Knowledge Management in Call Centers: The Impact of Routing on the Distribution of Expertise

Geoffrey Ryder • Christoph Heitz† • Kevin Ross

Baskin School of Engineering, University of California–Santa Cruz
1156 High Street, Santa Cruz, CA 95064, USA

† *Institute for Data Analysis and Process Design,*
Zurich University of Applied Sciences
CH-8401 Winterthur

gryder@soe.ucsc.edu • christoph.heitz@zhwin.ch • kross@soe.ucsc.edu

When customer inquiries arrive at a call center, management's routing rules bear the responsibility for assigning service agents to handle them. These rules must balance the need to minimize customer waiting time with the desire to match customers with the most skilled agents. A myopic rule for skill-based routing would consider an agent's skill level to be fixed, as determined by formal training or certification. However, the routing rule itself has an impact on skill levels: through on-the-job learning, the development of the agents' expertise depends on the calls they take.

In this paper we develop quantitative tools that allow us to reason about the process of acquiring on-the-job expertise by call center agents. First, we show how to link routing decisions to expertise level outcomes. We then define utility functions that depend on that expertise, and show how polices of evenly shared routing and extreme specialized routing affect those utilities. We use these utilities as metrics in an analysis of how to optimize expertise development while meeting service level targets. Finally, we introduce a measure of forced task sharing, and describe how it shapes the distribution of skills among the workforce of agents. Empirical data that describe the distribution of customer arrivals among task types guides our analysis.

Key words: call center, routing, expertise, learning, forgetting, turnover, nonlinear optimization, knowledge management, workforce planning

1. Introduction

1.1 Expertise in a Call Center Setting

A major influence on a customer's satisfaction at a call center is the knowledge level of the agent who takes their call. Knowledge management, in particular maintaining or increasing the cumulative knowledge of the agents, is therefore a key issue for ensuring service quality. This is especially

true when the call center operates within dynamic markets, and agents are required to keep pace with changes.

For the operational management, on the other hand, the knowledge of the employees is usually treated as *exogenous* to the service delivery process. Knowledge is considered to be a given and fixed resource, and is treated as such for routing and call assignments. This makes sense when all training happens *off-line*, but does not account for the case where knowledge and expertise are actually gained *on-the-job* through the service process itself. If we assume that learning-on-the-job takes place, then the operational rules have an impact on knowledge, and knowledge is therefore an *endogenous* rather than an exogenous variable. In particular, routing policies determine which agents work on which jobs, and thus may have a major impact on the learning of the agents and their expertise level attained. In our paper, we study how routing might influence the knowledge, how changing knowledge levels will affect customer experience, and how knowledge management and routing can be treated together.

We model the expertise of a service agent with simple dynamic equations, reflecting the essential features of gaining expertise through experience (learning) and lowering expertise through absence (forgetting). We find that, in the long run, the expertise level of an agent increases as the arrival rate to this agent increases. That is, a busy agent will maintain a higher level of expertise and therefore give the customers better average service.

As yet a third factor driving the results, frequent turnover will reduce the average expertise level within the firm. In practice turnover may indicate the process of agents quitting and new agents being hired; or, it may indicate changes in task content due to changes in demand patterns, which may make prior experience irrelevant. We introduce a form of discounting due to turnover to our model as well.

In a multi-agent environment, the arrival rates to each agent are influenced by the routing rules employed at the call center. Different routing rules may lead to different distributions of expertise, and therefore to different customer quality experiences. We describe how to scale up our analysis of an agent's on-the-job learning to systems with multiple agents and task types, and to design routing rules that meet management's goals for agent expertise in this setting.

1.2 Service Quality, Learning, and Turnover in Call Centers

A key paper that helps establish this area of research is by Pinker and Shumsky (2000). They analyze a Markov chain system model of learning and turnover that includes two types of specialist workers, and a set of cross-trained or flexible workers that can perform either task. Each workers

service quality improves with tenure, though specialists always provide the highest service quality. The optimal staffing arrangement, giving both high agent utilization and high average expertise, turns out to include a mix of specialists and flexible agents. They also specify how the staffing solution changed along the dimensions of arrival load and expertise development rate (or learning rate). High learning rates favor more specialists in large systems, and a precise, optimal mix of flexible and specialized agents in small systems. By contrast, low learning rates favor a staff mix with more flexible agents—the content of the work is simple enough that it can be mastered quickly, so agents can take on more tasks. The authors recommend the use of forgetting models in future work, and we aim to build on their results in this paper by including forgetting and learning together.

Gans and Zhou (2002) model learning and turnover effects using a Markov decision process, and demonstrate that the optimal hiring policy for each state of a firms agent roster is a "hire-up-to policy similar to the "order-up-to policies from the supply chain management literature. Whitt (2006) explores ways to characterize the employee retention distribution for call center agent populations. Improved retention results in a higher average expertise of agents in the center, because expertise improves with tenure. Among other findings, decreasing distributions such as the negative exponential are noted to be reasonable first approximations to real retention distributions. This is because in general employees are most likely to leave within a short time of starting, and the probability of leaving then decreases as tenure grows.

We share a key assumption with these three previous papers: we assume service quality, denoted here by variable X, improves with tenure; or more precisely with cumulative production. But we do not specify precisely what improved quality means to the customer. This allows us to reason about the relationship of routing rules and expertise in a general way that can be adapted to fit specific cases. In an application of our results to call center agent data, X may stand for the handle time of a call, which should decrease with expertise; or to the first call resolution rate (FCR), which should increase with experience. See Vericourt and Zhou (2005) for a discussion of the FCR metric. If SERVQUAL-type survey data is available, X may represent a measure of customer satisfaction (Parasuraman et al., 1988). Froehle (2006) conducts a statistical analysis of such data, and finds that agent preparedness, subject matter knowledge, and thoroughness are most important to customers' perceptions of service—three qualities that can be expected to increase with experience. Froehle also describes the alternatives modern agents have for communicating with customers, such as email and instant messaging. The customer call center is now rightly termed a *customer contact center* as well. We will use the terms interchangeably.

Our work has been guided by results in the literature from several areas of service operations research. For more on learning and turnover in call centers, see Bordoloi (2004), Gans and Zhou (2003), and Zohar et al. (2002). For an in-depth background on call center planning and operations, see Aksin et al. (2007), Brown et al. (2002), Cleveland and Mayben (2000), Gans et al. (2003), Hasija et al. (2005), Iravani et al. (2007), and Koole (1997). Due to the inherent complexity of call center operational models, high quality simulations are becoming important (Avramidis and L'Ecuyer, 2005).

For related results on learning and forgetting at work, Shafer et al. (2001) present a detailed study of empirical learning and forgetting data in an industrial application with worker service times roughly equivalent to call handle times in a call center. Badiru (1992) presents a survey of applied learning models, and Nembhard and Osothsilp (2001) do the same for forgetting models. Sikström and Jaber (2002) explore new ways of measuring the impact of production breaks on productivity. Sayin and Karabati (2007) develop a detailed optimization model for solving a rostering problem with learning and forgetting effects in a corporate setting involving several departments. A similar problem is explored by Eitzen et al. (2004), who note that forgetting effects require that worker skill levels be maintained through repetition in work assignments.

We do not discuss the details behind learning and forgetting effects here, but there is a body of work from the behavioral sciences that supports our operational models. See especially Globerson and Levin (1987), Howick and Eden (2007), and Schilling et al. (2003). Behavioral scientists see new opportunities opening up now for joint work with those in the operations field, in order to apply the growing catalog of behavioral results to service operations (Bodreau et al. 2003).

There is a growing set of work describing call center outsourcing contracts, and the competitive milieu faced by call center operators—see for example Aksin et al. (2008), Hasija et al. (2008), Ren and Zhou (2008), Shumsky and Pinker (2003). The ultimate goal of this research is to provide new avenues for productivity growth in call center operations, so that savvy operators may drive more profitable, or lower cost, service contracts, such as described in Reis (1991). Although in practice measuring and acting on learning curves requires care, productivity gains have contributed to business success in a variety of settings (Ghemawat 1985).

2. The Dynamics of Agent Expertise

2.1 Finding the Steady-State Value of Expertise

Consider the evolution of expertise in an agent answering calls to a call center. Let the expertise $X(t)$ of the agent at time t be on a scale $0 \leq X(t) \leq 1$, where $X(t) = 0$ indicates a novice, and $X(t) = 1$ corresponds to an expert. Define the average time between completed jobs to be τ_a, including the receiving and processing of a job, followed by some time until the next job arrives. The arrival *rate* of customers to the system is $\lambda = 1/\tau_a$. We assume that the agent learns while processing the job (on-the-job), thus increasing its expertise level X(t), and forgets while not processing, leading to a decrease of X(t).

In our learning model, the agent's expertise by processing one job increases on the average through

$$X(t) \mapsto X(t) + \alpha(1 - X(t))$$

where α is a learning parameter. That is, the experience gain is proportional to $(1 - X(t))$, and so becomes geometrically smaller as $X(t)$ approaches expert status. In the absence of forgetting, an agent will move from novice to roughly half of her maximum possible level by completing $1/\alpha$ jobs.

Skills need to be maintained through reinforcement; in the absence of work to occupy an agent, forgetting ultimately reduces the expertise of the agent to zero (novice level). We assume that forgetting occurs at a continuous rate β, so that for a period of length τ_a, the expertise is discounted by $e^{-\beta\tau_a}$. Taking learning events and continuous forgetting together, we get

$$X(t + \tau_a) = (X(t) + \alpha(1 - X(t))e^{-\beta\tau_a} \tag{1}$$

Learning is designed to be a geometrically decreasing concave function of time, and the forgetting exponential function is convex in time for positive τ_a, which holds for the cases we consider. Given these simple dynamics, asymptotic behavior of $X(t)$ will tend toward the fixed point X_∞ of this equation, with $0 \leq X(t) \leq 1$. The smaller τ_a (i.e. the more jobs per time unit the agent is handling), the higher the asymptotic expertise level X_∞, and vice versa.

$$X_\infty = \frac{\alpha}{e^{\beta\tau_a} + \alpha - 1} \tag{2}$$

The final detail we will include in the expertise model is a limit on forgetting. If we assume that forgetting *only occurs when the agent is idle*, we may modify the definition of τ_a in Equation (2)

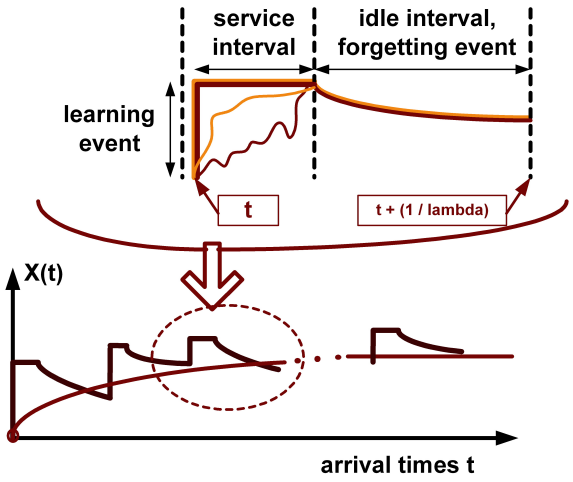

Figure 1: An illustration of how on-the-job experience is developed.

to be the expected time τ_I that the agent is not actively helping customers.

$$\tau_I = \left(\frac{1}{\lambda} - \frac{1}{\mu}\right), \text{ where } \frac{1}{\lambda} \geq \frac{1}{\mu} \tag{3}$$

To keep the sign of the exponent in the forgetting term negative, τ_I must be positive. If the agent's utilization is greater than 100%, or $\lambda > \mu$, then let τ_I be zero, and let the forgetting function be $e^0 = 1$.

Then the steady-state value of expertise becomes:

$$X_\infty = \frac{\alpha}{e^{\beta\tau_I} + \alpha - 1}. \tag{4}$$

From prior studies of learning and forgetting rates, and our own observations of call center data, we expect a reasonable range of interest for parameter α to be between 0.05 and 1e-4; and we take $\beta \leq \alpha$. See Appendix 6.1 for specifications under which this expertise function is concave.

2.2 Mapping Arrival Rates to Expertise Levels

Figure 1 depicts the evolution of these expertise equations over time: on-the-job experience grows through serving a sequence of customers. Expertise is increasing in an agent's relevant *cumulative production.* (Badiru 1991, Shafer et al. 2001). As the top diagram shows, a true accounting of the knowledge-building process would involve a highly nonlinear, complex function. We simplify this function in our stylized model by awarding an expertise increase at the time when a customer arrives, where the size of the boost depends on the proximity of current expertise $X(t)$ to asymptotic expertise X_∞.

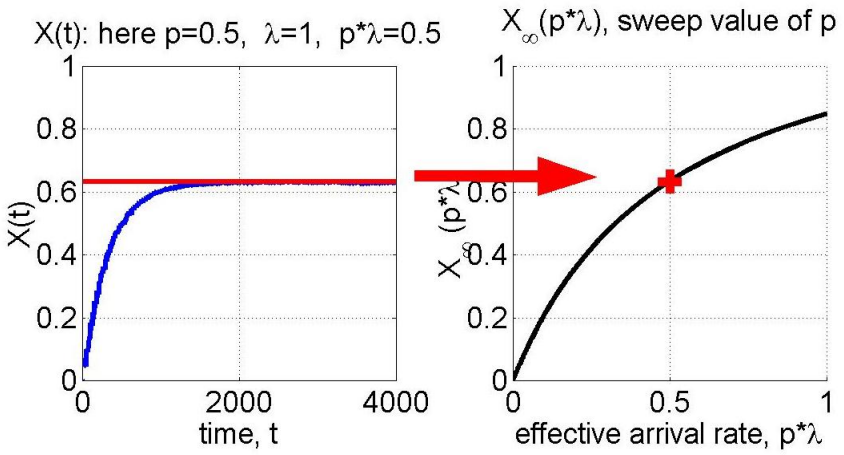

Figure 2: The asymptotic value of expertise. The key advantage of our formulation: expertise becomes a continuous function of the routing proportion p. Here $\alpha = 2e - 3$, $\beta = 8e - 4$, $\mu = 10$, $\lambda = 1$, and $p = 0.5$.

The increase is held constant during the service time; when service is done and the customer leaves, that increase is allowed to gradually diminish according to our negative exponential forgetting function. This is one type of function among several that have been used to fit the measured effects of forgetting (Nembhard and Osothsilp 2001); we apply it here for its simplicity and tractability.

The bottom of Figure 1 depicts the trend of expertise over time. Decisions by individual customers about when to call are unpredictable—customer arrivals appear random to the agent. A longer interarrival time will incur more forgetting, a short one less, but over many customer visits the variations average out and $X(t)$ settles to X_∞ of Equation (4).

Figure 2 illustrates a key feature of our expertise formulation: *we have established a link between the arrival rate of jobs to an agent and her expertise level.* Management can take advantage of this linkage to design routing rules that optimize the distribution of expertise among the workforce. The succeeding sections will discuss criteria for optimizing this distribution.

Note that an analysis of system capacity, the variable lambda (λ) traditionally stands for an arrival rate. Here we let it be the rate coming into the entire system, which may employ many agents. Then the rate to any single agent will be a fraction of λ. Let this fraction of λ arriving at a single agent be denoted by p, with $0 \leq p \leq 1$. Then we may modify the idle time from Equation (3) to be:

$$\tau_I = \left(\frac{1}{p\lambda} - \frac{1}{\mu} \right) \tag{5}$$

Management controls the proportion p for each agent through routing rules, hence controlling

the intensity of on-the-job learning experiences, and ultimately the agent's expertise level. We make this linkage explicit by writing the asymptotic expertise level as $X_\infty(p\lambda)$. Where the value of λ is understood to be a certain value, we can just write $X_\infty(p)$. Note that although $X(t)$ was a function of discrete interarrival times t, $X_\infty(p\lambda)$ may be a continuous function of a continuous real variable p, and is thus convenient for analysis.

In Figure 2, the left side shows $X(t)$ increasing as customers are served, until the asymptotic value $X_\infty(p\lambda)$ is reached—the horizontal line at about 0.63. This asymptotic limit maps to a single point on the plot at the right. The right side shows the value of $X_\infty(p\lambda)$ as the routing proportion p to this agent is swept from zero to one.

2.3 The Asymptotic Expertise Level Discounted for Turnover

Now we have an expression for the asymptotic expertise level of an agent that depended on routing rule proportion p, customer arrival rate λ, service rate μ, learning parameter α, and forgetting parameter β. However, in a service organization such as a call center, new agents join and veteran agents leave on a regular basis. Furthermore, product or policy changes that affect customers may cause the content of their inquiries to change, rendering an agent's current expertise obsolete. We refer to these interruptions in expertise development as *turnover events*.

In this section, the model records occasional turnover events as sudden losses of all expertise accumulated since the agent first started. For simplicity, we adopt the convention that the workforce remains at a constant size, and workers are either rehired or retrained (Pinker and Shumsky 2000, Gans and Zhou 2002, Whitt 2006).

Figure 3 shows two plots of expertise over time, with turnover events occurring at random intervals. The top figure uses an extreme routing rule, where one agent gets all the work; the routing proportions are $p_1 = 1, p_2 = 0$. The bottom figure uses an even routing rule, $p_1 = 0.5, p_2 = 0.5$. (Training and retraining periods are omitted.) It is evident that turnover events may prevent agents from reaching their limiting expertise values—in this example the average system expertise over time is closer to the midpoint of the experience curves, as indicated by the horizontal lines.

We can discount the asymptotic value of expertise in the presence of turnover events to reflect this effect, as follows. Let the variable n count each job completed; let $\theta(p)^- = e^{-\beta(1/(p\lambda)-1/\mu)}$; and let $\theta(p)^+ = e^{+\beta(1/(p\lambda)-1/\mu)} = e^{+\beta\tau_{idle}}$, where $\tau_{idle} = (1/(p\lambda) - 1/\mu)$.Then we can rewrite the previous asymptotic value from Equation (4) as

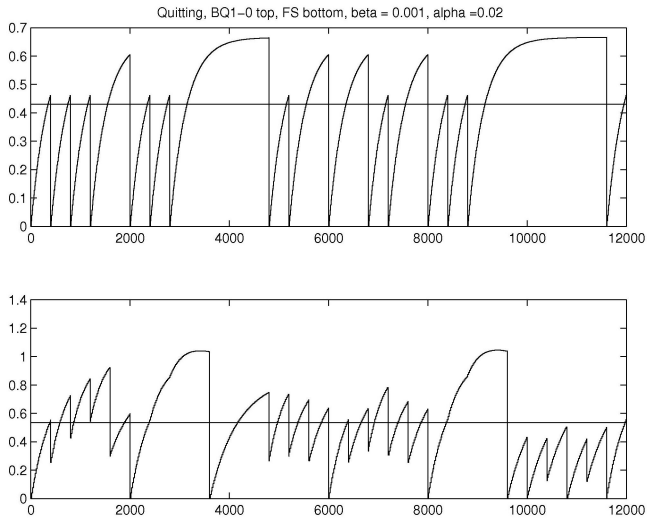

Figure 3: Time series plots of expertise versus time. Top: a policy of extreme routing. Bottom: a policy of even routing. The mean expertise in the system composed of one queue and two agents is given by the horizontal line in the middle of each plot. Sudden drops in expertise are due to randomly generated quitting events. We find that the results of Section 3 for utility function optimization still hold when turnover events are included in the model.

$$X_{n+1} = (X_n - \alpha \cdot (1 - X_n)) \cdot \theta(p)^- \tag{6}$$

$$X_\infty = \frac{\alpha}{\theta(p)^+ + \alpha - 1}. \tag{7}$$

By letting $\gamma = (\theta(p)^- - \alpha\theta(p)^-)$, and $d_n = X_\infty - X_n$, we can use Equation (7) to rewrite (6) as

$$(X_\infty - X_{n+1}) = \gamma \cdot (X_\infty - X_n) \tag{8}$$

$$d_{n+1} = \gamma \cdot d_n \tag{9}$$

If we let $d_0 = \alpha\theta(p)^-$, then $X_\infty = d_0 \sum_{n=0}^{\infty} \gamma^n$.

Define q be the probability that an agent quits or requires retraining during the interarrival interval $(n, n+1]$; let q be exponentially distributed and independent on each interval. Then we can define an expertise value $X_{p,q}$ that also depends on q.

$$\begin{aligned} X_{p,q} &= d_0 \sum_{n=0}^{\infty} (\gamma \cdot (1-q))^n \\ &= \frac{\alpha}{(\theta(p)^+ + \alpha - 1) + q - q\alpha} \\ &= \frac{\alpha}{(e^{+\beta\tau_{idle}} + \alpha - 1) + q(1-\alpha)} \end{aligned} \tag{10}$$

To see how q discounts expertise through Equation (10), consider these cases. If $q = 0$, we recover the expertise value without turnover, X_∞. If $q = 1$, $X_{p,q} = \alpha\theta(p)^-$, the value after serving one customer and quitting. For $\tau_{idle} = 2$, and $q, \alpha,$ and β all set to 1e-3, we have $X_\infty = 0.33$, but $X_{p,q} = 0.25$.

In a contact center, the expected value of elapsed time between turnover events, say T_q, would be estimated as an average over all agents. Since q is specified per interarrival interval, its distribution with respect to T_q must be calibrated for each agent's unique, routing-dependent interarrival time. Assuming turnover events occur as a Poisson process within each interval, the probability that more than zero turnover events occur is $q = 1 - P(\text{ 0 events } | \ T_q \ )$, or $q = 1 - e^{-1/(p\lambda T_q)}$. For a typical call center T_q will span hundreds to many thousands of customer service encounters, so q will be a small value.

We will forego presenting turnover results in Section 4 for multi-agent, multi-task scenarios. We only note here that our test results using Equation (10) instead of Equation (4) in those scenarios gave the same solution structure, but at appropriately lower levels of asymptotic expertise. However, an analysis of turnover effects using asymptotic expertise levels has certain limitations that we will discuss more in Section 5.

2.4 Expected Value of Expertise When Arrivals and Service are Exponentially Distributed

In Section 2.1 we derived a general expression for the asymptotic expertise level, Equation (4), that did not specify a distribution for the arrival rate λ or for the service rate μ. Here we apply the assumption underlying the most commonly used system capacity model, and take λ and μ to be exponential random variables: let interarrival times be $\sim$ Exp(λ), and let service times be $\sim$ Exp(μ). Then we determine the expected value of expertise as a function of these random variables. The resulting expressions provide insight, and may be used when arrivals and departures from a server are Poisson.

Let τ_a be the interarrival time of customers to this agent. Then $\tau_a \sim (p\lambda)e^{-p\lambda t}$.

$$
\begin{aligned}
X(n+1) &= (X(n) + \alpha \cdot (1 - X(n))) \cdot e^{-\beta\tau_a} \\
E[X(n+1)] &= (E[X(n)] + \alpha \cdot (1 - E[X(n)])) \cdot E[e^{-\beta\tau_a}] \qquad (11) \\
E[e^{-\beta\tau_a}] &= \int_{\tau_a=-\infty}^{+\infty} e^{-\beta\tau_a} \lambda p e^{-\lambda p \tau_a} d\tau_a \\
&= \frac{-\lambda p}{\beta + \lambda p} \int_{\tau_a=0}^{+\infty} e^{-\tau_a(\beta+\lambda p)} \cdot (-1) \cdot (\beta + \lambda p) d\tau_a \\
&= \frac{\lambda p}{\beta + \lambda p} \qquad (12) \\
E[X(n+1)] &= (E[X(n)] + \alpha \cdot (1 - E[X(n)])) \cdot \frac{\lambda p}{\beta + \lambda p}
\end{aligned}
$$

In the long run, $E[X(n)]$ settles down to the asymptotic value of expertise, $E[X(n+1)] = E[X(n)] = X_\infty(p)$. Letting $\theta = \frac{\lambda p}{\beta + \lambda p}$, we have:

$$
\begin{aligned}
E[X(n)] &= (E[X(n)] + \alpha \cdot (1 - E[X(n)])) \cdot \theta \\
&= \theta(E[X(n)] + \alpha\theta - \alpha\theta E[X(n)] \\
&= \frac{\alpha\theta}{1 - \theta + \alpha\theta} \\
&= \frac{\alpha\lambda p}{\beta + \lambda p - \lambda p + \alpha\lambda p} \\
&\boxed{E[X(n)] = EX = \frac{\alpha\lambda p}{\beta + \alpha\lambda p}} \qquad (13)
\end{aligned}
$$

To review, Equation (13) gives the asymptotic value of expertise when the arrival and service events are Poisson processes; the interarrival time between customers has a negative exponential distribution with rate $p\lambda$; and μ is insignificant compared to $p\lambda$, as when the encounter takes a few minutes, but the time between calls is measured in days. We expect the routing proportion $0 \leq p \leq 1$ to be fixed under management's control. Note that expertise tends towards the maximum value 1 when $\alpha\lambda p \gg \beta$.

The derivations of discounted expertise, Equations (6) to (10), and $E[X(n)]$, Equations (11) to (13), can be applied in a straightforward way to yield expertise level equations for other cases as well. Equation (13) when discounted for turnover is given by

$$E[X_{p,q}(n)] = \frac{\alpha\lambda p}{\beta + \alpha\lambda p q(1-\alpha)}. \qquad (14)$$

Let $\phi = p^2\lambda^2 + p\beta\lambda + p\mu\lambda$. Then when the service time *is significant* and affects the length of the idle time, the expected value becomes

$$E[X_\mu(n)] = \frac{\alpha \cdot \phi}{\alpha \cdot \phi + \beta\mu}. \qquad (15)$$

See Appendix 6.2 for a complete derivation of Equation (15). Finally, to determine expertise when including service time and discounting for turnover, Equation (15) becomes

$$E[X_{p,q,\mu}(n)] = \frac{\alpha \cdot \phi}{\mu\beta + \phi \cdot (q + \alpha - q \cdot \alpha)}. \tag{16}$$

Note that large values of β and μ increase the size of the denominator, and so decrease expertise. In this model, service that is extremely fast and efficient may increase the length of the idle time, and actually incur a greater loss of expertise due to forgetting.

3. Expertise Utility Functions of the Customer and the Firm

3.1 The Customer's Utility U_c, and the Supervisor's Utility U_s

The observation of a correlation between arrival rates and expertise leads to the natural question of how a call center should route calls to different agents. Consider the situation where all incoming jobs are divided between two agents, A_1 and A_2. Take λ to be the arrival rate of all jobs into the system. Parameter p_1 is the fraction of jobs routed to A_1, and $(1 - p_1)$ is the fraction routed to A_2.

We see that the value of p_1 chosen by our decision rule thus determines the two asymptotic expertise levels of the agents—and we can introduce the notation $X_1(p_1)$ and $X_2(p_1)$ to denote the dependence of asymptotic expertise on p_1. Equations (4), (13), or (46) may all be used to compute $X(p)$ here, but in order to take the most general approach we assume Equation (4) unless otherwise stated. Given this situation, we would like to know how one might select the ideal value for p_1.

Customers and firms have different objectives with respect to knowledge of the agents. Customers may prefer to have the maximum available service expertise; we will call a utility function that maximizes this objective the *customer's utility*, or U_c. This is a function of agent expertise as determined by the routing policy, so we will indicate that dependence using the notation $U_c(p)$.

Management, particularly those in charge of shift staffing, are on the other hand also interested in the overall knowledge and expertise available within the company. For example, having more than one trained agent mitigates the risk of one agent leaving (and taking their expertise with them). We refer to a utility function that maximizes the experience available as the *supervisor's utility*, or $U_s(p)$. Having agents with similar knowledge level leads to quality assurance whereby each customer receives equivalent service, which might be desirable. This argument favors $U_s(p)$.

Figures 4 and 5 illustrate the trade-off between customer and supervisor perspectives. As a simple example, let the customer's utility be $U_c(p_1) = E[X]$, where $E[\cdot]$ denotes the expectation

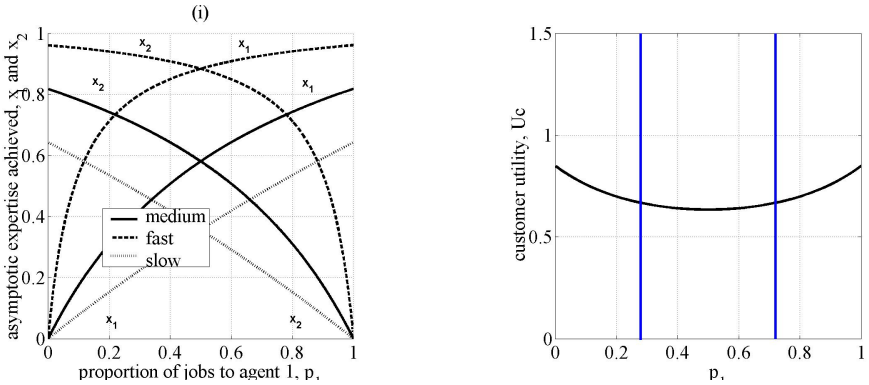

Figure 4: The customer's utility function U_c for the two agent case. Note the convex shape, with two optimal solutions residing at extreme values of the routing proportion p: $p = 0$, or $p = 1$. The vertical lines represent possible system constraints that limit the maximum utilization of an agent. Note that with the constraints, extreme points are still optimal, but the extreme values have been reduced.

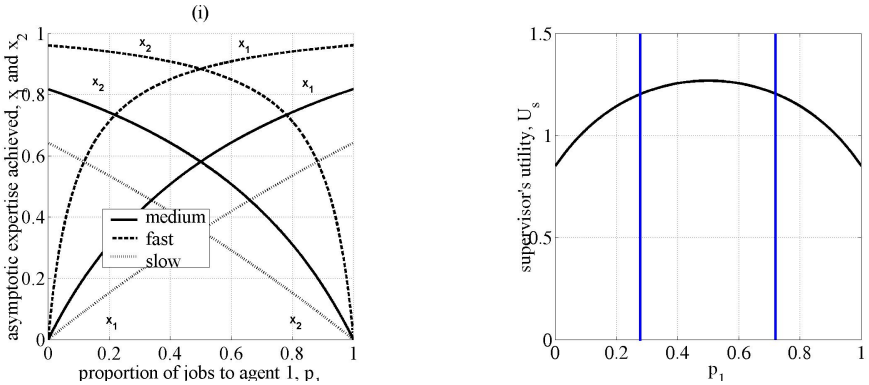

Figure 5: The supervisor's utility function U_s for the two agent case. Note the concave shape, with the optimal solution residing at the value where the routing proportions are equal: $p_1 = p_2 = 0.5$. The vertical lines represent possible system constraints that limit the maximum utilization of an agent. Note that with the constraints, the middle or even routing point is still optimal, and its value has been unaffected.

value; following the notation used in the figure, this is $E[X] = p_1 X_1(p_1) + (1 - p_1)X_2(p_1)$. Let the firm's utility be $U_f(p_1) = x_1(p_1) + x_2(p_1)$, corresponding to the total knowledge of the firm.

At the left, Figure 4 shows the asymptotic expertise attained by each of the two agents over the range of p_1. Here the forgetting rate for all pairs of curves is $\beta = 0.001$, and the learning rates from the top pair to the bottom pair are $\alpha = 0.011, 0.002$, and 0.0008—consider these *fast, medium,* and *slow* learning cases, respectively. The curve for the first agent grows with p_1, similar to the right side plot of Figure 2 in Section 2. As expected, the asymptotic expertise curve for agent 2 decreases in p_1. The right-hand plot shows the resulting customer's utility.

The left-hand side of Figure 5 repeats the two-agent expertise plot for reference, and the right-hand plot shows the supervisor's utility. Compare the plots of U_c and U_s, and note that the maxi-

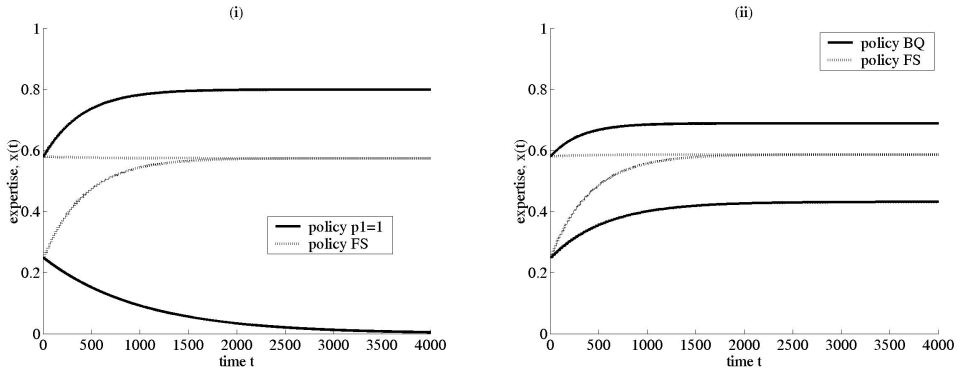

Figure 6: (i) Temporal evolution of the expertise levels $x_i(t)$ for the two agent case when always choosing the agent with maximum expertise (solid line) or splitting the jobs evenly between the agents, policy FS (dashed line). (ii) Temporal evolution of policies BQ (solid line) and FS (dashed line).

mum of the supervisor's utility U_s is a minimum of the customer's utility U_c. Further, note that if the firm chooses solely to increase the utility function for the customer, it destroys its own cumulative expertise.

For illustrating these two alternatives, assume that we have two agents with $X_1(0) = 0.58$, and $X_2(0) = 0.25$ for t=0, and consider the medium learning rate. Maximizing the customer's utility is equivalent to routing all jobs to agent 1 ($p_1 = 1$) whose expertise will increase while the expertise of agent 2 will decrease. Asymptotically, agent 1 will have an expertise of x=0.82, while the knowledge of agent 2 is zero. In Figure 6 (i), the temporal evolution is shown. In contrast, when choosing $p_1 = 0.5$, the company ends up with two equally trained agents.

In Figure 6 (ii) we compare how the expertise level of two agents will evolve in an M/M/2 queueing system with an average utilization of 28%. Each policy is work conserving, in that a call will never wait while an agent is free. They differ in that under best quality (BQ), if both agents are available the call is taken by the more proficient agent. This leads to $p_1 = 0.68$. In fair sharing (FS), the two agents will either alternate or be randomly assigned such a call with equal probability. As would be expected, BQ leads to one relative expert and one relative novice, while FS leads to two equally proficient agents.

The two-agent case thus provides an interesting insight: for maximizing the sum of the knowledge in the firm, it is better to route to the *less* experienced agents in order to give him or her the possibility to learn. This result is largely independent on the form of the learning-forgetting curves—as long as the increase of expertise ΔX by learning-on-the-job is a concave function as a function of X (less increase at higher expertise level), the gain of cumulative knowledge is always larger when routing to the less experienced agent. Thus, under a knowledge management

perspective, a balanced routing is always preferable.

3.2 Maximizing Utilities in the Many-Agent Case

3.2.1 Extreme Routing Optimizes the Customer's Utility

The following result says a routing rule that encourages specialization will maximize the expected value of expertise E[X], and hence the customer's utility function $U_c = E[X]$. In this section, we prove the N-agent case where the following conditions hold.

1. Let the asymptotic expertise value X_i of each agent i be a concave function of the proportion of customer traffic p_i that is routed to that agent. Let $X_i \in \Re^+$, and $0 \leq X_i \leq 1$.

2. We write $X_i(\alpha, \beta, \tau_a(p_i), \tau_s)$ to indicate the dependence of X_i on four parameters. The learning parameter α, the forgetting parameter β, and the mean service time τ_s are fixed values that are the same for all agents. Thus all agents have identical learning/forgetting curves.

3. The average interarrival interval $\tau_a(p_i) = (p_i\lambda)^{-1}$ is a function of fixed system arrival rate λ and routing fraction p_i, with $p_i \in \Re^+$, and $0 \leq p_i \leq 1$. Here p_i is the only independent variable, so we may write $X_i(p_i)$, and examine the properties of $X_i(p_i)$ as p_i is varied.

4. The first and second derivatives of $X_i(p_i)$ with respect to p_i exist.

5. Those derivatives satisfy the relationship $2X_i'(p_i) + p_i \cdot X''(p_i) \geq 0$.

6. An asymmetric routing rule is in force that routinely routes more traffic to one of the agents, say agent j.

Theorem 3.1. *Under the conditions stated above, the expected value of expertise seen by a customer, E[X], is decreased in steady state when work is transferred from the agent with the highest expertise to an agent with less expertise.*

Proof. Let $g(p_i) = p_i \cdot X(p_i)$. Then the expected value of expertise seen by an arriving customer in this system's steady state, $E_s[X]$, is given by

$$E_s[X] = p_1 \cdot X(p_1) + p_2 \cdot X(p_2) + ... + p_n X(p_n). \quad (17)$$

$$= g(p_1) + ... + g(p_n) \quad (18)$$

Recall that agent j receives a higher proportion of jobs than any other agent, so $p_j > p_i$, $\forall\ i \neq j$. Now, perturb the system to a new state using the routing rule to remove a small proportion of arrivals Δp from agent j's assignment, and add those arrivals to some other agent i's assignment. The new value of $E_s[X]$, or $E_{new}[X]$, becomes

$$E_{new}[X] = g(p_1) + ... + g(p_i + \Delta p) + g(p_j - \Delta p) + ... + p_n X(p_n) \tag{19}$$

Now we can analyze the difference $\Delta E = E_{new}[X] - E_s[X]$. If this difference is negative, the expected value of expertise dropped due to the perturbation. We can construct a first-order approximation to ΔE as follows:

$$\begin{aligned} g_i(p_i + \Delta p) &\approx g(p_i) + \Delta p \cdot g'(p_i) && (20)\\ g_j(p_j - \Delta p) &\approx g(p_j) - \Delta p \cdot g'(p_j) && (21)\\ \Delta E &\approx \Delta p \cdot (g'(p_i) - g'(p_j)). && (22) \end{aligned}$$

(Note that $X(p)$ is concave, so $X''(p) < 0$; because $p \geq 0$, we have that $p \cdot X''(p) \leq 0$.) Now $g(p) = p \cdot X(p)$, and $g(p)$ is convex if $g''(p) \geq 0$, so we can write:

$$g''(p) = 2X'(p) + pX''(p) \ \geq\ 0 \tag{23}$$

At the beginning we stipulated that Equation (23) was satisfied by $X(p)$. Therefore Equation (30) is negative due to the change in routing assignment Δp, and the theorem is proved.

□

Here is an example of the application of Theorem 3.1 using the asymptotic expertise function developed in Section 2.

Lemma 3.2. *The expertise function $X(p)$ given in Equation (4) satisfies the conditions of Theorem (3.1), under parameter ranges of interest in modeling learning and forgetting phenomena from Appendix 6.1. Therefore, for this function, $E[X]$ is maximized by extreme asymmetric routing distributions.*

Proof. Let the forgetting effect be given by $\theta(p) = e^{\beta \cdot ((p \cdot \lambda)^{-1} - (\mu)^{-1}}$. Now, we can apply the convexity test of Equation (23) to our expertise function $X(p)$.

$$\begin{aligned} g''(p) &\geq 0\\ 2X'(p) + pX''(p) &\geq 0 \end{aligned}$$

Expanding this using $X'(p)$ and $X''(p)$ from Appendix 6.1 gives:

$$\frac{2\alpha\beta\theta(p)}{p^2\lambda(\theta(p)+\alpha-1)^2}+\frac{2p\alpha\beta^2\theta(p)^2}{p^4\lambda^2(\theta(p)+\alpha-1)^3}-\frac{2p\alpha\beta\theta(p)}{p^3\lambda(\theta(p)+\alpha-1)^2}-\frac{p\alpha\beta^2\theta(p)}{p^4\lambda^2(\theta(p)+\alpha-1)^2}\geq 0.$$

Algebraically this reduces to the expression

$$\theta(p) \geq \alpha-1 \tag{24}$$

$$e^{\beta\cdot(\,1/(p\cdot\lambda)-\tau_\mu)} \geq \alpha-1 \tag{25}$$

Note that by construction $\alpha<1$, and $(p\cdot\lambda)^{-1}\geq\tau_\mu$; and so the left-hand side of (25) is greater than zero. Thus relation (25) will always hold for our asymptotic expertise function $X(p)$.

From the observations above and relation (30), we know that transferring some work Δp from a more proficient agent to a less proficient agent will always produce a negative change in $E[X]$.

□

3.2.2 Even Routing Optimizes The Supervisor's Utility

Here we alter Theorem 3.1 slightly to show that even routing is optimal for the supervisor's utility, $U_s(p)$. Let the same six preconditions hold as in Section 3.2.1.

Theorem 3.3. *When the firm's agents have an asymmetric distribution of expertise, then the sum of asymptotic expertise in the firm, S[X], is increased when work is transferred from an agent with the highest expertise to an agent with the lowest expertise.*

Proof. The sum of asymptotic expertise present in the firm, $S[X]$, is given by

$$S[X]=X_1(p_1)+X_2(p_2)+...+X_i(p_i)+...+X_j(p_j)+...+X_n(p_n) \tag{26}$$

We start with agent j receiving a higher proportion of jobs than any other agent, so $p_j>p_i$, $\forall\ i\ \neq\ j$. Now, perturb the system to a new state using the routing rule to remove a small proportion of arrivals Δp from agent j's assignment, and add those arrivals to some other agent i's assignment. The new value of $S[X]$, or $S_{new}[X]$, becomes

$$S_{new}[X]=X_1(p_1)+X_2(p_2)+...+X_i(p_i+\Delta p)+...+X_j(p_j-\Delta p)+...+X_n(p_n) \tag{27}$$

Now we can analyze the difference $\Delta S=S_{new}[X]-S[X]$. If this difference is positive, the sum of expertise in the firm increased due to the perturbation. We construct a first-order approximation to ΔS as follows:

$$X_i(p_i + \Delta p) \approx X_i(p_i) + \Delta p \cdot X_i'(p_i) \quad (28)$$

$$X_j(p_j - \Delta p) \approx X_j(p_j) - \Delta p \cdot X_j'(p_j) \quad (29)$$

$$\Delta S \approx \Delta p \cdot (X_i'(p_i) - X_j'(p_j)). \quad (30)$$

Note that $X_i(p_i)$ and $X_j(p_j)$ are concave functions that only differ in the independent variable p, and $p_i < p_j$. Thus $X_i'(p_i) > X_j'(p_j)$, and Equation (30) is positive due to the change in routing assignment Δp. This proves the theorem.

□

Lemma 3.4. *$U_s(p)$ is maximized by equal routing to all agents.*

Proof. Consider the sum of the expertise of a finite number of these agents who begin with an asymmetric distribution of expertise. That sum grows as the routing rule designer repeatedly implements routing changes as described in Equation (27), always taking away Δp from the agent with highest expertise, and giving it to the agent with the lowest expertise. After all possible changes are made, having increased $S[X]$ with each perturbation, there is no longer an agent with a higher expertise to choose from, and we find that $S[X]$ has been maximized by an even distribution of routing: $p_i = p_j \ \forall \{i, j\}$. □

Lemma 3.5. *When asymptotic expertise is given by $X_\infty(p)$ from Equation (4), $S[X]$ is maximized by equal routing to all agents.*

Proof. The expertise function $X_\infty(p)$ given in Equation (4) satisfies all the conditions of Theorem (3.3), and in particular it is concave under the parameter ranges of interest in modeling learning and forgetting phenomena from Appendix 6.1. Therefore sums of type $S[X]$ that are composed of expertise functions of type $X_\infty(p)$ are maximized by equal routing to all agents. □

4. Managing Expertise and Capacity in Contact Centers

In the day-to-day operation of a customer contact center, providing sufficient staffing levels of trained agents to handle call demand within contracted customer waiting time limits must be the primary concern. Skill development through on-the-job learning unfolds slowly, and is necessarily seen by line managers as a secondary priority to keeping customer waiting time low. Here we apply the theory developed in previous sections together with nonlinear optimization to explore

the interaction of expertise and capacity management goals: how much flexibility do we have in specific situations to shape the distribution of our agents' expertise? Given that flexibility, what is the optimal distribution?

Consider the *capacity* of a contact center to be the number of trained agents able to handle the incoming flow of customer traffic. If the traffic forecast is roughly correct, an M/M/C queueing system model gives a useful approximation of the waiting time performance. An M/M/C system inherently follows the supervisor's utility, U_s: every agent on average receives an equal share of the incoming workload.

In cases where we desire to improve the customer's utility U_c, we need to introduce a higher degree of task specialization within the system—to disaggregate the incoming job stream into subsets based on content, and route subsets of jobs to particular subgroups of agents. Yet this violates the assumption in the M/M/C model, and makes the waiting time analysis difficult. It also makes the waiting time worse, if the mean service times are constant. In order to specialize in a subset of the tasks, an agent must refuse to accept call types outside of his purview, and that may reduce the total capacity of the contact center. This may even force the agent to be idle if no calls of his assigned types are present. Thus the unfortunate side-effect of increasing the specialization of agents' work assignments is an accompanying increase in agent idle time, and customer waiting time.

As Pinker and Shumsky (2000) point out, the specialization versus cross-training trade-off varies according to the size of the organization. Very large centers with many customers have enough traffic to keep focused specialists busy. Small contact centers need cross-trained workers, and when learning rates play a significant role, it becomes important to optimize the blend of cross-trained and specialized workers.

Even in large centers, however, small groups of agents may be assigned ownership of specific customer inquiry types based on content. Therefore, optimizing expertise through smart routing rules may still benefit large centers, because with respect to work content, they may actually be a collection of small independent departments, each handling unique tasks, all under a common administrative umbrella.

Another property of knowledge-intensive contact centers is that agent utilization is surprisingly low. For high-end financial services firms we studied, the mean utilization of about 5,000 workers was between 30% and 40%, depending on which one of several geographically dispersed facilities the agent belonged to. This assumes eight-hour work shifts, and five-day work weeks. There are various reasons for lower utilization, including staff provisioning for peak traffic times, and

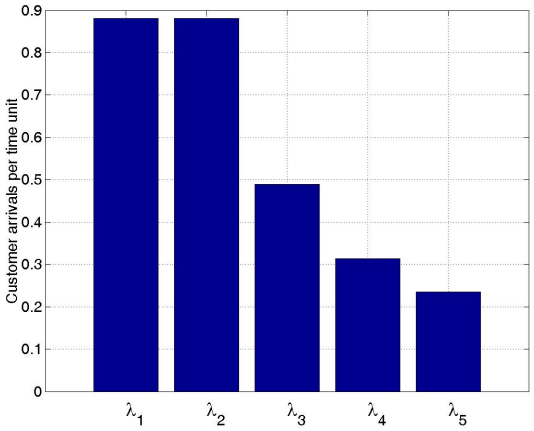

Figure 7: Test pattern of arrival rates for different job classes. This distribution is shown in Figure 11 as Case C. There are five job classes here, or learning dimensions, with arrival rates λ_j as shown. The service rate of each agent was set to $\mu = 1$. All arrival and service events are assumed be Markovian, with arbitrary time units.

automation that guides many customers to prerecorded answers. Note that in a low-utilization center, the balance of agents' time is occupied by background tasks, such as processing mail or email, research, updating web site answers for accuracy, and so on. Low utilization does not imply wasted time.

In addition, of course, there are peak periods when agents are fully utilized; but overall this indicates that there is a significant fraction of time when a customer arrives, and there is a *choice of several agents* to whom we may route the inquiry. To optimize the distribution of expertise among our agents, we take advantage of this opportunity to choose, and thus route customers to agents in a manner that develops and maintains expertise consonant with management's expertise development goals.

Here we define the optimal assignment with an objective function based on expertise—either U_c or U_s—and then constrain it using information about the system's capacity.

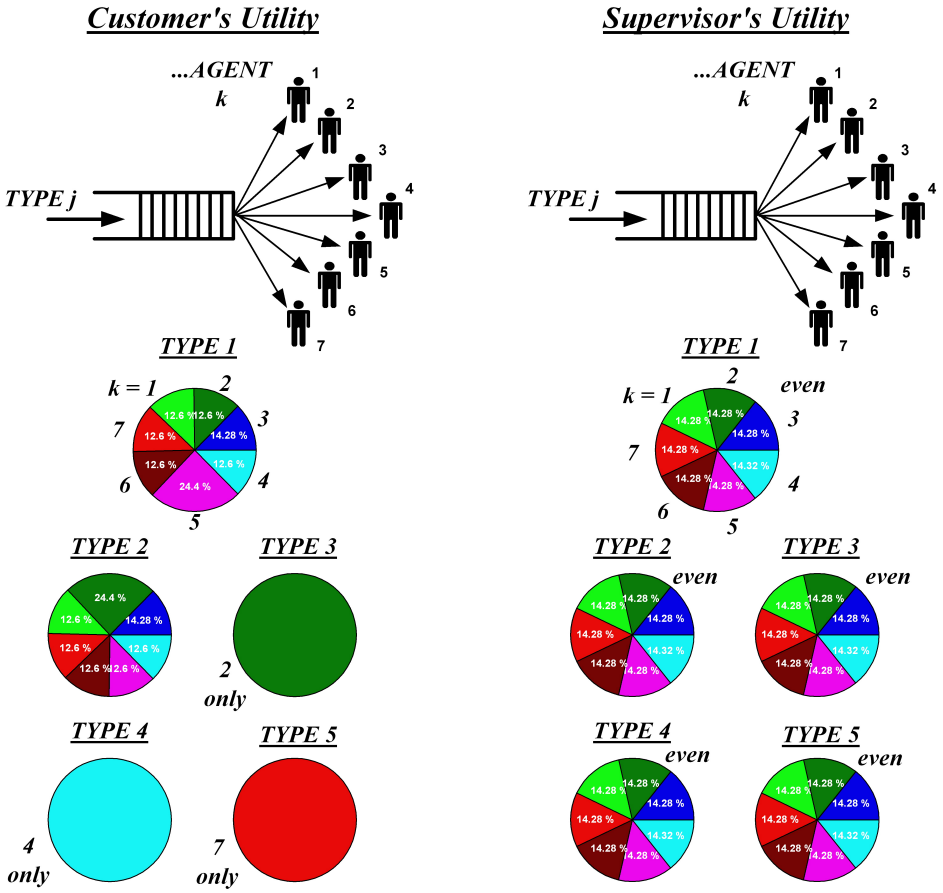

Figure 8: Left: details of the routing assignments for seven agents, using U_c as the routing objective. Each pie chart gives the proportion of the arrival rate of type j jobs assigned to agents 1 through 7. The arrival rates for the five types are shown in Figure 7. The arrows from queue to agents are drawn for one of the types, and the reader should imagine that five overlapping sets of arrows are present. Rates for types 1 and 2 are large, requiring all agents to share those tasks. Note that forced task sharing was not invoked for contact types 3, 4, and 5, allowing all of those customers to be routed to specialists under U_c. Right: the routing assignments using U_s as the objective. Note that all job types were routed in even proportions to all agents.

4.1 A Nonlinear Programming Approach with Sharing and Utilization Constraints

$$\max_{p_{jk}} \quad U_c = \sum_{j=1}^{J}\sum_{k=1}^{K} p_{jk} \cdot X_\infty(p_{jk}\lambda_{jk}) \tag{31}$$

such that

$$\forall \text{ j, k:} \quad \lambda_{jk} \cdot p_{jk} \geq \mathcal{L}_{jk} \quad \text{“capacity sharing”} \tag{32}$$

$$p_{jk} \geq 0 \tag{33}$$

$$\mathbf{L} \cdot \vec{p} \leq \vec{\mu} \quad \text{“utilization”} \tag{34}$$

$$\mathbf{L} \cdot \vec{p} = \vec{\lambda} \quad \text{“full service.”} \tag{35}$$

We divide the incoming tasks into J subgroups, each of which requires a skill set independent of other subgroups. For emphasis, we may refer to one of these subgroups as a *job class*, *task type*, *contact type*, or *learning dimension*. Then we will associate a separate learning curve for every agent k, for every job type j. The nonlinear program (NLP) with linear constraints of Equations (31) through (35) finds the best routing proportion p_{jk} of job type j to agent k. The objective chosen in Equation (31) is the customer's utility function; alternatively, the supervisor's objective of Equation (36) may be substituted here instead. These functions use Equation (4), so we will be reasoning about asymptotic expertise levels; in this section we may shorten the expression for asymptotic expertise from X_∞ to just X.

$$\max_{p_{jk}} \quad U_s \quad = \quad \sum_{j=1}^{J} \sum_{k=1}^{K} X_\infty(p_{jk}\lambda_{jk}) \qquad (36)$$

(such that...)

Three system constraints control the program's flexibility in choosing p_{jk}: the capacity sharing constraint, from Section 4.2; a utilization constraint, holding work assignment for every agent k to less than 100%, or $\sum_{j=1}^{J} p_{jk} \cdot \lambda_j \leq \mu_k$; and a full service constraint, such that all arriving jobs are assigned to some agent. Matrix $\mathbf{L}$ has K rows and $J \cdot K$ columns, vector $\vec{p}$ of p_{jk} values has $J * K$ elements, vector $\vec{\mathcal{L}}$ has K elements, vector $\vec{\mu}$ has K elements, and vector $\vec{\lambda}$ has J elements. We omit it here, but sometimes supervisors might prefer a fairness constraint be used as well, so the mean utilization of all agents is the same.

Here we assume management considers both U_c and U_s to have merit, and is interested in seeing the trade-offs involved in choosing either objective. Most of the work then involves computing and analyzing the customer's utility. It is easy to compute the even routing solution for U_s—just give every agent an equal share of each task type! As we saw in Figure 4, however, U_c has multiple optima, and any particular solution depends on the initial conditions we submit to the solver. Our initial condition is a vector of even routing probabilities, with a slight random bias added in to ensure that the solution for objective U_s consists of extreme points.

Given the distribution of arrival traffic in Figure 7, the NLP solver's routing assignments $\vec{p}$ for objectives U_c and U_s appear in Figure 8. This and subsequent experiments assume a 7-agent, 5-job class system. The system's utilization is 40% in this case, and there is some flexibility to choose routing assignments while meeting waiting time targets. Objective U_c's arrangement takes contact types not subject to forced sharing and develops specialists, and objective U_s's arrangement

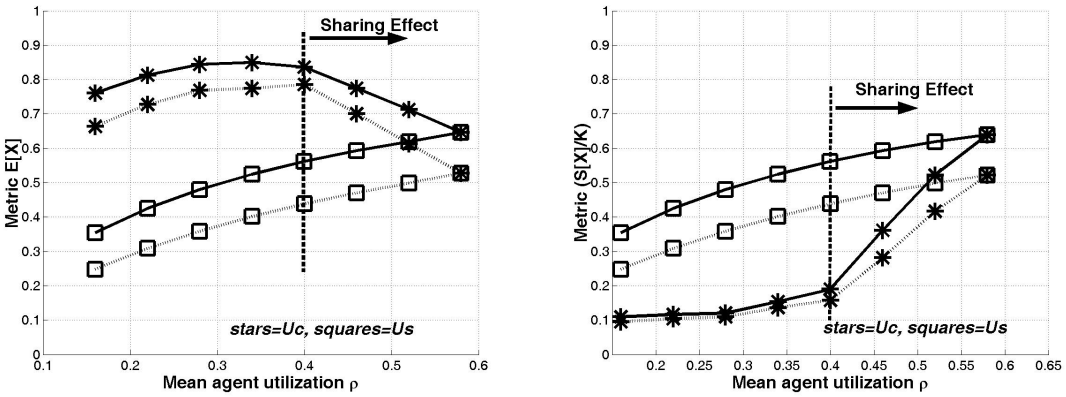

Figure 9: Left: Plot of the NLP output metric $E[X]$ versus system utilization ρ, when the NLP objective is U_c (stars) and U_s (squares). Right: Plot of the NLP output metric $S[X]$ versus system utilization ρ, when the NLP objective is U_c (stars) and U_s (squares). See Section 4.1 for details.

distributes jobs evenly. Note that if the service rates for all job types are the same, the assignment by objective U_s is equivalent to the *longest queue first* service discipline often used as the default routing policy in contact centers.

It should be noted that the routing rule targets given by the vector of proportions $\vec{p}$ can be implemented in many ways. For example, a strict schedule may be kept, and calls denied to agents who are ahead of schedule, so they do not take customers who would otherwise maintain the pre-planned level of their fellow agents' expertise. Yet denying any calls (i.e., forcing longer waits until the scheduled agents are available) would reduce the center's service level. Another implementation would be to set routing priorities based on $\vec{p}$, but never deny routing a call to an agent when higher priority agents are busy. Due to the stochastic nature of arrivals, the expertise targets based on $\vec{p}$ would not be met precisely, but the impact on service levels would be minimized. Due to space constraints our focus in this paper must remain the characterization of the NLP solutions, but we recommend thoroughly simulating alternate implementation strategies for $\vec{p}$ in practice to quantify these trade-offs.

Note that the customer's utility $U_c = E[X]$ and the supervisor's utility $U_s = S[X]$ may be objective functions to be optimized, or they may be performance metrics by which to judge the effects of routing assignments. To avoid confusion, from now on we will use the terms U_c and U_s when describing objectives, and $E[X]$ and $S[X]$ when discussing metrics.

Figure 9, on the left, shows a plot of the output result from our NLP solver of the output metric $E[X]$. Stars indicate the objective function was U_c; squares indicate the objective function was U_s. The top curves had a higher learning rate. See Table 2 for details of the parameters used. Starting

at the left of the curve, utilization grows as the arrivals to the system increase, and expertise also increases because there are more chances for agents to learn. However, midway the effect of the sharing lower bound $\mathcal{L}_{jk}$ manifests itself as well, forcing agents to share work assignments, and reducing opportunities for agents to gain specialized expertise. As the effect of $\mathcal{L}_{jk}$ grows for increasing ρ, capacity issues force the system to adopt even routing; then $E[X]$ drops and $S[X]$ rises. At the far right, the NLP solutions for objectives U_c and U_s have become the same.

4.2 The Capacity-Constrained Task Sharing Lower Bound

As Figure 9 shows, a key constraint in this setting is the lower bound on the number of jobs of a specific type that an agent must accept. If this lower bound is zero, we have the flexibility to assign some or none of this task type to this agent; in particular, this is useful when we desire to optimize the customer's utility U_c. If it is nonzero, we must assign at least that portion of type j tasks to the agent in order to meet the system's service level targets. We denote this lower bound by $\mathcal{L}_{jk}$. Due to its impact on routing flexibility, we call this quantity the level of *capacity-constrained task sharing*.

Given a number of different task types, individual agent assignments, and assumptions about the service and arrival time distributions, it may be a complex process to determine this lower bound precisely. A convenient means to make an estimate of $\mathcal{L}_{jk}$ for each job class j is as follows.

1. Design the system according to an M/M/C model, and set staffing levels to achieve the desired service level targets. See Gans et al. (2003) for a detailed discussion. According to their definitions, here we are dealing with systems in the *quality regime*.

2. Note that the probability that a customer waits for time t_w, given that she waits at all, is $\Pr\{t_w > t \mid t_w > 0) = e^{-t(\mu C - \lambda)}$.

3. Then the conditional expectation of this waiting time is

$$\begin{aligned} E[t_w > t \mid t_w > 0] &= ET_w(C, \mu, \lambda) = \int_{t=0}^{\infty} \tau \cdot (\mu C - \lambda) \cdot e^{-\tau(\mu C - \lambda)} d\tau \\ &= (\mu C - \lambda)^{-1} \end{aligned}$$

   For more details, see Gross and Harris (1998).

4. For each task type j, assume we assign a group of agents C_j to be full-time specialists who handle it, and these specialists have about the same utilization ρ as the rest of the agents. Then C_j may be found from $\rho = \lambda_j/(C_j\mu)$.

Type	λ_j	C_j	$\lambda - \lambda_j$	$ET_w(C - C_j, \mu, \lambda - \lambda_j)$	$\Delta ET_w > \eta$?	$\mathcal{L}_{jk}$
1	0.82	3.00	1.98	0.49	0.26	0.12
2	0.82	3.00	1.98	0.49	0.26	0.12
3	0.64	2.00	2.16	0.35	0.11	0
4	0.29	1.00	2.51	0.29	0.05	0
5	0.22	1.00	2.58	0.29	0.05	0

Table 1: A fast way to approximate what the sharing lower bound should be. This example uses the arrival pattern over learning dimensions of Figure 7. Here $\sum_{j=1}^{5} \lambda_j = 2.8$, $\mu = 1$, $C = 7$, $\rho = 0.4$, $\eta = 0.15$, and $ET_w(C, \mu, \lambda) = 0.238$.

5. If $ET_w(C - C_j, \mu, \lambda - \lambda_j) - ET_w(C, \mu, \lambda) > \eta$, for some cutoff η to be determined, then a policy of extreme routing to achieve high levels of specialized expertise in task type j will cause an unacceptable increase in waiting time.

6. For tasks that fail the test, set $\mathcal{L}_{jk} = \lambda_j / C$, or in our notation for the number of agents K, $\mathcal{L}_{jk} = \lambda_j / K$.

The goal of this procedure is to impinge only a small amount (determined by η) on the routing choices of the M/M/C system, in order to preserve its service-level performance. For tasks that make up a large proportion of the system arrivals, this will not be possible, and $\mathcal{L}_{jk}$ will be nonzero. Table 1 shows an example of how to estimate $\mathcal{L}_{jk}$. Again, this simple procedure is only a rough approximation, and we expect more thorough methods of waiting time analysis, including discrete-event simulation, to be used in applications.

By setting $\mathcal{L}_{jk} = \lambda_j / K$, we exclude one possible path to better solutions—we could instead assign the forced sharing to apply only to a subset of the agents, and try to find the optimal subset for each sharing group. On the other hand, extending the bound to all agents through our assignment $\mathcal{L}_{jk} = \lambda_j / K$ provides useful rostering flexibility for handling high-volume call types. In the next sections' results we retain our simple approach, and just note for future study the possibility of more combinatorial optimization work. See Iravani et al. (2007) for interesting new research along these lines.

4.3 Routing Rules Driven by the Distribution of Contact Types/Learning Dimensions

Because nonzero $\mathcal{L}_{jk}$ values force the system to follow a more even routing rule for type j tasks, its presence has an important implication for system-wide expertise optimization. If the system's traffic is defined by a small number of task types, each with a large enough arrival rate to force

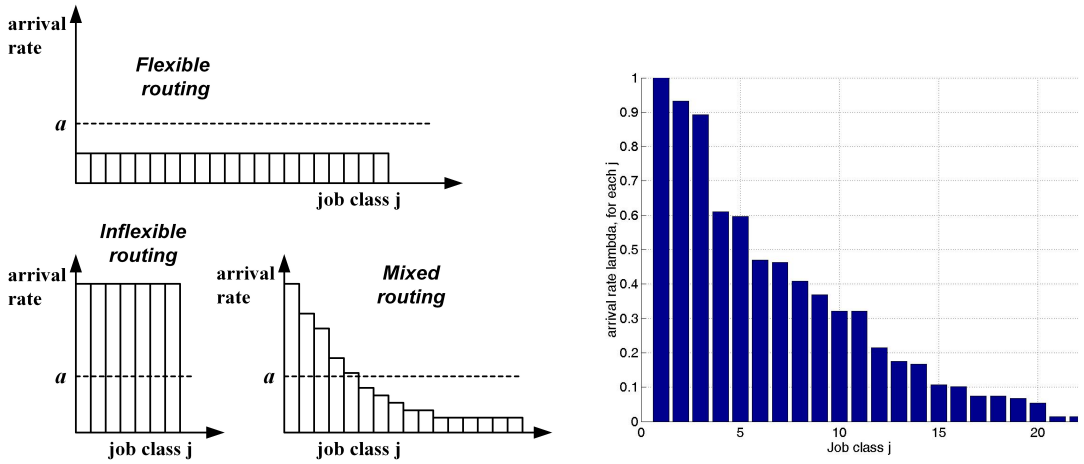

Figure 10: Left: Three hypothetical arrival patterns faced by a contact center. With respect to on-the-job learning, each job class represents a unique learning dimension, for which expertise must be developed separately. The service capacity-based limit a drives the lower bound of Section 4.3, and a combined with the arrival pattern determines the firm's flexibility in routing jobs to optimize agents' expertise. Right: Empirical pattern of arrival rates for different job classes for a financial services contact center.

a significant sharing lower bound, then there is no flexibility for specialization in routing assignments; the system defaults to even routing and the supervisor's utility U_s.

Figure 10 illustrates this dynamic. The constant a defines how much management will tolerate boosting specialized expertise and objective U_c at the cost of reducing service levels. The three sets of axes show three hypothetical arrival patterns faced by a contact center. With respect to on-the-job learning, each job class represents a unique learning dimension, for which expertise must be developed separately. At the top left, arrival rates are small enough that job classes do not need to be shared by all agents, permitting specialization, and allowing U_c to be optimized by extreme routing. At bottom left, arrival rates for each class are large, such that all job classes must be shared by all agents to meet waiting time goals. At right is an arrival pattern over learning dimensions seen in empirical contact center data. Some tasks with high arrival rates require a level of participation and proficiency by all agents; but others are small enough that service may be limited to selected agents, boosting the customer's utility U_c for those particular types.

Figure 11 summarizes our experiments related to the distribution of traffic over learning dimensions. The distributions at left appear with a dotted line to indicate the arrival rate λ_j above which forced task sharing is invoked. At right, the gain in the metric E[X] of objective function U_c over U_s appears in the first column, and the average value of the sharing lower bound in the second. Note these five cases of particular interest:

A. Here one learning dimension, λ_1, generates much more traffic than the others. The sharing

lower bound $\mathcal{L}_{1k}$ is so large that the agents have little spare capacity to become a specialist in the other tasks; all the other tasks must be shared as well. Given the precision limits of our NLP solver, a large enough λ_1 value he routing rule reverts to U_s.

B. One task generates just enough traffic to make $\mathcal{L}_{1k}$ nonzero. There is some forced sharing, but also significant flexibility to specialize in tasks λ_2 through λ_5.

C. Two lower bounds $\mathcal{L}_{1k}$ and $\mathcal{L}_{2k}$ are imposed on the agents, reducing the flexibility of agent k to specialize to $\mu_k - \mathcal{L}_{1k} - \mathcal{L}_{2k}$. In general, as more learning dimensions cross the forced sharing threshold, the NLP's flexibility to choose routing proportions decreases.

D. All learning dimensions are below the forced sharing threshold. $\mathcal{L}_{jk} = 0$ for all j; highly specialized task assignments are possible, and U_c can be very high.

E. All learning dimensions are above the forced sharing threshold, and the routing rule reverts to U_s.

When cases B, C, or D hold, management has the ability to boost the customer's utility if it wishes to. On the other hand, were it known that the distribution is A or E, management has no need to devise rosters of partial specialists to implement U_c—the firm will just end up adopting even routing and the supervisor's utility.

4.4 Routing Rules Driven by the Ratio of Learning to Forgetting Rates

From Equation (4), parameter α represents the learning rate and β the forgetting rate. Figure 12 demonstrates how the two expertise objectives change as the ratio α/β increases from 1 to 200. The test parameters were the same as those of Figure 9, with $\rho = 0.4$. At higher values of this ratio, learning is so fast that agents become experts from handling even a small amount of traffic, and so all agents may serve all contact types. Over most of this range, for this example extending from about 10 to over 100, the difference between the objectives on metric $S[X]/K$ is much greater than the difference on metric $E[X]$, making a policy of specialization less attractive.

When the ratio is less than about 10, and the impact of forgetting is high, ongoing maintenance of expertise levels through specialization and objective U_c is more attractive. With respect to on-the-job learning in call centers, these results suggest that specialization and extreme routing rules are most valuable when the forgetting rate is significant.

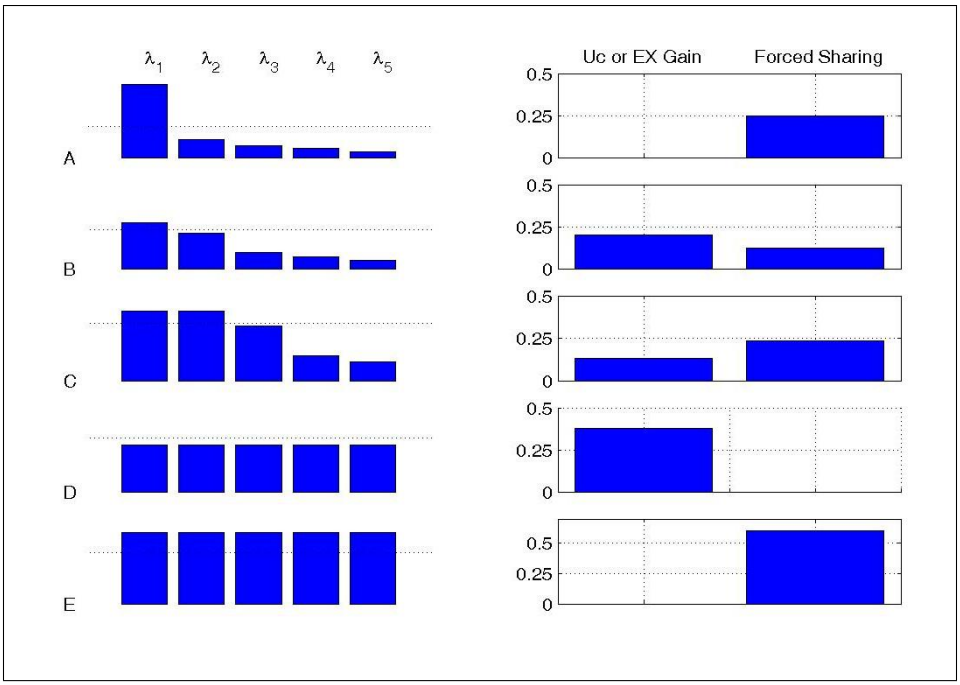

Figure 11: Left column: five traffic distributions, with the cutoff value for triggering task sharing shown as a dotted horizontal line. Right column: two quantities affected by the distribution. "$E[X]$ gain" is the expected value metric under objective U_c divided by its value under objective U_s, showing how much better a policy of specialization may do compared to a policy of even routing. "Forced sharing" is the average sharing lower bound per agent over all job types, indicating the proportion of an agent's service rate that must be dedicated to shared tasks. For distributions A through D the system utilization was 40%, and for E it was 60%. Note that the higher the sharing becomes, the lower the gain is reduced.

5. Conclusions, and Recommendations for Future Work

In this paper, we develop a method to quantify how task routing rules influence the long-run expertise of agents in a call center through on-the-job learning effects. We then prove how to obtain the optimal solutions for two conflicting objectives: the expected value of expertise seen by customers, or the *customer's objective*; and the sum of all the expertise within the firm, or the *supervisor's objective*. A policy whereby agents specialize in as few tasks as possible optimizes the expected value, while a policy of all agents handling all call types optimizes the sum.

Applying these insights, we describe a nonlinear programming solver to create routing rules of customer types to agents. This solver finds an optimal rule set that maximizes the distribution of expertise among agents within boundaries set by waiting time constraints, because call center management will only consider agent expertise targets to be useful if they do not significantly lengthen customer wait times. The most important means of communicating this capacity issue to the solver is through the *forced task sharing lower bound*. We see that a key driver of this lower

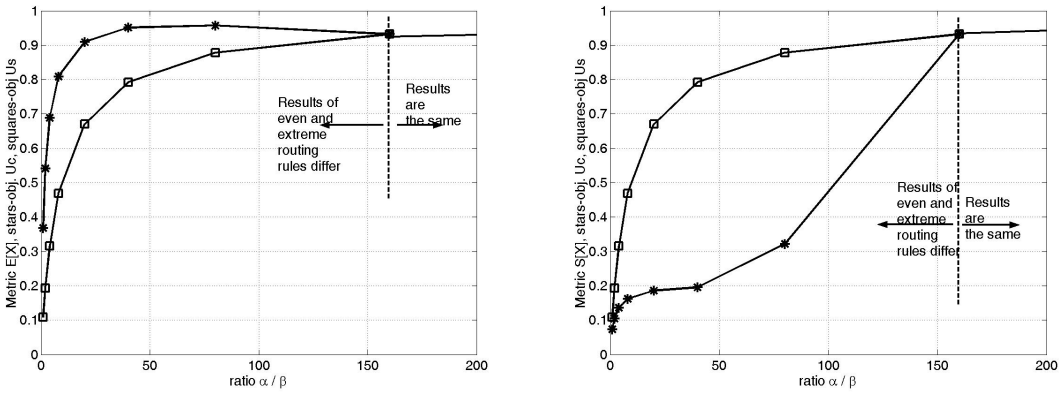

Figure 12: Left: Plot of the NLP output metric $E[X]$ versus the ratio of learning to forgetting rates α/β, when the NLP objective is U_c (stars) and U_s (squares). Right: Plot of the NLP output metric $S[X]/K$ versus α/β, when the NLP objective is U_c (stars) and U_s (squares). When learning is very fast compared to forgetting, both even routing and extreme routing result in about the same level of expertise.

bound is the distribution of traffic over the learning dimensions defined by the various task types, and we specify what mix of specialized and balanced routing rules is possible given a particular shape for this distribution.

Our major finding is the following: because the customer's utility for expertise is optimized by task specialization, any improvement in the customer's utility works against the waiting time reduction made possible by server pooling. Nevertheless, given the right traffic distribution, it is possible to introduce some specialization within capacity constraints to boost the customer's utility along selected learning dimensions.

These results suggest a number of useful areas for future work. First, the forced sharing lower bound may be made more precise, which will improve the quality of the solver's routing assignments. Second, a mixed objective function that rewards the expected value of expertise but penalizes the variance in expertise levels seen by customers may be able to create useful blends of balanced and specialized routing. Similarly, a minimax objective may be able to maximize the lowest asymptotic value of expertise that a customer would see. Third, there are many ways to implement the solver's assignments, causing more or less impact on customer waiting times. These should be explored using analytical methods and simulation, and the results used to improve the sharing lower bound.

Finally, our discussion of the effect of agent and task turnover is limited here because we only consider asymptotic expertise levels. Sometimes skilled agents quit, and inexperienced agents must take over their work. Such transient cases complicate our performance analysis of balanced

Default Parameters Used						
# Contact Types, J	# Agents K	Learning Rate α	Forgetting Rate β	Sharing Parameter η	System Utilization ρ	Service Rate Per Agent
5	7	3.5e-3	5e-4	0.15	0.4	1 (arbitrary time units)
Constraint Tolerance	Objective Fun. Tolerance	Output Vector $\vec{p}$ Tolerance	Maximum Solver Iterations	Solver Type		
1e-8	1e-8	1e-8	1000	Sequential Quadratic Program with Line Search		

Figure 10: Study of $E[X]$ and $S[X]$ Versus Utilization ρ			
Utilization Values	*α, top*	*α, bottom*	*Other Parameters*
0.16, 0.22, 0.28, 0.34, 0.4, 0.46, 0.52, 0.58	6e-3	3.5e-3	Default

Figure 12: Study of Arrival Distributions	
Arrival Rate Distribution	*Other Parameters*
1.74, 0.43, 0.28, 0.21, 0.14	Default
1.09, 0.84, 0.39, 0.29, 0.19	Default
0.82, 0.82, 0.64, 0.29, 0.22	Default
0.56, 0.56, 0.56, 0.56, 0.56	Default
0.84, 0.84, 0.84, 0.84, 0.84	Default

Figure 13: Study of $E[X]$ and $S[X]$ Versus Utilization ρ		
Forgetting Values	*α*	*Other Parameters*
1e-5 * (400, 200, 100, 50, 20, 10, 5, 2.5)	4e-3	Default

Table 2: Summary of parameters used in experiments.

versus extreme routing rules. One straightforward extension would be to create a discrete Markov chain using sampled versions of our learning-forgetting experience curves, together with states for task and agent quitting. Then appropriate penalties or rewards may be assigned to the states. Multiplying the stationary distribution and the reward vector would provide a measure of turnover's impact.

All too often on-the-job learning is considered to be an exogenous parameter by designers of call center operations. We hope to contribute towards a better understanding of this phenomenon, so that call center operators may include it in their planning process and use it as a means of competitive advantage in the markets they serve.

6. Appendix

6.1 Appendix: The Concave Property of Asymptotic Expertise

Under the range of parameters that are of interest in studying on-the-job productivity changes, the expertise function of Equation (4) is a concave function in p, as Figure 2 suggests. This concave property is a way of capturing the well-documented observation that real learning curves exhibit

diminishing returns with increased production. For example, empirical results have been fit to functions that show equal increments of improvement for every doubling of a worker's cumulative production. The arguments below discuss the conditions under the asymptotic expertise model of Section 2.1 is concave.

To simplify notation, let $\theta(p) = e^{\beta \cdot ((p \cdot \lambda)^{-1} - (\mu)^{-1})^+}$. Then the expertise function and its derivatives with respect to p are given by:

$$X(p) = \frac{\alpha}{\theta(p) + \alpha - 1} \tag{37}$$

$$X'(p) = \frac{\alpha\beta\theta}{(\theta(p) + \alpha - 1)^2 \; p^2\lambda} \tag{38}$$

$$X''(p) = \frac{\alpha\beta\theta(p)[-\beta\theta(p) + 2p\lambda\theta(p) + 2p\lambda\alpha - 2p\lambda + \alpha\beta - \beta]}{(\theta(p) + \alpha - 1)^3 \; p^4 \; \lambda^2} \tag{39}$$

Applying the condition $X''(p) \leq 0$ allows us to identify the range of parameters needed for $X(p)$ to be concave, resulting in the following relationship between learning parameter α, forgetting parameter β, job routing assignment $(p\lambda)^{-1}$, and service rate μ.

$$\alpha \geq 1 - \frac{\theta(p) \; (2p\lambda - \beta)}{(2p\lambda + \beta)} \tag{40}$$

Let $\beta = K \cdot p \cdot \lambda$; then (40) becomes

$$\alpha \geq 1 - e^{K(1-(p\lambda/\mu))} \cdot \left[\frac{(2-K)}{(2+K)}\right]. \tag{41}$$

For modeling learning-based performance in environments such as call centers, we would like the asymptotic value of expertise to be achieved after cumulative production N_c has reached tens, hundreds, or thousands of jobs. N_c is inversely proportional to α, so α will typically range between $a_{hi} = 0.05$ and $a_{low} = 1e-4$.

With this in mind, Inequality (41) defines a range of α values for which the asymptotic expertise function is concave. For $a_{low} \leq \alpha \leq a_{hi}$, we have $K < 2$, and $\beta < 2p\lambda$. We take a single agent's utilization to be 100% or less, so $(p\lambda/\mu) \leq 1$. Under these conditions, Inequality (41) always holds, and thus expertise is concave, for $\alpha \geq \beta$. In fact, at lower utilization rates β can grow much larger than α and still preserve concavity; but unless otherwise indicated, we will only consider examples of asymptotic expertise where the learning rate α is equal to or greater than the forgetting rate β.

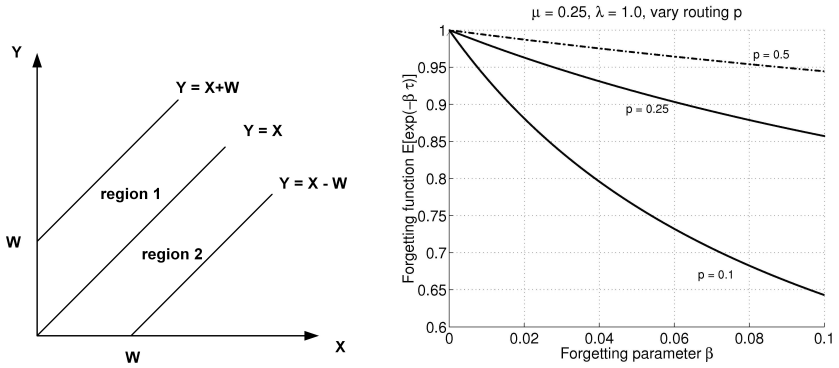

Figure 13: Left: regions of integration for the integral of $W = |X - Y|$. Right: behavior of the forgetting function of Equation (45).

6.2 Appendix: Expertise Based on Idle Times

In some cases we would like to base the amount of forgetting on the length of idle time τ_I, rather than the interarrival time τ_a, so expertise is not lost during the service time. To do this, we can adapt a well-known distribution for the difference between two exponential random variables (this assumes arrivals and service are Markovian). First, consider $X \sim \nu_1 e^{-\nu_1 x}, Y \sim \nu_2 e^{-\nu_2 y}$, and $W = |X - Y|$. Then we can find the density function of W with the help of Figure 13 (left), as follows.

$$\begin{aligned}
F_W(w) &= P\{|X - Y| \le w\} \\
&= P\{X - w \le Y \le X + w\} \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy \\
&= \int_0^{\infty} \left[\int_x^{x+w} \nu_1 e^{-\nu_1 x} \nu_2 e^{-\nu_2 y} dx \right] dy + \int_0^{\infty} \left[\int_y^{y+w} \nu_1 e^{-\nu_1 x} \nu_2 e^{-\nu_2 y} dx \right] dy \\
&= \frac{\nu_1}{\nu_1 + \nu_2}(1 - e^{-\nu_1 w}) + \frac{\nu_2}{\nu_1 + \nu_2}(1 - e^{-\nu_2 w}) \qquad (42)
\end{aligned}$$

Taking the derivative of Equation 42 gives the distribution of W:

$$f_W(w) = \frac{\nu_1}{\nu_1 + \nu_2} \nu_2 e^{-\nu_2 w} + \frac{\nu_2}{\nu_1 + \nu_2} \nu_1 e^{-\nu_1 w} \qquad (43)$$

Recall that the idle time is given by $\tau_I = (\frac{1}{p\lambda} - \frac{1}{\mu})^+$, where $p\lambda$ is the mean interarrival time and $1/\mu$ is the mean service time. We expect that most of the random arrival intervals will be greater than most of the random service times; the superscript "+" indicates that τ_I is never negative—but it may occasionally be zero when a service encounter lasts so long that another customer ready to be

seen by this agent. We can adapt Equation 43 to describe this by altering it to be a mixed random variable, with a discrete probability mass at the event $\tau_I = 0$. Then we can determine the expected value for our negative exponential forgetting function, $E[e^{-\beta\tau_I}]$.

$$\begin{aligned} f_{\tau_I}(\tau_I) &= \frac{\nu_2}{\nu_1+\nu_2}\nu_1 e^{-\nu_1\tau_I} + \frac{\nu_1}{\nu_1+\nu_2} \\ E[e^{-\beta\tau_I}] &= \int_0^\infty e^{-\beta\tau_I} f_{\tau_I}(\tau_I) d\tau_I \\ &= e^{-\beta\cdot 0}\frac{\nu_1}{\nu_1+\nu_2} + \int_0^\infty e^{-\beta\tau_I}\frac{\nu_2}{\nu_1+\nu_2}\nu_1 e^{-\nu_1\tau_I} d\tau_I \end{aligned}$$

$$E[e^{-\beta\tau_I}] = \frac{\nu_1}{\nu_1+\nu_2}\left(1 + \frac{\nu_2}{\nu_1+\beta}\right) \tag{44}$$

Now let $\nu_1 = p\lambda$, and $\nu_2 = \mu$ in Equation (44), so we can represent the *expected* impact of forgetting on the agent's expertise level as:

$$\boxed{E[e^{-\beta\tau_I}] = \frac{p\lambda}{p\lambda+\mu}\left(1 + \frac{\mu}{p\lambda+\beta}\right)} \tag{45}$$

Figure 13 (right) shows the behavior of Equation 45.

Consider $\theta = E[e^{-\beta\tau_I}]$, as given by Equation 45. Then the asymptotic value of expertise becomes:

$$\begin{aligned} \theta &= \frac{p\lambda}{p\lambda+\mu}\left(1 + \frac{\mu}{p\lambda+\beta}\right) \\ \theta^{-1} &= \frac{(p\lambda+\mu)(p\lambda+\beta)}{p\lambda(p\lambda+\beta+\mu)} \\ X_\infty &= \frac{\alpha}{\theta^{-1}+\alpha-1} \\ &= \frac{\alpha}{\frac{(p\lambda+\mu)(p\lambda+\beta)}{p\lambda(p\lambda+\beta+\mu)}+\alpha-1}. \end{aligned}$$

$$\boxed{X_\infty(p) = \frac{\alpha(p^2\lambda^2+\beta p\lambda+\mu p\lambda)}{\alpha(p^2\lambda^2+\beta p\lambda+\mu p\lambda)+\beta\mu}.} \tag{46}$$

This result is more complex than Equation (13) due to the introduction of the service rate μ. Now forgetting occurs only when the agent is idle. Note that the impact of forgetting grows as the product $\beta\mu$ gets larger—when the service rate μ is high, the agent is idle for longer periods of time.

References

Aksin, Z., M. Armony, V. Mehrotra. 2007. The modern call center: a multi-disciplinary perspective on operations management research. *Production and Operations Management Science* **16**(6) 665 - 688.

Aksin, Z., F. de Vricourt, F. Karaesmen. 2008. Call center outsourcing contract analysis and choice. *Management Science* **54**(2) 354 - 368.

Avramidis, A., P. L'Ecuyer. 2005. Modeling and simulation of call centers. *Proceedings of the 2005 Winter Simulation Conference*. Dec. 4-8, Orlando, Florida. 611–620.

Badiru, A. 1992. Computational survey of univariate and multivariate learning curve models. *IEEE Transactions on Engineering Management* **39**(2) 176–187.

Bodreau, J., W. Hopp, J. McClain, L. Thomas. 2003. On the interface between operations and human resource management. *Manufacturing and Service Operations Management* **5**(3) 179 – 202.

Bordoloi, S. 2004. Agent recruitment planning in knowledge-intensive call centers. *Journal of Service Research* **6**(4) 303 – 323.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2002. Statistical analysis of a telephone call center: a queueing science perspective. Technical Report, Wharton Financial Institutions Center, University of Pennsylvania, Philadelphia, PA.

Cleveland, B., J. Mayben. 2000. *Call Center Management on Fast Forward: Succeeding in Today's Dynamic Inbound Environment*. Call Center Press, Annapolis, Maryland.

Eitzen, G., D. Panton and G. Mills. 2004. Multi-skilled workforce optimization. *Annals of Operations Research*. **127**(1) 359 – 372.

Froehle, C. 2006. Service personnel, technology, and their interaction in influencing customer satisfaction. *Decision Sciences* **37**(1) 5–38.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: tutorial, review, and research prospects. *Manufacturing and Service Operations Management* **5**(2) 79–141.

Gans, N., Y. Zhou. 2002. Managing learning and turnover in employee staffing. *Operations Research* **50**(6) 991 – 1006.

Gans, N., Y. Zhou. 2003. A call-routing problem with service-level constraints. *Operations Research* **51**(2) 255 – 271.

Ghemawat, P. 1985. Building strategy on the experience curve. *Harvard Business Review* **63**(2) 143 – 149.

Globerson, S., N. Levin. 1987. Incorporating forgetting into learning curves. *International Journal of Operations and Production Management* **7**(4) 80–94.

Gross, D., C. Harris. 1998. *Fundamentals of Queueing Theory*. Wiley Interscience, New York, NY, page 73.

Hasija, S., E. Pinker, R. Shumsky. 2005. Staffing and routing in a two-tier call center. *Int. J. Operational Research* **1**(1/2) 8 – 29.

Hasija, S., E. Pinker, R. Shumsky. 2008. Call center outsourcing contracts under information asymmetry. *Management Science* **54**(4) 793 – 807.

Howick, S., C. Eden. 2007. Learning in disrupted projects: on the nature of corporate and personal learning. *International Journal of Production Research* **45**(12) 2775–2797.

Iravani, S., B. Kolfal, M. Oyen. 2007. Call-center labor cross-training: it's a small world after all. *Management Science*. Vol. 53, No. 7, pp.1102–1112.

Koole, G. (1997) 'Assigning a single server to inhomogeneous queues with switching costs',*Theoretical Computer Science*, Vol. 182, No. 1, pp.377–332.

Nembhard, D.A., N. Osothsilp. 2001. An empirical comparison of forgetting models. *IEEE Transactions on Engineering Management* **48**(3) 283–291.

Parasuraman, A., V. Zeithaml, L. Berry. 1988. SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing* **64**(1) 12–40.

Pinker, E., R. Shumsky. 2000. The efficiency-quality trade-off of cross-trained workers. *Manufacturing and Service Operations Management* **2**(1) 32–48.

Reis, D. 1991. Learning curves in food services. *Journal of the Operational Research Society* **42**(9) 623 - 629.

Ren, Z., Y. Zhou. 2008. Call center outsourcing: coordinating staffing level and service quality. *Management Science* **54**(2) 369 – 383.

Ross, S. 2002. *Simulation*. Elsevier, San Diego, CA.

Sayin, S., S. Karabati. 2007. Assigning cross-trained workers to departments: a two-stage optimization model to maximize utility and skill improvement. *European Journal of Operational Research* **176**(3) 1643 – 1658.

Schilling, M., P. Vidal, R. Ployhart, A. Marangoni. 2003. Learning by doing something else: variation, relatedness, and the learning curve. *Management Science* **49**(1) 39 – 56.

Shafer, S., D. Nembhard, M. Uzumeri. 2001. The effects of worker learning, forgetting, and heterogeneity on assembly line productivity. *Management Science* **47**(12) 1639–1653.

Shumsky, A., E. Pinker. 2003. Gatekeepers and referrals in services. *Management Science* **49**(7) 839 – 856.

Sikström, S., M. Jaber. 2002. The power integration diffusion model for production breaks. *Journal of Experimental Psychology: Applied* **8**(2) 118–126.

Vericourt, F., Y. Zhou. 2005. Managing response time in a call-routing problem with service failure. *Operations Research* **53**(6) 968 – 981.

Whitt, W. 2006. The impact of increased employee retention on performance in a customer contact center. *Manufacturing and Service Operations Management* **8**(3) 235–252.

Zohar, E., A. Mandelbaum, N. Shimkin. 2002. Managing learning and turnover in employee staffing. *Management Science* **48**(4) 566 – 583.