which we assume is in equilibrium (see Ross 1983, Proposition 3.4.5). If, for a randomly selected server, y is the time already spent on the job, then the probability density function for y is:

$$g_y(t) = \frac{G_x(t)}{E[x]}. \quad (4)$$

Several related alternative stochastic models of tenure appear in the literature (see Bartholomew et al. 1991) such as the inverse Gaussian model, the log-normal model, and the log-logistic model. The mixed exponential model has been effective in some cases, and the network structure is intuitively appealing and relatively easy to integrate into our framework. In addition, its intuitive interpretation suggests a greater ease of application and estimation.²

Now that we can describe the server's experience, the next step is to translate that experience into quality and revenue with a learning function.

2.3. Service Quality and the Value to the Firm of Worker Experience

When considering the role of experience-based learning in staffing decisions, a for-profit firm is ultimately concerned with the actual monetary value of worker experience. In a service setting, service quality is the mechanism by which experience is translated into firm value. In this section we construct a model of the relationship between worker experience and revenue by combining a model linking experience and service quality with a model linking service quality and firm revenue.

In the literature on service management, service quality has typically been defined as a multidimensional quantity. Parasuraman et al. (1991) describe five key dimensions to service quality (tangibles, reliability, responsiveness, assurance, and empathy) that can be measured using the proposed SERVQUAL survey instrument. To our knowledge, there have not been any empirical studies of the direct relationship between the customer's perceived service quality and worker experience, i.e., a service quality learning curve. However, it is clear that on some dimensions of service quality, such as reliability and assurance, experience

often has a positive effect. Here we assume that the service quality learning curve is of a similar form to that found in production environments. In the following paragraphs we describe the function in some detail and discuss the attributes that make this a reasonable choice.

The typical learning curve found in the production literature (Yelle 1979) is a log-linear function that estimates the labor input required to produce the k th unit of a product. Naturally, it is decreasing in k . In a service setting there can be enormous variability in the requirements of customers. A worker does not gain the same experience from each customer served, and the gain in experience is often greater for more complex and time-consuming cases. If we use service time as a surrogate measure of task complexity, we can recast the cumulative experience of a worker as the cumulative time b spent performing a task. For this application, the learning curve represents the quality of service for a particular task provided by a worker after b time units of work with that task. We define experience-based service quality as

$$Q(b) = Lb^{n_l}, \quad (5)$$

where L is the quality of service provided by a worker with one time unit of experience and n_l is a parameter determining the rate at which the quality improves (both L and n_l may depend on the server and customer types). We believe that it is reasonable to assume that the function is increasing in b and that there is a diminishing marginal rate of return to experience, so that $0 < n_l < 1$.

The value to the firm of increased service quality comes from many sources. Higher quality interactions with customers prevent rework in downstream business processes and increase the likelihood of sales of products and services during those interactions. Higher service quality can also lead to increased customer loyalty, thereby increasing future sales. Certain financial service firms consider their inbound call centers to be profit, rather than cost, centers. Revenue is generated by retaining customers, selling products to existing customers, and attracting new customers by establishing a reputation for service quality (Duncombe et al. 1996). Here we assume that the monetary benefits of service quality are again positive

²Our model, as well as those described in the literature, assumes that worker replacement is instantaneous. If there are significant, random delays between a worker's departure and the arrival of a replacement, then the firm faces the additional challenge of maintaining a given level of service in the face of variable capacity. Such conditions will affect the relative benefits of flexible and specialized workers. A solution to this problem is beyond the scope of this paper.

and show diminishing marginal returns. In accordance with these assumptions and to simplify the model formulation we assume that the monetary value to the firm of providing a customer with q units of quality is:

$$V(q) = Mq^{n_2}, \quad (6)$$

where M is the value of a single unit of service quality, $0 < n_2 < 1$, and both M and n_2 may depend on the server and customer types. Combining Equations (5) and (6), we find that the revenue to the firm generated by a worker with b time units of experience in a particular task is

$$W(b) = Kb^n, \quad (7)$$

where $K = L^{n_2}M$, $n = n_1n_2$ and $0 < n < 1$. In the remainder of this paper we refer to Equation (7) as the "learning curve."

To date, there are few empirical studies in the literature that would specify the precise forms of functions that link worker experience, service quality and firm revenue, and there are scenarios in which Equation (7) may not be appropriate. When job burnout is significant, quality may have a negative slope over some values of b (Cordes and Dougherty 1993). When the service task is repetitive and customer volume is high, workers may quickly reach a peak in the learning curve. Then, some minimum utilization may be needed to maintain quality, but the length of tenure is irrelevant (in this case, a simpler form of this model would be appropriate). However, there are numerous service systems in which a server's accumulated experience has a positive impact on the service interaction. Examples include the medical staff in a hospital and call centers for medical services, hardware and software technical support, and financial services (which often require a complex mixture of customer support and sales). Function (7) has properties that are reasonable for these service systems: It is positive, increasing in experience, and offers diminishing marginal rates of return.

Given K , the parameter n may be determined by specifying a percentage increase in revenue during a proportional increase in experience b . If a factor of 10 increase in experience leads to a $\nu\%$ increase in quality-derived revenue, then

$$n = \frac{\ln(1 + \nu)}{\ln(10)}. \quad (8)$$

In the rest of the paper we refer to the parameter ν as the "learning rate." A high ν says two things about service quality: That a server learns quickly and that a server has much to learn. Higher values of ν imply greater revenue per customer after a given amount of time on the job.

Note also that $W(b)$ is unbounded from above, suggesting that in theory revenue could increase forever. When the traditional learning curve refers to task time, so that the greater the experience the faster the throughput, it is unreasonable to have an unbounded learning curve as most tasks have a minimum duration determined by physical constraints. An unbounded learning function is more appropriate when we consider improving quality through experience, because it is difficult to place an upper-bound on quality.³ We note that even if ν were 100%, the learning rate slows dramatically as time advances. For example, if we measure time in weeks and set $\nu = 100\%$, $n \approx 0.3$. Given this parameter, revenue rises by approximately 23% during the first week of experience and by approximately 3% in the tenth week.

2.4. Objective Function

We define the objective of the firm to be the maximization of expected gross profit (expected revenue less costs that are variable in the number of employees). We define the revenue of the system as the product of the number of customers served per unit time, and the monetary value of service experienced by these customers. In particular, the total quality output of the system = $\sum_i \sum_j R_{ij} E[W(b_{ij})]$, where $i = A$ or B and $j = A, B$ or F , R_{ij} is the number of type i customers served by type j servers and $W(b_{ij})$ is defined above in (7). Because we are also concerned with efficiency, we are interested in the costs of inputs. Our model contains

³While standard techniques for measuring service quality such as the SERVQUAL instrument use finite scales, these instruments do not assume that service quality is a bounded quantity. These customer survey instruments do not measure service quality per se but rather the degree to which a firm is meeting customer's expectations of what the service quality should be. Customer expectations are always changing, and in many competitive environments customers will tend to demand increasing service quality.

only costs that vary directly with labor. If N_j is the number of type j workers where $j = A, B,$ or $F,$ and c_j is the corresponding cost per unit time of such workers, which may include supporting technology and training, then the total labor cost is defined as:

$$c_A N_A + c_B N_B + c_F N_F. \quad (9)$$

The following is the expected profit of the firm:

$$Z = \sum_i \sum_j R_{ij} E[W(b_{ij})] - (c_A N_A + c_B N_B + c_F N_F). \quad (10)$$

We calculate the average quality, $E[W(b_{ij})]$, by linking the models for tenure, learning, and service. We find it convenient to condition the probability distribution of b_{ij} on y , the total time spent on the job by the server, so that the expected value is

$$E[W(b_{ij})] = K_{ij} E_y[E[b_{ij}^n | y]], \quad (11)$$

where on the right-hand side, the outer brackets define an expectation with respect to y and the inner brackets are an expectation with respect to b_{ij} given y .

To evaluate (11), we first note that because the conditional intensity of the arrival process to any group of servers is independent of the tenure state of the servers, arrivals to the process see time averages. This is the ASTA property, described in Melamed and Whitt (1990). Therefore, the distribution of y is $g_y(t)$, as defined in (4).

As for the distribution of b_{ij} given y , we show that when the expected time it takes to provide service ($1/\mu_j$) is significantly shorter than the time on the job (y), then $E[b_{ij}^n | y]$ is closely approximated by $(\rho_{ij}y)^n$ where ρ_{ij} is the long-run fraction of time that a type j worker spends on a type i job. Therefore,

$$\begin{aligned} E[W(b_{ij})] &\approx \int_0^\infty K_{ij} (\rho_{ij}t)^n g_y(t) dt \\ &= \frac{\lambda_3(\lambda_1 + \lambda_2) K_{ij} \Gamma(n+1) \rho_{ij}^n}{(\lambda_2 - \lambda_3)\lambda_3 + \lambda_1(\lambda_1 + \lambda_2)} \\ &\quad \cdot \left\{ \frac{(\lambda_2 - \lambda_3)}{(\lambda_1 + \lambda_2)^{n+1}} + \frac{\lambda_1}{\lambda_3^{n+1}} \right\}. \end{aligned} \quad (12)$$

To justify this approximation, we show that as the time on the job by a server increases, the relative approximation error vanishes. In addition, we demonstrate by

simulation that convergence occurs quickly, so that for our application approximation error is negligible.

Let Y represent the time spent on the job by a server. Here, we assume that Y is a given, constant value, and we will examine the accuracy of our approximation as Y increases. Let $\epsilon(Y)$ be the proportional deviation of the true expectation from the approximation:

$$\epsilon(Y) = \frac{E[b_{ij}^n | Y] - (\rho_{ij}Y)^n}{(\rho_{ij}Y)^n} = E\left[\left(\frac{b_{ij}}{\rho_{ij}Y}\right)^n \middle| Y\right] - 1. \quad (13)$$

The result of the following lemma ensures that this relative approximation error vanishes. For convenience, in the lemma and proof we write b_{ij} as b and ρ_{ij} as ρ , although the results apply to any type of service i and server j .

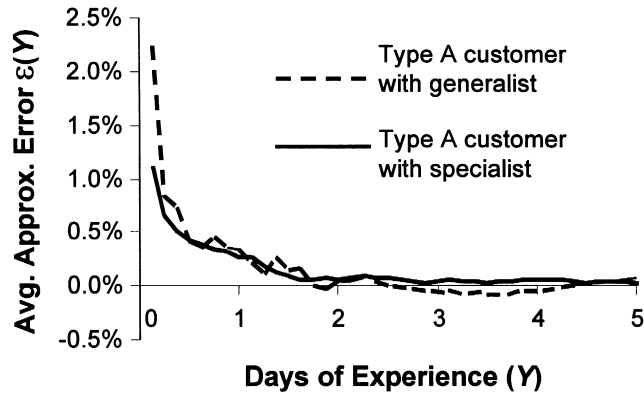
LEMMA 1. *Let Y be the total time on the job for a given server, and b represent the total time spent by the server on a particular task, and ρ be the long-run proportion of time the server spends on the particular task. Then,*

$$\lim_{Y \rightarrow \infty} E\left[\left(\frac{b}{\rho Y}\right)^n \middle| Y\right] = 1.$$

PROOF. The proof follows from an application of renewal theory. Appendix A restates the lemma and contains the details of the proof.

This lemma implies that $\lim_{Y \rightarrow \infty} \epsilon(Y) = 0$. The relative error of the approximation eventually disappears as the length of tenure grows. However, the lemma does not tell us how quickly the convergence occurs. We conducted numerous simulation experiments in which we compared the observed revenue $W(b_{ij})$ to the approximate value $K_{ij}(\rho_{ij}Y)^n$, and averaged the errors over multiple simulation runs. Figure 3 displays the results of one set of simulations with one server of each type, expected service times of approximately four minutes, and a learning rate of $\nu = 100\%$. The figure displays values of $\epsilon(Y)$ as Y grows, averaged across 500 simulation replications. We see that the relative approximation error falls off quickly. In fact, the average error for both specialist and generalist servers is less than 0.05% within two days of the beginning of each server's tenure. Additional simulations demonstrate that the convergence rate is faster for larger systems and for systems with smaller values of ν . For our application, because there is a high probability that customers observe values of the random variable y that

Figure 3 Simulation Results: Average Approximation Error as Total Time on the Job Increases



are measured in months or years, while the service times are on the order of minutes or hours, the error from the approximation (12) is negligible.

Finally, note that our definition of total quality output, $\sum_i \sum_j R_{ij} E[W(b_{ij})]$, is most appropriate for service environments in which revenue is generated directly from the quality of each service encounter, e.g., if a high-quality service experience is more likely to lead to an additional sale. There are some environments in which revenue is more closely related to the *average* quality of service experienced by customers (such would be the case if the firm wishes to attract new customers by establishing a reputation for good service). It is not difficult to alter the revenue function for such cases, and we have found that our results, as presented in §4, remain essentially unchanged under these variations.

3. Service Process Approximation Method

In this section we describe procedures to generate the service process performance statistics, R_{ij} and ρ_{ij} , needed for calculating total quality throughput and overall system profit. Because the maximum number of customers in the system is finite, it is possible to generate these quantities by simple numerical techniques. Define a three-dimensional state space in which each dimension represents the number of customers visiting each type of server (*A*-specialist, *B*-specialist, and flexible). Then describe the balance equations between states and use a numerical procedure,

such as the Gauss-Seidel procedure, to find the desired steady-state probabilities (Gross and Harris 1985). While easily implemented for small numbers of servers, this procedure is impractical when the system is large. For example, a system with 50 servers of each type requires a state space with over 10^5 elements.

Because these numerical techniques quickly become impractical, telecommunications engineers have studied these overflow systems for decades and have developed a variety of useful approximation procedures. The most common approximations are described by Wilkinson (1956), Kukzura (1972), Fredericks (1980), and Whitt (1984). Jagerman (1984) and Guerin and Lien (1990) provide summaries of the most important methods and test the accuracy of each.

The simplest, and perhaps the most well-known, approximation is Hayward's method, which estimates the blocking probability at the overflow queue with Erlang's formula after adjusting for the variability of the arrival process (Fredericks 1980). While we have tested this method, we propose an alternative. We adapt the method proposed by Kukzura (1972) in which each overflow process is approximated by an Interrupted Poisson Process (IPP). This approximation provides accurate results with a relatively small computational burden and is substantially more accurate for our application than Hayward's method.

The IPP can be described as a Poisson arrival stream which is switched between "on" and "off" periods. The stream is on for an exponential period of time with mean γ^{-1} and is off for an exponential period of time with mean ω^{-1} . The first step in the approximation procedure is to determine the values of these parameters, given the parameters of the arriving traffic and the specialized servers. Kukzura (1972) derives equations that match the parameters of the actual overflow process with the parameters of the IPP, and these equations are summarized in Jagerman (1984).

In our system, the flexible servers see two overlapping IPPs. To describe the number of customers in service with a flexible server, we define a two-dimensional, continuous-time Markov process. One dimension represents the number of customers in service, while the other dimension represents the state of the IPPs. Appendix A describes the balance equations for this Markov process. The state space is relatively

small, and steady-state probabilities for extremely large systems can be found in less than a second by numerical techniques.

To test the accuracy of this procedure, we found the system blocking probability,

$$\text{Fraction not served} = 1 - R/(\lambda_A + \lambda_B), \quad (14)$$

by using both exact numerical and approximate procedures over a wide range of system configurations and customer loads. In experiments with a system with 20 servers overall, $N_A = N_B$, $\mu_A = \mu_B$, $\lambda_A = \lambda_B$, and load factors $(2\lambda_A)/(20\mu_A)$ from 0.6 to 1.2, the difference between the exact and approximate solutions was quite small. For example, when there are 10 flexible servers and 5 specialists of each type and the load factor is 0.8, the approximation underestimates the fraction not served by less than 0.0002. In tests with systems up to 100 servers, the approximation was always within 0.0005.

The IPP approximation proved to be more accurate than the simpler Hayward's method. Hayward's method proved to be particularly inaccurate for systems with relatively few flexible servers and an overflow process characterized by long pauses and short bursts. For these systems, the error associated with the Hayward's method was 5 to 10 times larger than that of the IPP approximation.

4. Numerical Experiments

In the previous sections we developed a model for estimating the performance of a service system in terms of efficiency and quality of service. The model incorporates both worker learning and a description of the turnover process. In this section we conduct numerical experiments to derive insights into the benefits and costs related to the use of cross-trained workers in service systems. The first set of experiments (§4.1) illustrates the basic dynamics of the model by investigating the impact of staff mix on cost and quality. Using a generic system with equal numbers of *A* and *B* customers, we examine how the model parameters affect the optimal choice of staff mix, and we identify the range of parameters over which this choice has a significant impact on system performance. We also investigate

how the tenure process affects the workforce configuration.

While the first set of experiments develops general insights, the second (§4.2) demonstrates how the model would aid in the design of a particular service center. We use data collected from the call center of a consumer bank to set certain parameters, such as the system size, the ratio of *A* and *B* customers, and the length of tenure. We then examine the impact of service system design on efficiency and quality. We compare the performance of an "optimal" design (in terms of the objective function *Z*) with a variety of other designs, including the design now in use by the service center.

4.1. Insights from a Symmetric System

For the following example we assume we are modeling a call center with two types of customers, *A* and *B*, who arrive according to Poisson processes. Because all parameters are the same across the two types of customers, we call the system "symmetric." Among the following parameters, the arrival rate, service rate, tenure, and cost parameters are based on survey data that were collected under the auspices of the *Call Center Forum* of the *Wharton Financial Institutions Center* (Evensen et al. 1998). We examine a more specific system culled from the results of the survey in §4.2. Here, our baseline system has the following parameters:

- $\mu_A = \mu_B = \mu_F = 16/\text{hour}$.
- $\lambda_A = \lambda_B = 800/\text{hour}$. Therefore, the arrival load on the system is equivalent to $2 \cdot 800 / 16 = 100$ servers.
- $C_A = C_B = C_F = \$25/\text{hour}$ (for salary, benefits, training, and the costs of supporting telecommunications equipment).
- The initial rate at which workers' tenure ends (λ_2) is 0.0125 per week and the rate at which they move into the more committed stage (λ_1) is 0.05 per week as well. We assume that the rate at which workers' tenure ends from the committed stage (λ_3) is 0.006 per week. This implies that 20% of employees end tenure within 4 months, while the overall average tenure is 3 years.
- The learning curve parameters, K and ν , are the same for both types of customers and for all server types, with weeks as the time unit. We assume that $\nu = 40\%$.

To determine K , the initial revenue supplied by a

server after the first full week of work, we assume that he provides sufficient quality to break even. During this initial week the server can see 16 customers per hour and supplies each with a quality level of approximately $L = 1$. This hour of service costs the firm \$25. Therefore, we set $K = \$25/16 = \1.56 . Note that sensitivity analyses indicate that neither doubling nor halving K changes substantially the insights provided by the following experiments.

4.1.1. Impact of Staff Mix on Cost and Quality. To determine staffing requirements, we assume that the call center maintains the service standard that at least 99% of all customers must be served. There is a wide range of staff configurations that satisfy this 99% throughput constraint. The firm could staff only cross-trained workers, in effect merging the two customer streams into one. The firm could staff only specialists and manage the two customer streams independently. Staffing with any amount of cross-trained workers between 0 and the amount needed for a purely flexible work force (117 workers) is feasible. Figure 4 displays the cost of staffing the system per customer arrival, $(c_A N_A + c_B N_B + c_F N_F)/(\lambda_A + \lambda_B)$, over the entire range in the number of flexible workers. Note that the jaggedness of the plot is due to the integrality of the staff size.

It is clear from the figure that flexible workers are more efficient, in the sense that the same throughput can be achieved with fewer workers or, equivalently,

at lower labor cost. This phenomenon has been explored in the queueing literature. For example, see Whitt (1990) for a discussion of similar economies of scale in a service system. Here we see that there are diminishing benefits to increasing the flexible component of the workforce. In fact, in this example there is little benefit to having more than 34 flexible workers out of a total workforce of 118, or 29%.

In Figure 5 we plot the monetary value of the average quality experienced by an arriving customer. In the notation of Equation (10), this is $(\sum_i \sum_j R_{ij} E[W(b_{ij})]) / (\lambda_A + \lambda_B)$, or revenue divided by the arrival rate.

In general, the specialists provide higher quality service than the flexible workers. However, an entirely specialist workforce provides lower quality service than one with some flexible workers, the quality maximizing point being 7 flexible workers, or 6% of the total. This phenomenon can be explained by the fact that in each staffing arrangement the system is constrained to produce approximately the same output (i.e., at least 99% of all customers are served). When the system is staffed with only specialists, we require more workers to do the same amount of work as a system staffed with some flexible workers. This lowers server utilization, dilutes the experience gained by each worker, and reduces the quality of service delivered by the specialists. This effect is strongest when there are few flexible workers.

It is possible to increase the quality of service delivered generalists by using alternative routing schemes. For example, generalists might be segmented into

Figure 4 Labor Cost per Customer as the Number of Flexible Workers Varies

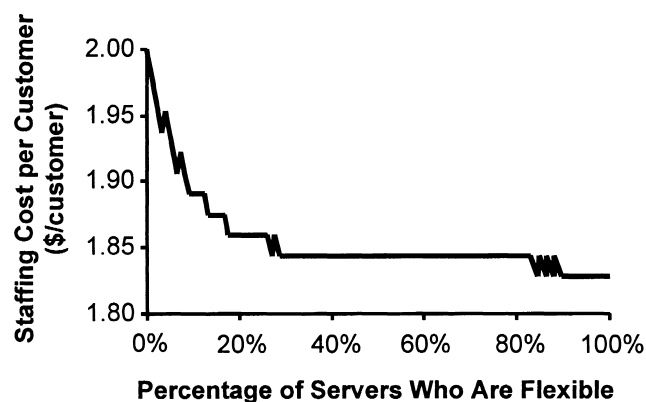
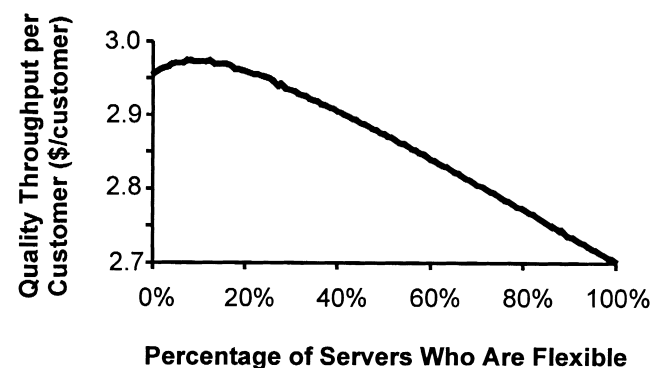


Figure 5 Average Quality Output as the Number of Flexible Workers Varies



pools. Each generalist pool serves the overflow from a single specialist queue and only sees another type of customer when there is an overflow from the other generalist pool. The routing scheme we use is simpler to manage and appears frequently in practice, but it is suboptimal in some cases. Determining optimal routings for all staffing levels is beyond the scope of this paper.

4.1.2. When is the Optimal Staff Mix Important? A manager designing a service center must determine which customer types to group together for service by cross-trained workers and which types to service independently with specialists. Once this decision has been made, the service standard determines the number of workers of each type to staff. When there are two types of customers, the choice is often whether to staff only generalists to achieve high efficiency, or only specialists to achieve high-quality service. While a manager’s intuition might suggest that an intermediate solution is best, such a choice is difficult to specify precisely, or to justify, without a model such as ours. In this section we determine when optimizing the staff mix is important by comparing the performance of optimal staffing with the two extreme strategies, 100% specialists or 100% generalists.

The results from §4.1.1 also suggest that the choice of workforce configuration involves a trade-off between the efficiency of cross-trained workers and the higher experience and quality of specialized workers. The dynamics of queueing systems show that the size of the system influences the efficiency gains created by cross-trained workers. On the other hand, the rate at which workers improve their quality of service through experience influences the impact of the staffing decisions on the quality of service experienced by the customer. Similarly the tenure process also affects the quality of service.

Here we first assume that the system under consideration is identical to the baseline system introduced at the beginning of §4.1. For this baseline system, the staffing configuration that maximizes profit, the Z-optimal system, has 49 specialized servers of each type and 21 flexible servers. The performances of this system and of the extreme configurations are summarized in Table 1.

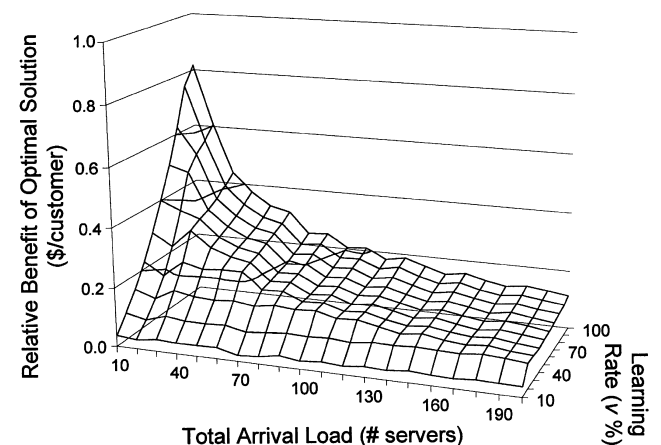
For the baseline system, the optimal configuration has a profit advantage of \$0.14 per customer over the best of the extreme solutions, while the two extreme solutions are \$0.08 apart. Of course, these results may change given different system sizes and different rates of learning, and to examine this possibility we now vary the arrival rates of both customer types and the learning rates of all employees. While the model developed in this paper also allows for differences in the service rates, learning curves, costs, and tenure processes of different classes of workers, we assume that these parameters are the same for all workers. This assumption enables us to focus attention on the questions of staff mix posed above.

Figure 6 compares the optimal staff mix with the best of the two extreme staffing levels over a wide range of ν values and arrival loads. We can see that as the size of the system increases, the importance of choosing the optimal staff mix diminishes; and as the learning rate

Table 1 Staffing Configurations and Performances for the Baseline System

Type of System	N_A	N_B	N_F	Profit (\$/Hour)	Profit (\$/Customer)	Difference
Z-Optimal	49	49	21	\$1,765	\$1.10	–
All Specialized	64	64	0	1,530	0.96	\$0.14
All Flexible	0	0	117	1,400	0.88	0.08

Figure 6 Difference in Performance (Profit per Customer) between Optimal Solution and Best Extreme Solution



increases, the advantages of the optimal system increase.

The figure suggests that for large systems it may be more important for the firm to know when to prefer one extreme over another, rather than the optimal mix of the two classes of workers. Figure 7 illustrates the benefits of making the correct choice between all-flexible and all-specialized workers for different values of ν and arrival loads. The figure is a contour chart of the performance difference between the best and worst extreme solutions. The benefits are most significant in the corners. For example, when the arrival load is low and the learning rate is just 10%–20%, an all-flexible system is superior, with a profit advantage of \$0.40–\$0.60 per customer. When the system size and learning rates are high, specialized servers gain the advantage.

Table 2 summarizes these perspectives on the importance of making the correct staffing choices.

The table provides information about how the system parameters affect the relative performance of the

Figure 7 Difference in Performance (Profit per Customer) between Best and Worst Extreme Solutions

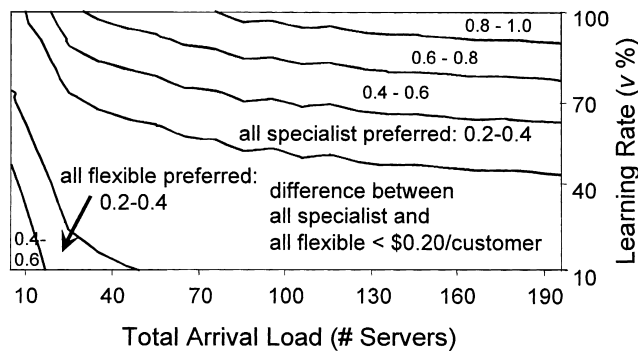


Table 2 Insights from the Numerical Experiments with a Symmetric System

		Arrival Load	
		Low (small system)	High (large system)
Learning Rate	Low	Favoring flexible important	Staff mix less important
	High	Optimal mix important	Favoring specialized important

optimal staff mix and each extreme solution. Figure 6 suggests that we distinguish between systems with arrival loads fewer or greater than 70 workers. Note, however, that if we change the cost, turnover, or performance measures, or allow the parameters to vary among the different types of workers, the determination of the precise mix of staff may become more or less critical. In the next section we examine more closely how the optimal mix varies with the tenure process.

4.1.3. Changing the Tenure Process. Recall that the parameters λ_1 , λ_2 , and λ_3 specify the distribution of tenure for each server. It stands to reason that a change in any of these parameters will change the rate of turnover in the service system and therefore will affect the quality of service. In the example introduced above, the rate of departure from the committed state was $\lambda_3 = 0.006$, and the overall average tenure was three years. A reduction in λ_3 should increase the average tenure, provide the average server with more time to “climb the learning curve,” and produce higher quality service, on average.

Numerical experiments confirm this hypothesis. An increase in average tenure generally raises system performance, although all systems show decreasing marginal returns as the average tenure grows. In addition, when the learning rate is slow (small ν), a change in the tenure process has a negligible impact on quality. When learning rates are relatively high, an increase in tenure can have a significant impact, particularly when the average tenure is already quite low.

While it is clear that the tenure process has a direct effect on the system performance, the impact on the model’s optimal staffing decision is not immediately obvious. Sensitivity analysis demonstrates that when all worker types have the same tenure parameters, changes in tenure parameters have little effect on the optimal staffing configuration. For example, the optimal configuration of the baseline case does not change as the average tenure varies from 6 months to 12 years. The reasons for the relative insensitivity of staffing to tenure follow from these three facts:

1. For the symmetric system described above, the staffing configuration that maximizes revenue (the first term in Z , Equation (10)) is independent of the values of λ_1 , λ_2 , or λ_3 . In Appendix A we restate this as a proposition and provide a proof.

2. Because staffing costs do not depend on the tenure process, the staffing configuration that minimizes staffing cost (the final terms in Z) is independent of the values of λ_1 , λ_2 , or λ_3 .

3. The form of the learning curve used in the model causes there to be diminishing marginal rates of return in revenue from increases in average tenure.

Given these three facts and the results summarized in Table 2, we see that in a small system with little learning, where the cost benefits of flexibility dominate the objective function, staffing is insensitive to the tenure process. In large systems the revenue term dominates and by fact 1, staff configuration is insensitive to the tenure process. Only in small systems with high learning rates does the tenure process have an impact on the optimal staff configuration because it affects the balance between revenue and cost in the objective function. However, even for these more sensitive systems, the diminishing marginal rate of return of lengthened tenure (fact 3) dampens the influence of the tenure parameters on the optimal staffing configuration.

4.2. A Case Study of an Asymmetric System

We now consider a call center for a consumer bank that received two types of calls: A -customers are service calls with questions about consumer bank accounts, while B -customers are customers with questions about obtaining or using a new personal computer (PC) banking service. For this call center, A -customer volume is five times as large as B -customer volume, and the learning curve for B -customers is steeper than the

A -customer learning curve, so this system is asymmetric. The following experiments demonstrate that while the small volume of B -customers may make flexible servers attractive, the efficiency gains of flexible servers can be achieved only by significantly diminishing the quality of service offered to B -customers.

We assume that the center has approximately 70 servers on duty at any one time, so that it falls between the "small" and "large" systems described in Table 2. As was true for our baseline system, the following parameters are based on survey data from Wharton's *Call Center Forum*. All parameters are identical to those in the baseline system, except for the following:

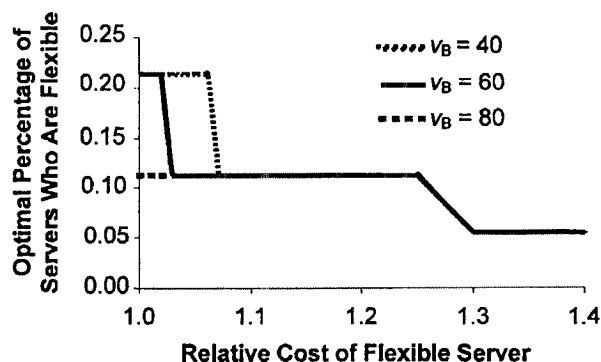
- $\lambda_A = 800/\text{hour}$, $\lambda_B = 160/\text{hour}$.
- $C_A = C_B = \$25$, but C_F , the cost of a flexible worker, is a free parameter that is adjusted below.
- Let $\nu_A = 20\%$ be the learning rate of A -customer service by specialists or generalists (this is a relatively flat learning curve—after two months, the weekly increase in quality is less than 1%).
- ν_B is a free parameter that is adjusted below, but we assume that $\nu_B \geq \nu_A$.

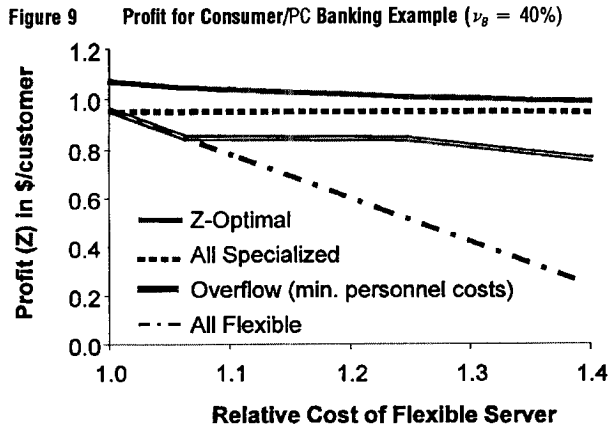
To determine staffing requirements, we assume that the call center maintains the service standard that at least 97% of all customers must be served. If queueing were allowed, a system with these arrival, service, and staffing parameters would serve approximately 90% of calls within 20 seconds. This service level exceeds the bank's target service level of 80% of calls handled within 20 seconds.

In the following experiments, we test the staffing configurations described in §4.1: all-specialists, all-flexible, and a mixture of specialist and flexible servers. We also examine the performance of a single-overflow system, which does not include any B -specialists. These systems are very common in practice, for they provide the efficiency gains of flexibility while splitting employees into two, rather than three, subgroups. In this system, all B -customers are served by a flexible server. Arriving A -customers are first routed to specialists and, if no specialist is available, the customer may be served by a flexible server.

4.2.1. Optimal Staffing Configurations. First assume that flexible servers are no more costly than specialists ($C_F = \$25/\text{hour}$) and that the learning curve for B -customers is steeper than the A -customer curve (ν_B

Figure 8 Optimal Percentage of Flexible Servers





= 40%). Then the configuration that maximizes our performance measure (Z) has 46 A -specialists, 9 B -specialists, and 15 flexible servers. Therefore, 21% of all servers are flexible. Figure 8 begins with this case on the left and shows how the optimal percentage of flexible servers declines as the cost premium for flexible servers rises. If flexible servers are 20% more expensive than specialists, just over 10% flexibility is prescribed. The percentage of flexible servers drops to 5% when the cost premium rises above 30%. Figure 8 also demonstrates how the value of flexibility declines as ν_B , the learning content of B -service, rises. The prescription that specialists should be reserved for those tasks for which substantial learning occurs is consistent with the insights derived in §4.1.

In practice, many call centers are organized as single-overflow systems with generalists and a single pool of specialists (in our case, A -specialists). Queueing models are used to find staffing levels that meet service objectives while minimizing staffing costs, $c_A N_A + c_B N_B + c_F N_F$. These models ignore the quality measures. Using a service process model for a single-overflow system, for each c_F we found the staffing configurations that minimize staffing costs. With no cost premium for flexible servers, the cost-optimal overflow system has 91% flexible servers. With a cost premium of 40%, the percentage of flexible servers declines to 30%. Note that for single-overflow systems, the percentage cannot decline to zero since in the absence of B -specialists, flexible servers are needed to serve the B -customers. In the next section we examine

Table 3 Impact of Nonoptimal Configurations on the Quality of Service for B -Customers

	$V_B = 20$	$V_B = 40$	$V_B = 60$	$V_B = 80$
All Specialized	0%	0%	-1%	-1%
Overflow (with minimum staffing costs)	-4%	-7%	-10%	-13%
All Flexible	-11%	-20%	-27%	-32%

the differences in performance among these systems.

4.2.2. System Performance. There are a wide variety of feasible staffing configurations for this simple call center: All-flexible, all-specialist, single-overflow (optimized to minimize personnel costs or optimized to maximize Z), and optimal combinations of all types of servers. At the time of the Wharton study, this particular call center had chosen an all-specialist system. Does this choice make sense? Does the configuration make a difference? In this section we examine the performance of the possible configurations by comparing each with the optimal systems described in Figure 8. First, we examine the aggregate performance measure, Z . Figure 9 displays the profit (in \$/customer), while holding $\nu_B = 40\%$. In this case, an all-specialist system performs almost as well as the optimal configuration, while all-flexible systems tend to have poor results. The performances of the single-overflow system falls between those of the all-specialist and all-flexible systems.

The poor performance of the all-flexible system in Figure 9 has one obvious cause: When flexible servers are costly, flexible system are costly as well. But that is not the whole story, for system design also has an impact on the quality of customer service. Table 3 shows the average quality of B -customer service over a range of ν_B (we have set the cost premium for a flexible server to 20%). The table shows the percentage change in average B -customer quality for three non-optimal configurations. Again, an all-flexible system is the worst performer. Recall that the volume of A -customers is five times as large as B -customer volume, so B -customers who arrive to an all-flexible system are likely to see an inexperienced server.

Because the single-overflow system tends to have a

larger number of flexible servers, *B*-customers in the single-overflow system also experience lower-quality service, on average.

In general, our results demonstrate that over a wide range of parameters, the current all-specialist design is reasonable. However, as the volume of PC banking customers (our type *B*-customer) grows, alternate configurations such as overflow systems may be more attractive. This model provides a method for evaluating these configurations for their impact on both efficiency and quality.

5. Conclusions

We have integrated a service system model with models for tenure and experience-based service quality. When examined in isolation, the service system model confirms a well-known rule of thumb from the queueing literature: Flexible or cross-trained servers provide more throughput while using fewer workers. However, in the integrated model these economies of scale are tempered by a loss in quality. Given multiple tasks, flexible workers may not gain sufficient experience to provide high-quality service to any one customer, and what is gained in efficiency is lost in quality.

This result may seem to suggest that an all-specialist workforce would guarantee the highest quality service. However, the model also demonstrates that low utilization in an all-specialist system can also reduce quality; employees may leave before they have accumulated a significant amount of experience. According to this model, the optimal staff mix is therefore a mixture of flexible and specialized workers.

The model also highlights when the performance of the system is sensitive to the staffing configuration choice. For small systems with high learning rates, the optimal staff mix is a significant improvement over either extreme case (a completely specialized or completely flexible workforce). If the rate of learning is slow and the system is small, flexible servers are preferred. For large systems with high learning rates, the model leans towards specialized servers. A uniform change in the tenure process (or turnover rate) among all employees has little effect on the optimal staffing decision. However, an increase in the average tenure

significantly improves overall quality, particularly if the learning rate is high and the original average tenure is low.

In practice, service systems are substantially more complex than the system described here. For example, we have assumed that customers who find all appropriate servers busy will leave the system. Because many service systems allow customers to queue for service, we may wish to incorporate queueing into the underlying service model. This extension would add to the complexity of the process model (see Green 1985 for one example of such a process), and the performance measure *Z* must be altered to reflect the impact of waiting time on perceived service quality. However, because our model of experience-based quality is driven exclusively by server utilization and employee turnover, the fundamental insights about the relationships between system size, learning rates, and turnover are not likely to change if queueing were incorporated into the service process model.

In some settings, forgetting is an important phenomenon as well as learning (Bailey 1989). The evidence suggests that this is a concern when lengthy work interruptions take place after intense periods of learning. One possible extension to the model is to add a "forgetting" component to the learning curve so that when utilization drops below some threshold, quality would decline. This would tend to accentuate the quality differences between servers who are heavily utilized for a particular task and those who see the task only rarely.

In general, the existing model can be readily transformed into a more elaborate representation of a service facility. A design issue currently under investigation is the choice of alternative customer routing schemes. Extensions that are easily made include three or more customer and service classes and heterogeneous service rates and tenure processes among server classes. By adjusting learning rates and the costs of the servers, the model can be used to assess the benefits of information technology such as expert systems and databases that transfer knowledge from specialists to flexible servers. The model can also serve as a platform for examining training programs and personnel assignment decisions. For example, many firms rotate employees through a variety of functions to prevent

burnout at any one job. While this has the potential to increase overall tenure, such a rotation system may reduce the average experience level at each job position. The tenure and learning processes under a rotation system would be more complex than the simpler model proposed in this paper. However, the model proposed here is sufficient to demonstrate the fundamental structure of interactions between learning, turnover, efficiency and quality in a service center.⁴

Appendix A. Lemma and Proposition

LEMMA. Let Y be the total time on the job for a given server, b represent the total time spent by the server on a particular task, and ρ be the long-run proportion of time the server spends on the particular task. Then,

$$\lim_{Y \rightarrow \infty} E \left[\left(\frac{b}{\rho Y} \right)^n \middle| Y \right] = 1.$$

PROOF. Consider the points in time that the system clears (all servers idle). These points are event times of a renewal process because the time periods between these events constitute a sequence of non-negative independent random variables with a common distribution function. Renewal theory tells us that with probability 1, $b/Y \rightarrow \rho$ as $Y \rightarrow \infty$ (Ross 1983).

It will be convenient to restate this in measure-theoretic terms. Let $b(\omega, Y)$ denote the random variable defined on (Ω, \mathcal{F}, P) , where $\omega \in \Omega$ is an elementary outcome and Y is the parameter described above. Then,

$$\frac{b(\omega, Y)}{Y} \rightarrow \rho \quad \text{as } Y \rightarrow \infty \text{ for all } \omega, \omega \notin A,$$

where event A has measure zero. Because $0 < \rho < 1$ and $0 < n < 1$, ρ^n is a continuous function of ρ . Therefore,

$$\left(\frac{b(\omega, Y)}{Y} \right)^n \rightarrow \rho^n \quad \text{as } Y \rightarrow \infty \text{ for all } \omega, \omega \notin A.$$

Because $0 \leq (b(\omega, Y)/Y)^n \leq 1$, the function $(b(\omega, Y)/Y)^n$ is integrable for all ω and Y . By the Lebesgue Dominated Convergence Theorem (Rao 1973, p. 136),

$$\lim_{Y \rightarrow \infty} E \left[\left(\frac{b}{Y} \right)^n \middle| Y \right] = \lim_{Y \rightarrow \infty} \int_{\Omega} \left(\frac{b(\omega, Y)}{Y} \right)^n dP = \rho^n.$$

Dividing both sides by ρ^n gives us the desired result.

PROPOSITION (THE RELATIONSHIP BETWEEN TURNOVER AND REVENUE). If (i) all worker types A, B or F have identical tenure processes

⁴The authors thank Gregory Dobson, Marshall Freimer, the senior editor, associate editor, and two anonymous referees for valuable comments. We also thank Frances Frei and Patrick Harker for giving us access to data from the Call Center Forum of the Wharton Financial Institutions Center.

determined by the parameter vector $\bar{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$, (ii) all worker types have identical learning curves determined by K and n , and (iii) $\lambda_2 > \lambda_3$ or $(\lambda_3 - \lambda_2)\lambda_3^{n+1} > \lambda_1(\lambda_1 + \lambda_2)^{n+1}$, then the staffing configuration that maximizes revenue, $N' = (N'_A, N'_B, N'_F)$, is independent of the value of $\bar{\lambda}$.

PROOF. By combining Equations (10) and (12) we find the following expression for total revenue:

$$\begin{aligned} & \sum_i \sum_j R_{ij} E[W(b_{ij})] \\ & \approx \frac{\lambda_3(\lambda_1 + \lambda_2)K\Gamma(n+1)}{(\lambda_2 - \lambda_3)\lambda_3 + \lambda_1(\lambda_1 + \lambda_2)} \left\{ \frac{(\lambda_2 - \lambda_3)}{(\lambda_1 + \lambda_2)^{n+1}} + \frac{\lambda_1}{\lambda_3^{n+1}} \right\} \\ & \cdot \{R_{AA}(\rho_{AA})^n + R_{AF}(\rho_{AF})^n + R_{BB}(\rho_{BB})^n + R_{BF}(\rho_{BF})^n\} \\ & = A(\lambda, K, n) \{R_{AA}(\rho_{AA})^n + R_{AF}(\rho_{AF})^n \\ & + R_{BB}(\rho_{BB})^n + R_{BF}(\rho_{BF})^n\}. \end{aligned} \quad (A.1)$$

The first term in the product, $A(\bar{\lambda}, K, n)$, depends on the tenure and learning curve parameters but does not vary with the staffing configuration. Condition (iii), above, ensures that this term is always positive. Therefore, if the configuration N' maximizes the second term of the expression, then it must maximize total revenue for any fixed $\bar{\lambda}$, K , and n . By definition, the condition $\lambda_2 > \lambda_3$ is satisfied for our application (see §2).

Appendix B. The Flexible Server Approximation

The flexible servers see one stream of customers that overflows from the A -specialists and another stream that overflows from the B -specialists. The N_i specialists ($i = A, B$) see customers who arrive according to Poisson processes with rates λ_i , and the specialists serve customers at rates μ_i . The overflow processes are not Poisson, but each can be approximated by an interrupted Poisson process (IPP). Given λ_i , N_i , and μ_i , Kuczura's formulas calculate the following parameters (Kuczura 1972 and Jagerman 1984):

γ_i^{-1} = average length of an "on" period in the overflow process (for $i = A, B$),

ω_i^{-1} = average length of an "off" period in the overflow process,

λ_{oi} = rate of arrivals during an "on" period.

We assume that the lengths of both "on" and "off" periods are distributed as exponential random variables, and arrivals to the flexible servers follow a Poisson process with rate λ_{oi} during an "on" period. Note that these arrival rates are represented by λ_{oA} and λ_{oB} rather than by λ_A and λ_B to distinguish them from the original arrival rates.

Represent the state of the flexible servers with a pair of numbers, (m, n) , where

m = number of customers in service with a flexible server ($0 \leq m \leq N_F$),

n = an index state signifying which overflow processes are "on" ($0 \leq n \leq 3$).

In particular,

$$n = \begin{cases} 0 & \text{when neither overflow process is on,} \\ 1 & \text{when the } A\text{-overflow process is on,} \\ 2 & \text{when the } B\text{-overflow process is on,} \\ 3 & \text{when both overflow processes are on.} \end{cases}$$

In the following balance equations,

$$\delta(m) = \begin{cases} 0 & \text{if } m = 0, \\ 1 & \text{otherwise.} \end{cases}$$

For $0 \leq m \leq N_F$,

$$\begin{aligned} [\delta(m)m\mu + \omega_A + \omega_B]p(m, 0) \\ = \gamma_A p(m, 1) + \gamma_B p(m, 2) \\ + \delta(m - N_F)(m + 1)\mu p(m + 1, 0). \end{aligned} \quad (B.1)$$

$$\begin{aligned} [\delta(m)m\mu + \omega_B + \delta(m - N_F)\lambda_{oA} + \gamma_A]p(m, 1) \\ = \omega_A p(m, 0) + \delta(m)\lambda_{oA} p(m - 1, 1) + \gamma_B p(m, 3) \\ + \delta(m - N_F)(m + 1)\mu p(m + 1, 1). \end{aligned} \quad (B.2)$$

$$\begin{aligned} [\delta(m)m\mu + \omega_A + \delta(m - N_F)\lambda_{oB} + \gamma_B]p(m, 2) \\ = \omega_B p(m, 0) + \delta(m)\lambda_{oB} p(m - 1, 2) + \gamma_A p(m, 3) \\ + \delta(m - N_F)(m + 1)\mu p(m + 1, 2). \end{aligned} \quad (B.3)$$

$$\begin{aligned} [\delta(m)m\mu + \delta(m - N_F)(\lambda_{oA} + \lambda_{oB}) + \gamma_A + \gamma_B]p(m, 3) \\ = \omega_B p(m, 1) + \omega_A p(m, 2) \\ + \delta(m)(\lambda_{oA} + \lambda_{oB}) p(m - 1, 3) \\ + \delta(m - N_F)(m + 1)\mu p(m + 1, 3). \end{aligned} \quad (B.4)$$

With $4(N_F + 1)$ states, solving for the steady-state probabilities takes little time. When $N_F = 250$, the Gauss-Seidel procedure, implemented in the C programming language and running on a Pentium 120 processor, takes less than one second to converge.

Given steady-state probabilities $p(m, n)$, the throughput of A -customers through the flexible servers is:

$$R_{AF} = \lambda_{oA} \sum_{i=0}^{N_F-1} (p(i, 1) + p(i, 3)), \quad (B.5)$$

and the throughput of B -customers is:

$$R_{BF} = \lambda_{oB} \sum_{i=0}^{N_F-1} (p(i, 2) + p(i, 3)). \quad (B.6)$$

References

Bailey, C. 1989. Forgetting and the learning curve: A laboratory study. *Management Sci.* **35**(3) 340–352.

Bartholomew, D. J., A. F. Forbes, S. I. McClean. 1991. *Statistical Techniques for Manpower Planning*, 2nd ed. John Wiley and Sons, New York.

Bordoloi, S. K., W. W. Cooper, H. Matsuo, H. 1998. Human resource planning in knowledge-intensive assembly operations. Unpublished Mimeo. Management Department, Graduate School of Business, The University of Texas at Austin, Austin, TX.

Cordes, C. L., T. W. Dougherty. 1993. A review and an integration of research on job burnout. *Acad. Management Rev.* **18**(4) 621–656.

Duncombe, V., G. Griffiths, D. LaBrie, et al. 1996. Moving from basic call centers to world class call centers. *Bank Marketing* **28**(2) 19–24.

Fredericks, A. A. 1980. Congestion in blocking systems—a simple approximation technique. *Bell System Tech. J.* **59**(6) 805–827.

Evensen, A., P. T. Harker, F. X. Frei. 1998. Effective call center management: Evidence from financial services. Working Paper 98–25, The Wharton Financial Institutions Center, University of Pennsylvania, Philadelphia, PA.

Gaimon, C., G. Thompson. 1984. A distributed parameter cohort personnel planning model that uses cross-sectional data. *Management Sci.* **30**(6) 750–764.

Gerchak, Y., P. Mahmut, S. Sengupta. 1990. On manpower planning in the presence of learning. *Engrg Costs and Production Econom.* **20** 295–303.

Green, L. 1985. A queueing system with general-use and limited-use servers. *Oper. Res.* **33**(1) 168–182.

Grinold, R. C. 1976. Manpower planning with uncertain requirements. *Oper. Res.* **24** 387–399.

———, R. E. Stanford. 1974. Optimal control of a graded manpower system. *Management Sci.* **20** 1201–1216.

Gross, D., C. M. Harris. 1985. *Fundamentals of Queueing Theory*, 2nd ed. John Wiley and Sons, New York.

Guerin, R., L. Y.-C. Lien. 1990. Overflow analysis for finite waiting room systems. *IEEE Trans. Comm.* **38**(9) 1569–1577.

Jagerman, D. L. 1984. Methods in traffic calculations. *AT&T Bell Laboratories Tech.* **63**(7) 1283–1310.

Kleinrock. 1975. *Queueing Systems*, Vol. 1. John Wiley and Sons, New York.

Kukzura, A. 1972. The interrupted Poisson process as an overflow process. *Bell System Tech. J.* **52**(3) 437–448.

Melamed, B., W. Whitt. 1990. On arrivals that see time averages. *Oper. Res.* **38**(1) 156–172.

Parasuraman, A., L. Berry, V. Zeithaml. 1991. Understanding, measuring, and improving service quality. Findings from a multi-phase research program. *Service Quality, Multidisciplinary and Multinational Perspectives*. S. W. Brown, B. Edvardsson, E. Gustmesson, and B. Gustavsson, eds., Lexington Books, Lexington, MA.

Rao, C. R. 1973. *Linear Statistical Inference and Its Applications*. John Wiley and Sons, New York.

Ross, S. M. 1983. *Stochastic Processes*. John Wiley and Sons, New York.

Whitt, W. 1984. Heavy-traffic approximations for service systems with blocking. *AT&T Bell Laboratories Tech. J.* **63**(5) 689–708.

———. 1990. Understanding the efficiency of multi-server service systems. *Management Sci.* **38**(5) 708–723.

Wilkinson, R. I. 1956. Theories for toll traffic engineering in the U.S.A. *Bell System Technical J.* **35**(2) 421–518.

Yelle, L. 1979. The learning curve: Historical review and comprehensive study. *Decision Sci.* **10** 302–328.

The consulting Senior Editor for this manuscript was Lawrence M. Wein. This manuscript was received on July 24, 1998, and was with the authors 3 months for 2 revisions. The average review cycle time was 71.7 days.