

Knowledge Management in Call Centers: How Routing Rules Influence Expertise in the Presence of On-the-Job Learning

Geoff Ryder

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

geoff.ryder@sap.com

Service agents in a call center are assigned to help customers by routing policies that seek to balance several objectives. Usually, these policies follow myopic rules in order to minimize the waiting time or maximize the quality experienced by the next customer. However, there is a secondary effect of the routing assignment: by learning-on-the-job, the development of the agents' expertise depends on the calls they take.

In this paper we address the effect that routing rules have on agent learning. We develop a nonlinear optimization framework for two kinds of expertise objectives: one that seeks equal distribution of experience across the workforce (effectively cross-training), and one that aims to develop specialized expertise by prioritizing the routing of specific customer inquiries to specific agents. Analytical models of call center operations are inadequate to handle this task, so instead we turn to discrete-event simulation, and evaluate the effect of routing policies on agent expertise with a custom simulator developed in the ExtendSim modeling environment. Simulation results describe an efficient frontier in routing policies that depends on the underlying expertise objective function.

1. Introduction: Modeling On-the-Job Learning

1.1 Motivation for an Optimization/Simulation Approach

Suppose an operations manager determines that on-the-job learning is significant among the workforce of agents she is responsible for. She observes her agents developing expertise with additional exposure to specific calls, exhibiting trends such as those of Figure 1 . How may she best take advantage of those trends to improve her group's performance? We provide a two-part answer to this question for call centers operating at medium utilization levels, where decisions about learning involve the most tradeoffs—assume a call center with average agent utilization rates between 30% and 60%, belonging to the *quality-driven regime* (see Gans et al. [2003], page 100). Empirical data suggest that such utilization levels are realistic over the time scale at which learning effects should be managed, e.g. multiple days or weeks.

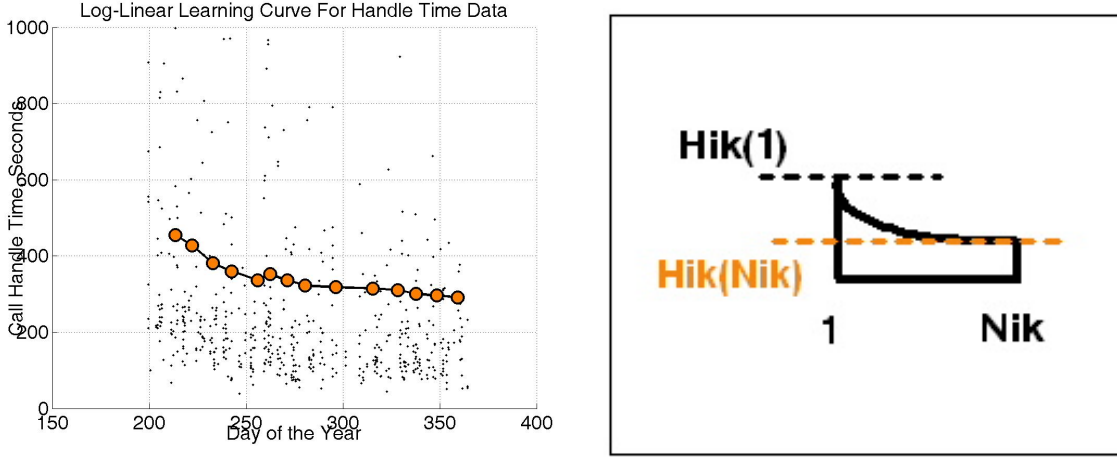


Figure 1: An example of cumulative average handle time performance improving with experience. Left: an agent gets faster and decreases her average handle time from 8 minutes to 5 minutes, a $\approx 40\%$ improvement, after handling about 560 calls over five months. Right: our model of this process, the log-linear learning curve, where the call handle time for agent i and call type k is a decreasing function of cumulative production, $H_{ik}(N_{ik}) = H_{ik}(1) \cdot N_{ik}^{-b_{ik}}$. The learning exponent in this case is $b = 0.07$.

For simplicity, we focus on improvements in call handle time with cumulative expertise, where handle time is assumed to start slow and to get faster with experience. We have several reasons for selecting the handle time–cumulative experience model here. First, we noted that trends of handle time getting faster with experience were common among the agents and call types observed in empirical data provided by our industrial partners. Trends of learning with cumulative production were significant on a wide scale. Improvements in handle time may also have several interpretations, making them useful in different contexts. For instance, they may be taken to be improvements in service quality—a fast response may indicate a knowledgeable agent. They also have a noticeable operational impact because they increase the call center’s capacity, and lower customer waiting times, which can be measured in queueing simulations.

Section 2 discusses mathematical programming approaches for generating optimal work assignments. These assignments are in the form of call routing targets for each agent that depend on their skill and expected learning curve in cumulative production for each call type. To be useful, the routing targets need accurate assessments of agents’ potential to improve—derived for instance from human resources data about an agent’s education, background, and training period observations—and an accurate demand forecast for incoming calls. We will assume sufficient accuracy on both measures for our purposes here.

Section 3 describes routing rules that implement these optimal routing targets in a stochastic model of a small contact center, where customer arrival times and service times are randomly distributed. In the normal case of multiple agents serving multiple queues, the observed inherent randomness of these systems will cause agents to be ahead of schedule with respect to some routing targets, and behind with respect to others, at any single point in time. Recognizing this, we define five routing rules that control agent activity with respect to their pre-set targets in different ways. We then conduct experiments using discrete-event simulations to evaluate the rules along dimensions of expertise development, variation of expertise among agents, and maintenance of

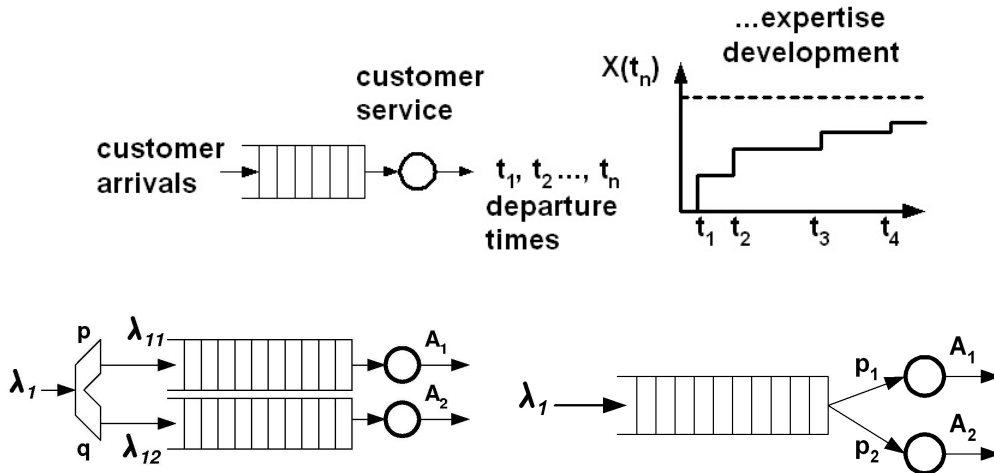


Figure 2: Top: Plot of expertise development for a single agent responsible for a queue of customers. Here, the agent’s expertise develops according to the irregular stair-step plot, with an increase recorded at the instant of each customer departure t_n . The amount of this increase diminishes as the production record grows, and is bounded above by the dotted line, which represents a perfect expertise value of one. Bottom left: an analytically tractable arrangement where each agent handles her own M/M/1 queue. Bottom right: a more realistic model where agents share responsibility for a single queue. But because of the difficulty of finding analytical results for priority M/M/c systems, we turn to randomized simulations enabled by ExtendSim to investigate the realistic model at bottom right.

waiting-time performance. Taken together, the optimization and simulation approaches provide a contact center manager with a toolkit to set both routing targets and target-implementing rules such that the development of agents’ expertise can be fully incorporated in operational planning.

1.2 Embedding Expertise Development in Queueing Models—Some Limitations

Contact centers employ a large workforce of agents. Tractable analytics for such multi-agent systems require that we use models like that on the bottom left of Figure 2, where each agent is responsible for service to a separate M/M/1/∞ queue. Yet in practice contact centers assign multiple agents to the same queue to keep waiting times low. To model waiting times accurately, we must turn to the M/M/c system on the right (Cleveland [2000], pages 89–90), or to related models such as M/G/c—in that case the general service time distribution G is often taken to be the lognormal distribution, following empirical results (Gans et al. [2003], page 127).

M/M/c systems are the most tractable of the family of multi-server queueing system models. But as with all the others, it is difficult to compute and write down expressions for steady state results where routing proportions p are different for every agent; further, the mean service time must remain stationary (Gross and Harris[1999], page 156). But in conflict with this requirement, we know that service times may change with the experience our agents. For these reasons, we turn to randomized discrete-event simulations enabled by ExtendSim to generate results for different

routing rules.

2. A Nonlinear Program to Find Optimal Work Assignments

2.1 Design of the Objective Function

Suppose our operations manager is responsible for a total of I agents who handle K different call types. The handle time performance changes by agent and by type, so let subscripts i and k be the indices specifying each agent and type, respectively. More customers typically call in than a single agent can handle, so she seeks to divide N_{total} calls into work assignments N_{ik} for each {agent, type} pair:

$$N_{total} = \sum_{k=1}^K N_k = \sum_{i=1}^I \sum_{k=1}^K N_{ik}. \quad (1)$$

Her first step is to define an objective function for the optimization. She notices that her agents perform better with experience, and thus wishes to consider learning-based improvements in efficiency as part of her work allocation model. We consider two fundamental functions of expertise that she can use.

- The first is the the expected value of agent expertise seen by a customer, which we call the *customer's utility function of expertise*, or U_c . This function is optimized by an extreme routing rule that concentrates the highest possible volume of work on the smallest feasible subset of agents.
- The second is the sum of expertise present at the call center, which we call the *supervisor's utility function of expertise*, or U_s . This function is optimized by a routing rule that divides work assignments evenly among agents.

In the approach taken here, one of these utility functions becomes the optimization objective, and its influence is represented in the optimization solver output by a set of target percentages for routing calls to agents. Subsequent simulation results show us which of a selection of routing rules characterized by high to low evenness among call distribution outcomes does best to implement the objective's target values in a stochastic model of the service system.

The agents' circumstances will favor one of these utilities. If the arriving tasks favor specialized expertise—expert agents demonstrate critical efficiency and accuracy improvements compared to unseasoned agents—and if on-the-job learning is important for manifesting that expertise, then U_c is the better choice. Some agents may have strong latent task-specific potential that benefits the group if it can be tapped; specialized routing targets to steer the right calls to those agents are preferred.

On the other hand, the manager may have several reasons for wanting to develop agents' expertise equally. Agents may cover for each other more easily. Server pooling to reduce system congestion becomes easier. Agents perceive their own workload to be more fair when everyone else receives the same assignment. The firm may reclaim some bargaining power with respect to salaries by distributing necessary skills more evenly. If such reasons predominate, objective U_s is the better choice.

An additional subtlety exists in the present context regarding the supervisor’s utility U_s , the straight sum of expertise values in a group. Agents face a variety of call types, and each type is characterized by different average performance measures. In fact, we find that each agent-type pair exhibits a unique performance trend. We take a broad approach and accommodate both multiple call types and multiple agents when pursuing objective functions for expertise development. *Given differing call types and agent abilities, the optimal value of the sum of expertise may not be produced by a completely even assignment of calls to agents.* But from our experience of how to maximize a concave sum, we can expect that the best assignment for U_s will be more even than the best assignment for U_c .

Our manager knows that during her career she may face a different balance favoring one or the other for a specific agent group and time period. Here we will consider both U_c and U_s as objective functions, and explore the ramifications of both choices.

As a first step, our manager collects data to determine an accurate estimate of her agents’ potential to improve, such as the data shown in Figure 1 that plots one agent’s change in performance on a specific task. She has confidence in her data, and can estimate well the capabilities of untried agents by matching their human resources profiles with past work histories of veterans who were once at the same stage of expertise.

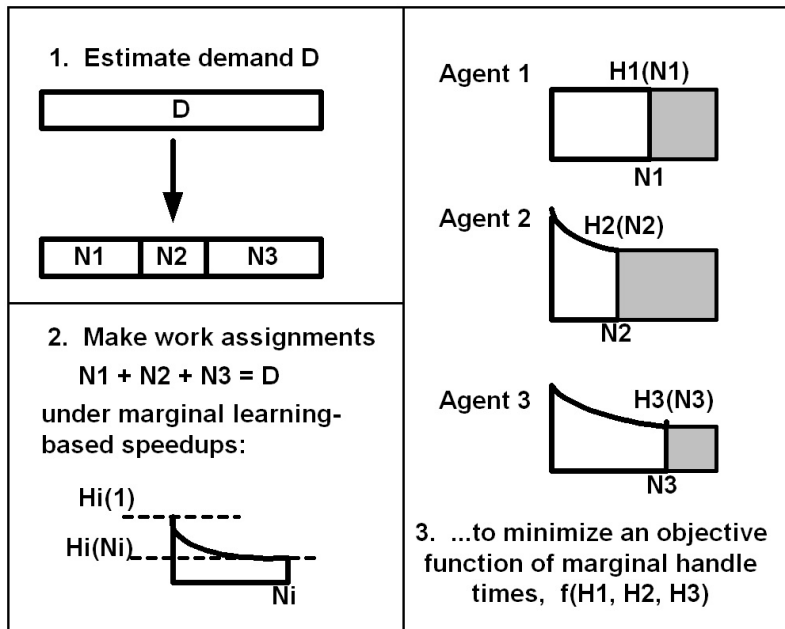


Figure 3: Given a demand forecast D , we must assign N_i jobs to agents $i \in \{1, 2, 3\}$, such that $\sum_{i=1}^3 N_i = D$. We require that our optimization objective take into account the dynamic of agent handle time decreasing through cumulative experience.

The model of expertise development appears at the right of Figure 1. This model is the commonly used log-linear learning curve applied to improvements in an agent’s handle time, where the call handle time for agent i and call type k is a decreasing function of cumulative production, $H_{ik}(N_{ik}) = H_{ik}(1) \cdot N_{ik}^{-b_{ik}}$. Figure 3 illustrates this dynamic in the presence of multiple agents. The organization also conducts research for quality assurance purposes on agent performance, and

finds that improved efficiency often correlates with improved expertise on other dimensions of service quality—by encouraging efficiency improvements, she also hopes to boost quality metrics that are harder to quantify.

Our manager considers whether or not it is useful to specify routing priorities, and works out a simple way to estimate their importance. Say $H_{ik}(N_M)$ is the handle time after an agent takes some number of calls N_M , where N_M is the maximum number of calls of type k that could be routed to i in one period. Then to attain a percentage p of this improved efficiency, the agent needs to be routed N_p calls, where

$$\begin{aligned} H_{ik}(N_p) &= H_{ik}(N_M) + p \cdot [H_{ik}(1) - H_{ik}(N_M)] \\ H_{ik}(1) \cdot N_p^{-b_{ik}} &= H_{ik}(1) \cdot N_M^{-b_{ik}} + p \cdot [H_{ik}(1) - H_{ik}(1) \cdot N_M^{-b_{ik}}] \\ N_p &= \left((1 - p) + p \cdot N_M^{-b_{ik}} \right)^{(-1/b_{ik})} \end{aligned} \quad (2)$$

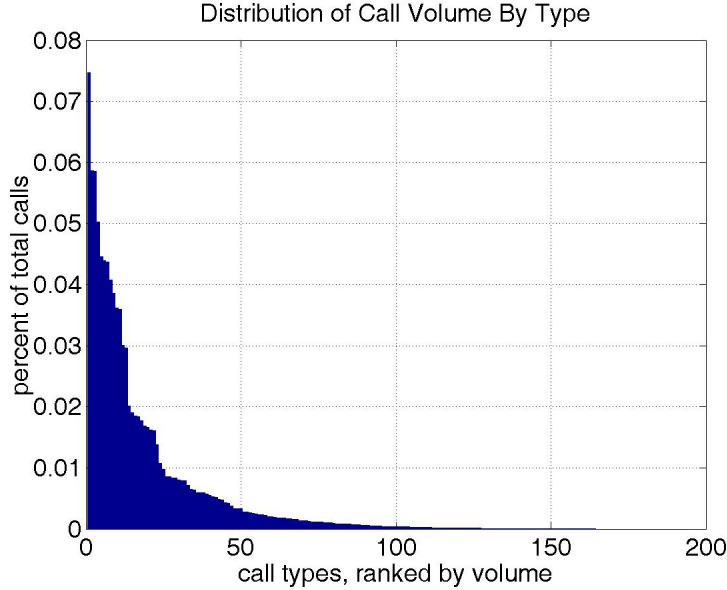


Figure 4: Different call types are characterized by different mean arrival rates. Routing rules that encourage priorities or specialization are most important when significant learning occurs among the low-volume call types.

For example, if $b_{ik} = 0.01$ and $N_M = 300$, to attain $p = 75\%$ of the potential efficiency improvement this agent must be routed a total of $N_p = 70$ calls. In general it would be easy to route N_p calls to i if the arrival rate λ_k of call type k is large, so that i receives frequent opportunities for service through a policy of agents taking the first waiting customer—then routing priorities may be unnecessary. In our example, consider the volume of other call types to be so large that they crowd out our agent’s favored type, making it unlikely a no-priority policy will cause agent i to get 70 calls within the optimization period. Were our manager to ascertain that many examples like this exist in her group—that certain of her agents will progress rapidly on specific lower-volume call types—then objective U_c is preferred, because her assignment objective must introduce routing priorities to match capable agents to favored task types that they would otherwise rarely see.

2.2 Maximizing the Customer's Utility Function U_c

Here we can also define the total customer's utility U_c over all agents and call types as the *expected value of expertise*. A quantity $0 \leq X \leq 1$ represents a positive experience level. Now, consider instead the slowest mean handle time H_k among the agent-type pairs to be zero expertise; and let positive quantities of expertise be represented by the difference between H_k and faster values H_{ik} .

- We can write the expertise level as $X_{ik} = H_k - H_{ik}$.
- In the cumulative production model $H_{ik}(N_{ik}) = H_{ik}(1) \cdot N_{ik}^{-b_{ik}}$.
- Substitution in the first equation gives $X_{ik} = H_k - H_{ik}(1) \cdot N_{ik}^{-b_{ik}}$.
- The customer's utility is defined as $U_c = \sum_j p_j X_j$ for some j , requiring routing proportions p_j . Here, proportions p_{ik} for agent i and call type k may be obtained by splitting the number of calls assigned N_{ik} into $N_{ik} = p_{ik} \cdot D_k$, where D_k is the demand forecast for type k calls.

The customer's objective, the expected value of expertise, may now be expressed as

$$\sum_i^I \sum_k^K p_{ik} X_{ik} = \sum_{i=1}^I \sum_{k=1}^K p_{ik} \cdot \left[H_k - \left(H_{ik}(1) \cdot D_k^{-b_{ik}} \cdot p_{ik}^{-b_{ik}} \right) \right] \quad (3)$$

Note that the routing proportions sum to one for every call type, or $\sum_i^I p_{ik} = 1$. Also note that expertise now has units of time (say, minutes) and its values lie in the range $0 < X_{ik} \leq H_k$.

Our manager observes that Equation (3) would better reflect the value of agents' expertise to the contact center if they were weighted by the relative arrival rates of the different call types. She therefore assigns weights to each term $w_k = \frac{\lambda_k}{I \cdot \sum_{k=1}^K \lambda_k}$, where λ_k represents the mean arrival rate of type- k customers to the center. The term I in the denominator here allows the weights to sum to 1 over all terms in the double summation, but it also reduces the numerical accuracy of a solver. Since it is present identically in all terms, it can be discarded. Note that the weights w_k could potentially incorporate other factors, such as the profitability of serving each of k different classes of customers. Weights w_{ik} could also be defined to account for unique costs or other characteristics of individual agents.

We now define the final form of the **customer's utility function** as

$$U_c(p_{ik}) = \sum_{i=1}^I \sum_{k=1}^K w_k \cdot p_{ik} \cdot \left[H_k - \left(H_{ik}(1) \cdot D_k^{-b_{ik}} \cdot p_{ik}^{-b_{ik}} \right) \right]. \quad (4)$$

When our manager decides the customer's utility—the expected value of expertise—is the right quantity to optimize, she would like to maximize Equation (4) as the objective, written

$$\max_{p_{ik}} (U_c(p_{ik})), \text{ or equivalently } \min_{p_{ik}} (-1 \cdot U_c(p_{ik})).$$

Unfortunately we see in Lemma 2.1 below that $U_c(N_{ik})$ is a convex function, and so it presents the potential for multiple local maxima. It does not have convenient optimality properties—the local maximum that is returned depends on the initial conditions the solver sees, and it may not be the true global maximum.

As a technical matter before stating the next result, we note that all routing proportions p_{ik} are constrained to be strictly greater than zero in the targets we seek from a solver, although p_{ik} is allowed to be very small, so that it is acceptable for $p_{ik} \cdot D_k \approx 0$.

Lemma 2.1. *The objective function of Equation (4), $U_c(p_{ik})$, is a **convex function** of p_{ik} .*

Proof. To show that $U_c(p_{ik})$ is convex in p_{ik} , we must show that $\vec{v}^T \nabla^2 U_c(p_{ik}) \vec{v} \geq 0$ for any real-valued vector \vec{v} of length $I * K$ (see Boyd and Vandenberghe [2004], page 74, or Nash and Sofer [1996], pages 22-23). Here $\nabla^2 U_c(p_{ik})$ is the Hessian matrix, I is the total number of agents, and K is the total number of call types. First, note that

$$\begin{aligned}
U_c(p_{ik}) &= \sum_{i=1}^I \sum_{k=1}^K w_k \cdot p_{ik} \cdot \left[H_k - \left(H_{ik}(1) \cdot D_k^{-b_{ik}} \cdot p_{ik}^{-b_{ik}} \right) \right] \\
&= \sum_i \sum_k w_k \cdot p_{ik} \cdot H_k - \sum_{i=1}^I \sum_{k=1}^K w_k \cdot H_{ik}(1) \cdot D_k^{-b_{ik}} \cdot p_{ik}^{(1-b_{ik})} \\
\text{Constant } C_1 &= \sum_{i=1}^I \sum_{k=1}^K w_k \cdot p_{ik} \cdot H_k = \sum_k w_k \cdot H_k \tag{5}
\end{aligned}$$

$$\begin{aligned}
U_c(p_{ik}) &= C_1 - \sum_{i=1}^I \sum_{k=1}^K w_k \cdot H_{ik}(1) \cdot D_k^{-b_{ik}} \cdot p_{ik}^{(1-b_{ik})} \\
&= C_1 - \sum_i \sum_k C_{ik} p_{ik}^{(1-b_{ik})} \tag{6}
\end{aligned}$$

where terms $C_{ik} = w_k \cdot H_{ik}(1) \cdot D_k^{-b_{ik}}$ are constant with respect to p_{ik} , and Equation (5) holds because $\sum_i p_{ik} = 1$. $U_c(p_{ik})$ is now a constant minus a sum, and the nonzero Hessian terms are

$$\begin{aligned}
\frac{\partial U_c(p_{ik})}{\partial p_{ik}} &= -(1 - b_{ik}) \cdot C_{ik} \cdot p_{ik}^{-b_{ik}} \\
\frac{\partial^2 U_c(p_{ik})}{\partial p_{ik}^2} &= b_{ik} \cdot (1 - b_{ik}) \cdot C_{ik} \cdot p_{ik}^{-(1+b_{ik})} \geq 0 \\
\nabla^2 U_c(p_{ik}) &= \text{diag} \left(b_{ik} \cdot (1 - b_{ik}) \cdot C_{ik} \cdot p_{ik}^{-(1+b_{ik})} \right). \tag{7}
\end{aligned}$$

The Hessian matrix $\nabla^2 U_c(p_{ik})$ is thus a diagonal matrix of size $(I * K, I * K)$, with strictly positive elements on the diagonal (because $p_{ik} > 0$). Therefore all of its eigenvalues are positive; $\nabla^2 U_c(p_{ik})$ is positive definite (see for example Lay [1994], pages 289 and 416); and the quadratic form $\vec{v}^T \nabla^2 U_c(p_{ik}) \vec{v} > 0$. It is therefore also positive semidefinite. Thus we have shown that $U_c(p_{ik})$ is a convex function of p_{ik} , and the lemma is proved. \square

Finding the best solution to objective $U_c(p_{ik})$ involves submitting the problem ($\max_{p_{ik}} U_c(p_{ik})$) to a nonlinear solver; but as discussed above, the global optimum may not be returned, or the solver may have trouble converging to a solution. Our operations manager devises the following work-around. The first-order Taylor series approximation to the nonlinear term $p_{ik}^{-b_{ik}}$ is just the linear term p_{ik} . Her data reveal the values of the learning exponents to be $b_{ik} \leq 0.1$ among the agent-type combinations, so the maximum linearization error for one p_{ik} term in the sum is about 4% at $p_{ik} = 0.349$ —and the error is much less for other p_{ik} values. She accepts the trade-off of losing a small amount of accuracy to guarantee the solver’s convergence to a good solution, and therefore the new **linearized objective function for the customer’s utility** becomes

$$U_c(p_{ik}) = \sum_k^K w_k \cdot H_k - \sum_{i=1}^I \sum_{k=1}^K \left(w_k \cdot H_{ik}(1) \cdot D_k^{-b_{ik}} \right) \cdot p_{ik}. \quad (8)$$

2.3 Maximizing the Supervisor’s Utility Function U_s

To maximize the sum of expertise U_s in the current context, where a high level of expertise is expressed as fast service, we need to minimize the sum of final marginal handle times. Thus our manager starts with an objective function to minimize the sum of $H_{ik}(N_{ik})$ values over all agent-type pairs, and again includes the weighting by call type arrival rate $w_k = \frac{\lambda_k}{I \cdot \sum_{k=1}^K \lambda_k}$. This gives

$$U_s(p_{ik}) = \sum_{i=1}^I \sum_{k=1}^K \left(w_k \cdot H_{ik}(1) \cdot D_k^{-b_{ik}} \right) \cdot p_{ik}^{-b_{ik}}. \quad (9)$$

We seek the values p_{ik}^* that minimize this function, or $\min_{p_{ik}} (U_s(p_{ik}))$. An optimal routing target p_{ik}^* may show some unevenness, and increase assignments according to high-volume call types given by w_k values, and according to high-performing agents as determined by low $H_{ik}(1)$ and high b_{ik} values. Nevertheless as we show in Lemma 2.2 below this is a convex function we wish to minimize; in a solver that may be transformed to the problem of maximizing a concave function, which is naturally maximized by making **even assignments** when agents were equally capable. Our manager therefore expects the routing targets derived from $U_s(p_{ik})$ and the associated best routing rule to promote more even assignments than would be the case for the U_c objective, which explicitly seeks the most extreme specialized assignments.

Lemma 2.2. *The objective function of Equation (9), $U_s(p_{ik})$, is a **convex function** of p_{ik} .*

Proof. Proceeding as in Lemma 2.1, we have

$$\begin{aligned}
U_s(p_{ik}) &= \sum_{i=1}^I \sum_{k=1}^K \left(w_k \cdot H_{ik}(1) \cdot D_k^{-b_{ik}} \right) \cdot p_{ik}^{-b_{ik}} \\
&= \sum_{i=1}^I \sum_{k=1}^K C_{ik} \cdot p_{ik}^{-b_{ik}} \\
\frac{\partial U_s(p_{ik})}{\partial p_{ik}} &= -b_{ik} C_{ik} p_{ik}^{-(1+b_{ik})} \\
\frac{\partial^2 U_s(p_{ik})}{\partial p_{ik}^2} &= b_{ik}(1+b_{ik}) C_{ik} p_{ik}^{-(2+b_{ik})} \geq 0 \\
\nabla^2 U_s(p_{ik}) &= \text{diag} \left(b_{ik}(1+b_{ik}) C_{ik} p_{ik}^{-(2+b_{ik})} \right). \tag{10}
\end{aligned}$$

The Hessian matrix $\nabla^2 U_s(p_{ik})$ is thus a diagonal matrix of size $(I * K, I * K)$, with positive elements on the diagonal. Therefore as we saw for the Hessian matrix in Lemma 2.1, page 8, $\vec{v}^T \nabla^2 U_s(p_{ik}) \vec{v} \geq 0$, and the lemma is proved. \square

2.4 Design of the Constraints

Our operations manager notes that three constraints are implied by Figure 3. First, the sum of the calls must equal the demand forecast D_k , or $\forall k, \sum_{i=1}^I N_{ik} = D_k$. Second, every agent must get some calls—Agent 2 appears the least efficient, but still receives a nonzero allocation. This is because substantive changes in expertise accrue slowly and in parallel to the daily peaks and valleys of incoming traffic, and even the less promising agents will be needed during the peak times to keep customer waiting times reasonable. Thus some reserve capacity for peak times must be built into the solver’s work assignment—and informed by observations from empirical data, we take that to mean that every agent must be taking calls during at least 20% of the time during the optimization period. This capacity is captured in a set of *minimum utilization constraints*, which we will also refer to simply as *minimum constraints*. This discussion also informs the question of what time span the work assignments should cover. Since meaningful expertise gains occur after completing tens or hundreds of similar tasks, while agents only get a few chances each day to engage in the same task, the time horizon for the manager’s expertise-aware routing assignments will be some number of months.

Third, agents must not be overloaded. This is represented by the gray areas next to agent allocations that show unused capacity. Both behavioral and operational issues may be at stake: large differences in volumes of work assigned will strike some agents as unfair, and generating extremely tight work schedules through routing targets leaves agents with no flexibility to cope with day-to-day fluctuations around the average demand forecasts. These ideas are captured in *maximum utilization constraints*, or just *maximum constraints*. Empirical data suggest a maximum utilization constraint where an agent is between 50% and 60% utilized during the optimization period.

To express the minimum and maximum constraints, the manager evaluates the total time taken by agent i to handle all calls in his assignment of type k , for every agent-type pair. This is accomplished by the following steps.

- If an agent's handle time does not change, he takes $T_{ik} = \sum_{n=1}^N H_{ik}(n)$ minutes to handle N jobs, where $H_{ik}(n)$ is the time needed for the n th call; if $H_{ik}(n)$ is constant, $H_{ik}(n) = H_{ik}(1)$, and $T_{ik} = N \cdot H_{ik}$.
- Assume a unique log-linear learning curve function exists for every agent-type pair. $H_{ik}(n)$ decreases monotonically as we have seen compared with the first handle time $H_{ik}(1)$ as cumulative production n increases.
- As before, the cumulative production total N_{ik} and the routing target proportion p_{ik} are related by $N_{ik} = p_{ik}D_k$, where D_k is the demand forecast for type k calls. Here it is convenient to work with N_{ik} .
- Specific limits can be placed on the value of the learning exponent b_{ik} : $0 \leq b_{ik} \leq 0.1$, and therefore $0.9 \leq (1 - b_{ik}) \leq 1$.
- In the presence of a learning curve, the time taken to handle N jobs becomes $T_{ik} = \sum_{n=1}^N H_{ik}(n) \approx \int_{v=0}^N H_{ik}(v)dv$.

A good closed-form approximate expression for T_{ik} is:

$$T_{ik} \approx H_{ik}(1) \int_{v=0}^N v^{-b_{ik}} dv = \frac{H_{ik}(1) \cdot N^{(1-b_{ik})}}{(1-b_{ik})}. \quad (11)$$

Now the set of maximum utilization constraints may be written as

$$\forall i \sum_{k=1}^K T_{ik} \leq T_{max}. \quad (12)$$

and the set of minimum utilization constraints may be written as

$$\forall i \sum_{k=1}^K T_{ik} \geq T_{min} \quad (13)$$

The manager sets the values of T_{min} and T_{max} by looking at empirical trends of agent utilization, as those reveal the organization's consensus about the sustained rates at which agents should be busy over multiple-month time periods. Again, the minimum utilization might be 20%, and the maximum 60%. Multiplying the standard number of minutes worked per day by the number of working days in the optimization period gives the total time T_{work} corresponding to 100% utilization. Let ρ_{max} be the busiest allowed mean utilization rate over the period; then $T_{max} = \rho_{max} \cdot T_{work}$. Similarly, ρ_{min} is the lowest allowed mean utilization rate over the period, and $T_{min} = \rho_{min} \cdot T_{work}$.

Figure 5 compares the forms of the convex marginal handle time function $H_{ik}(N_{ik})$ and the concave total time $T_{ik}(N_{ik})$. Note that $T_{ik}(N_{ik})$ appears in two constraint functions with opposite

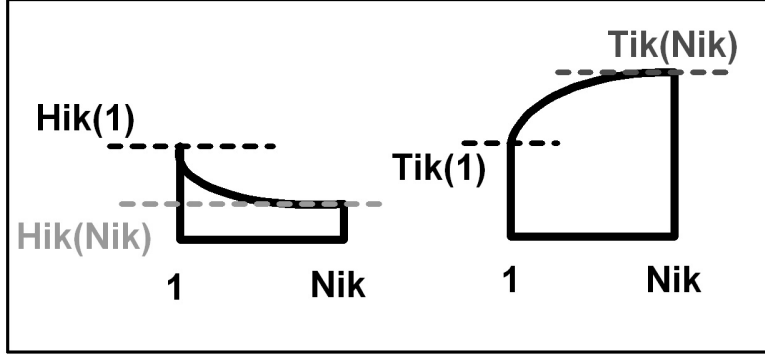


Figure 5: Left: plot of a convex log-linear learning curve $H_{ik}(N_{ik}) = H_{ik}(1)N_{ik}^{-b_{ik}}$. The curve shows marginal values of handle time per call decreasing with cumulative production N_{ik} , for agent i and call type k . Right: the sum of all the handle times from the left-hand curve becomes a concave function, the total time $T_{ik}(N_{ik})$ —note the scale of the peak is attenuated to fit on the plot.

relational symbols—greater than or equal to T_{min} , and less than or equal to T_{max} —but when solving an optimization problem, constraints must be able to be expressed in a standard form where all inequalities point in the same direction. We take that direction to be less-than-or-equal-to, which means the minimum constraint must be inverted. When that occurs, the sum of concave functions in the minimum constraint is also inverted, resulting in the convex function $\frac{T_{min}}{\sum_{k=1}^K T_{ik}}$.

2.5 A Nonlinear Program to Find Good Call Routing Targets

Below we present two convex programs with linearized maximum constraints, known as Program CP- U_c and Program CP- U_s . These two programs are used to generate the optimal call routing targets for Section 3. The constraint functions are the same for both objective functions; $U_c(p_{ik})$ maximizes the linearized customer’s utility function and $U_s(p_{ik})$ maximizes the supervisor’s utility function over all agent-type pairs by varying call routing proportions p_{ik} . Here all functions are written in terms of the optimization variables p_{ik} .

$$\boxed{\text{CP-}U_c} \quad \max_{p_{ik}} U_c(p_{ik}) = \sum_k^K w_k \cdot H_k - \sum_{i=1}^I \sum_{k=1}^K \left(w_k \cdot H_{ik}(1) \cdot D_k^{-b_{ik}} \cdot p_{ik} \right) \quad (14)$$

such that

$$\forall i, k: N_{ik} = p_{ik} \cdot D_k \quad \text{“number of calls”}$$

$$\forall i, k: T_{ik}(p_{ik}) = \frac{H_{ik}(1) \cdot (p_{ik} \cdot D_k)^{(1-b)}}{(1-b)} \quad \text{“time taken”}$$

$$\forall k: w_k = \frac{\lambda_k}{\sum_{k=1}^K \lambda_k} \quad \text{“arrival rate weighting”}$$

$$\forall i \sum_{k=1}^K \beta_{ik} H_{ik}(1) (p_{ik} \cdot D_k) \leq T_{max} \quad \text{“max. utilization”} \quad (15)$$

$$\forall i \sum_{k=1}^K T_{ik}(p_{ik}) \geq T_{min} \quad \text{“min. utilization.”} \quad (16)$$

$$\forall k \sum_{i=1}^I p_{ik} \cdot D_k = D_k \quad \text{“full service”} \quad (17)$$

$$\forall i, k \quad p_{ik} > 0 \quad \text{“}p_{ik} \text{ positive”} \quad (18)$$

$$\boxed{\text{CP-}U_s} \quad \min_{p_{ik}} U_s(p_{ik}) = \sum_{i=1}^I \sum_{k=1}^K w_k H_{ik}(1) (p_{ik} \cdot D_k)^{-b_{ik}} \quad (19)$$

such that...(as above)

We adopt the form of a convex program for each objective in order to guarantee the existence of a global minimum objective value. A convex program is defined as an optimization program that minimizes a convex function over a convex set, where the convex set is formed from a set of non-linear convex inequality constraints and linear equality constraints. See Boyd and Vandenberghe [2004], page 136; Bertsekas [2003], page 208; or Nash and Sofer [1996], page 473 for proofs and examples. It is known that convex programs yield unique global solution values, and commonly used interior point solvers such as Matlab’s *fmincon()* function can find these global solutions in a reasonable time by means of a variant of Newton’s method (see Coleman and Zhang [2009], pages 3–18).

3. Simulation of Routing Policies to Implement Optimal Work Assignment Targets

3.1 Approach to Simulation

Our nonlinear program outputs precise targets for the agents’ workloads. Yet it is unclear how such precise targets may be implemented in the unpredictable daily call volume seen by a typical contact center. To gain insight into methods for implementing our ideal targets, we now describe results generated by a contact center simulator. This simulator generates streams of random numbers that imitate the random time points at which calls arrive, and the random lengths of time required to serve each customer.

The goal here is to develop and analyze routing rules that both show fidelity to our optimization targets, and show high performance on day-to-day operational metrics. Call centers depend on metrics that are not easily amenable to analytical methods and optimization, such as mean customer waiting time, the average service level, the average number of abandonments or dropped calls per period, and so on. All time-based metrics correlate fairly well with waiting time — more waiting time leads to predictably higher abandonment rates, and lower service levels — so we focus primarily on the average waiting time \bar{W} as our key queueing simulation output.

In a single experiment, an optimization solver implementing Program CP- U_c or CP- U_s runs first. It defines the size and scope of the problem, and all relevant parameters. As described in Section 2.5, it outputs a list of target proportions for each agent-call type pair. It also generates two critical priority routing tables: the *agent-to-type map*, or just *agent map*, and the *type-to-agent map*, or just *type map*. The simulator contains a routing subprogram that uses these tables to guide new customers to free agents during the run. When output from Program CP- U_c or CP- U_s , both routing tables follow exactly the ordering of target proportions. For instance, if an agent is assigned tasks A and B in proportions 50.1% and 49.9%, task A comes first in his row of the agent map.

Figure 6 shows elements of the ExtendSim program used to generate results. The entire solution is a hybrid of the Matlab optimization toolkit, spreadsheets, ExtendSim, custom Java code, and postprocessing scripts. In ExtendSim, routing rules were implemented using hierarchies of customized equation blocks.

3.2 Priority Routing Rules

The simulator uses the proportions and tables differently, depending on a key parameter: the *routing rule* or *policy*. Figure 7 illustrates the differences among the routing rules we will be exploring. In the top figure, the system does not prioritize calls by type, but only considers waiting time: a “no priority” policy. When available to take a call, an agent chooses the customer with the longest waiting time according to first-in-first-out order. The figure assumes the agents’ mean service times are the same, so in the long run they each end up serving half of the incoming arrivals.

Without priorities, on-the-job learning specific to each call type still takes place, but it is not tracked. As we have seen before, this leads to expertise distributions among agents driven solely by the proportion of call types within the arrival stream, and each agent’s potential to learn by

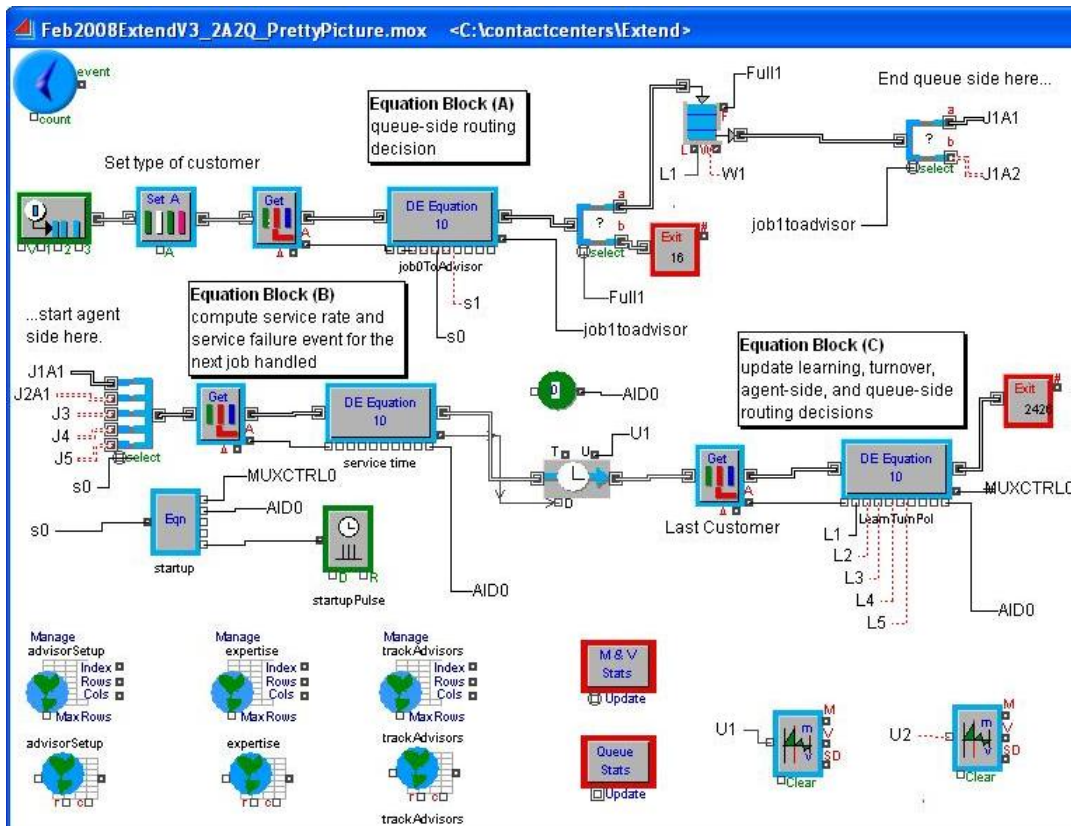


Figure 6: Simplified diagram of the ExtendSim simulator implementation used to test routing rules.

serving those types. Agents serve many types, and spend a high proportion of their time in the steep starting phase of each learning curve. This policy both minimizes waiting time from server pooling, and maximizes the sum of expertise over all agents and types in the system.

Yet management may determine that more specialized expertise is needed than can be obtained under a no-priority rule. The bottom of Figure 7 illustrates policies that develop specialized expertise. Maximum specialization occurs when agents serve a single type, and sit idle when their preferred customer type is absent. However, this extreme rule is rarely feasible except in the largest systems. Instead, we may use rules that trade-off some waiting time performance to obtain a higher degree of specialized expertise. Such rules employ agent-to-type priority routing schedules that limit idling.

Consider a priority routing rule for Figure 7. Call types A and B arrive at the system requiring separate kinds of expertise—in a the financial services setting, A may represent questions about tax forms, and B may represent questions about credit card payments. Agents 1 and 2 have their own preferred call type, but may also assist by taking calls of the other type. Let the service rates μ_1 and μ_2 be constant at one customer per arbitrary time period. Then the arrival rates λ_A and λ_B and the routing rule determine the degree of task sharing, and thus the trade-off of waiting time for specialized expertise.

- **Routing Rule Example** Assume the lower diagram of Figure 7 is a system with Markovian

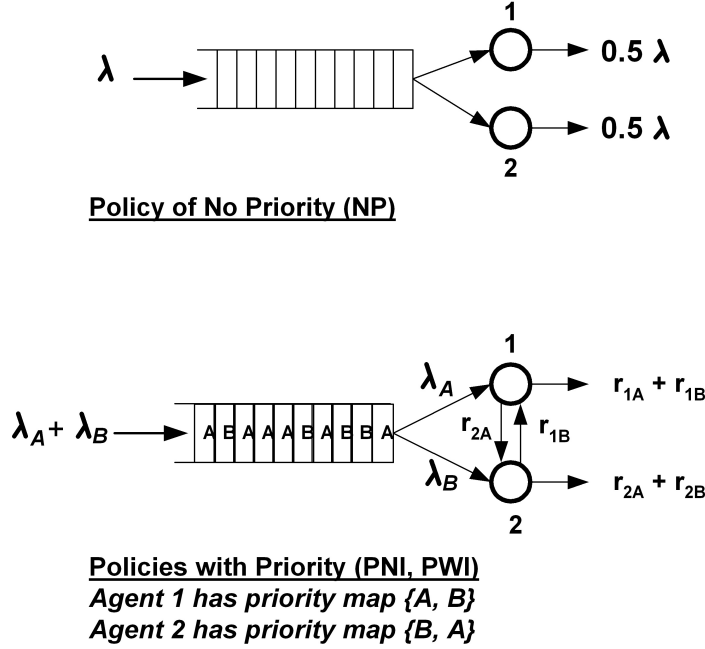


Figure 7: Three rules for routing calls to agents, and an example of policy-dependent priority maps. Top: When no call type priorities are identified, policy NP, agents take any waiting call. Bottom: When priorities are present, agents choose the first waiting call of their preferred type. If no preferred customers are present, the agent may offer to serve other types of customers (priority-no-idle policies), or he may idle and wait for his preferred type (priority-with-idle policies). Here λ_A and λ_B are Poisson arrival rates for call types A and B; r_{ik} are rates of type k served by agent i .

arrival and service processes. Mean arrival rates are $\lambda_A = 0.45$ customers per arbitrary time period, $\lambda_B = 0.25$, and service rates are $\mu_1 = \mu_2 = 1$. The priority schedule, or priority map, for agent 1 is $\{A, B\}$. For agent 2 the map is $\{B, A\}$. We adopt the routing rule *priority-no-idle* (PNI): serve the highest priority customer from the priority map, if a customer of that type is present in the system; otherwise check if the other agent is busy and a customer of the other type is waiting—if so, then serve the other type.

These parameters suffice to determine the steady-state flows in the system: $r_{1A} = 0.365$, $r_{1B} = 0.029$, $r_{2A} = 0.085$, $r_{2B} = 0.221$. Figure 8 shows a simulation procedure's convergence to the final answer. Agent 2 is free more often, and the task sharing rule allows him to give substantial assistance to Agent 1. Occasionally, Agent 1 assists Agent 2 with type B calls as well — about 7% of Agent 1's service is used to help Agent 2. \square

For our system simulations, let us consider the following five operational routing rules. Compared to the first default rule, the other four pay a higher cost in waiting time in return for a better value of the customer's objective function, $U_c(p_{ik})$.

1. No priority, or NP. Under this rule, on-the-job learning plays no role in workload assignments. We calculate a staffing level sufficient to maintain a desired average service level with respect to a call volume forecast, and do not consider skills or the call type when routing calls to agents.

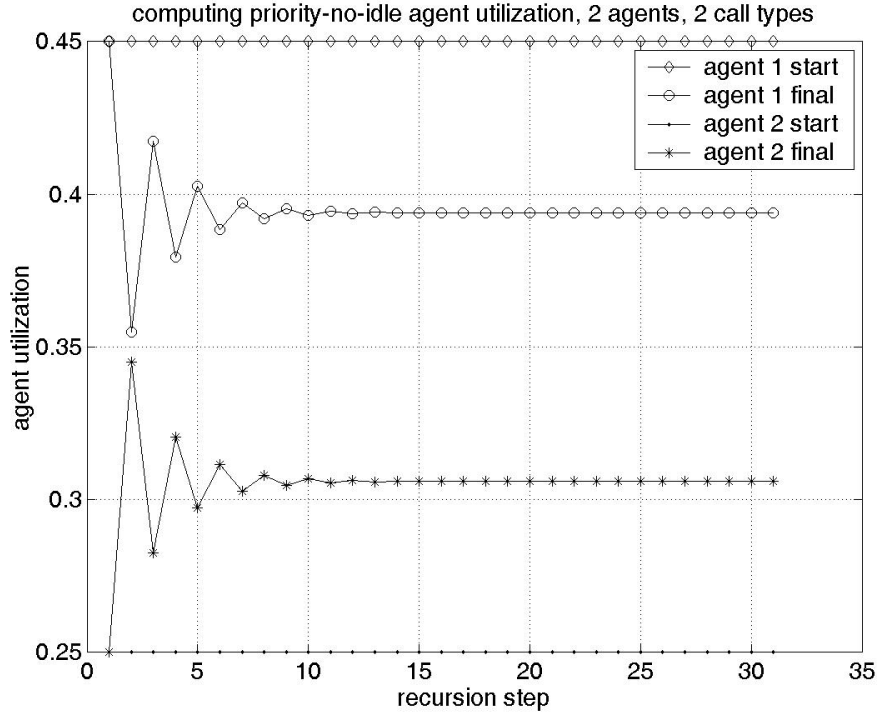


Figure 8: Utilization for two agents under Idle and No-Idle routing—example 3.2 results. The two results in the middle are achieved under policy PNI, while the extreme top and bottom results are obtained under PWI.

2. Priority-No-Idle-Constant, or PNI-Constant. When free, an agent will take his next call from the queue that is highest on his priority list, but all call types appear on his list. So he favors the call types in a certain order, but will not be idle if a customer is ready to be served.

3. Priority-No-Idle-Swap, or PNI-Swap. This is the same as PNI-Constant, with one difference — the priorities of call types in the routing tables are changed dynamically, such that the agent-type pair with the greatest error between the routing target proportion and the currently calculated proportion occupies the top agent map and type map positions. The pair with the second greatest error occupies the second two positions, and so on. The recalculation takes place following every service event.

4. Priority-With-Idle-N-Constant, or PWI-N-Constant. An agent will idle until a call type from his assigned priority list arrives, even at the cost of making other customers wait. The implementation drops all non-priority call types from the rows of the agent map and the type map. If all types were present, this would be the same as PNI-Fixed. The value N indicates the lengths of the rows on the agent and type maps.

5. Priority-With-Idle-N-Swap, or PWI-N-Swap. This is the same as PWI-Constant, except that the priorities among the agent-type pairs still present in the reduced agent and type maps are changed dynamically, as with PNI-Swap.

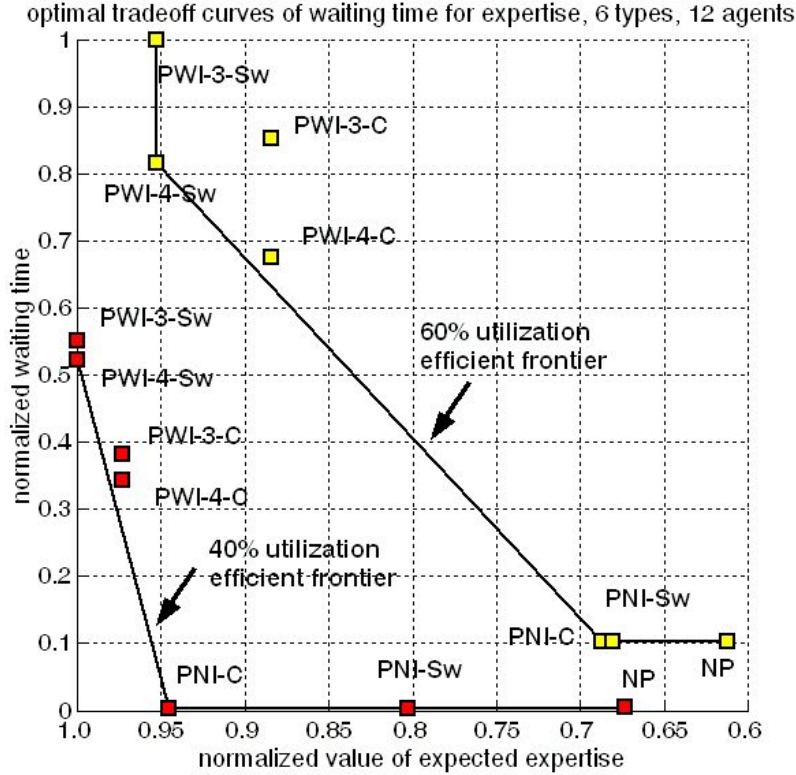


Figure 9: Optimal trade-off curves (efficient frontiers) defined by routing rules. These are for a system of 6 call types and 12 agents. Each point marks the performance of one of the types of routing rules. The rules shown define the efficient frontier for average system utilization rates of 40% and 60%. Normalizations are done separately with respect to the values generated by each utilization series.

3.3 The Efficient Frontier Defined By Five Routing Rules

Figure 9 plots the performance of our five rules in a such a way that they define an *efficient frontier* describing the trade-off between the normalized value of the customer's objective function $U_c(p_{ik})$, as defined on page 13, and the normalized mean waiting time per customer \bar{W} . An ideal policy, unachievable in practice, would land on the origin: zero waiting time, and maximum customer utility.

Each point represents the tradeoff at the final value of a year's operation (240 business days) of a twelve-agent, six-call-type system. All agents begin the year with the same mean service time of seven minutes for every call type; their learning exponents b_{ik} are all set to a moderate value of 0.01, and the system assumes Markovian arrivals and service. The arrival rates for each call type are equal, and are determined by the mean utilization value for the system: 40%, or 60%. Again, for this system the arrival rates among call types are equal, and learning rates among agents are equal.

Under these conditions, we see how the introduction of priorities increases both measures. Rule NP lies at one extreme, providing the lowest waiting time. The forced idling rule PWI-3-Swap lies on the other extreme, providing the largest gain in objective $U_c(p_{ik})$. Both priority-no-idle

rules offer good compromises for the 40% utilization case. When utilization increases to 60%, interruptions in agent work patterns caused by prioritizing call types result in longer queues. They also result in more service opportunities for ascending specialized learning curves; PWI-3-Swap at 60% achieves a higher $U_c(p_{ik})$ gain than PWI-3-Swap at 40%.

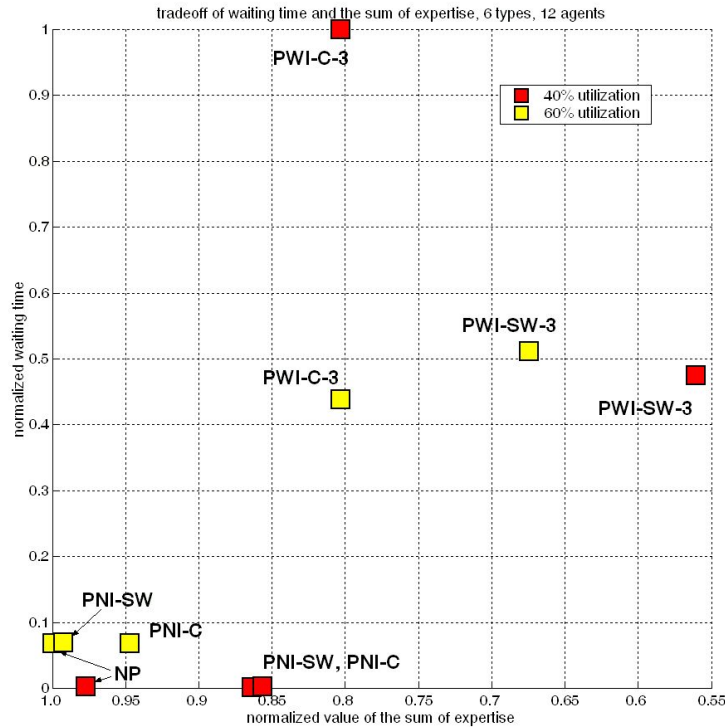


Figure 10: **Routing Rules and the the Supervisor’s Utility.** Optimal tradeoff curves for 6 call types and 12 agents, with the sum of expertise (the supervisor’s utility U_s) on the horizontal axis, and waiting time on the vertical axis. Each point marks the performance of one of the routing rules. Note that the optimal trade-off curves consist of a single point for both the 40% and 60% utilization cases. That point corresponds to routing rule NP, or no-priority.

By contrast, Figure 10 shows the equivalent efficient frontier curve for the same system using the supervisor’s utility as the objective function, as defined by $U_s(p_{ik})$ on page 13. Note that the curve consists of one point for both 40% and 60% utilization values: rule NP achieves both the lowest total waiting time and the best value of the objective function. This supports the intuition that the supervisor’s utility is optimized by even routing, as rule NP routes calls evenly among agents. It also naturally weights routing proportions according to the arrival rate ratio $\frac{\lambda_k}{\sum_{k=1}^K \lambda_k}$, which we also included in the objective $U_s(p_{ik})$. In contrast to rule NP, the priority routing rules

do not appear to be attractive options when our manager chooses the supervisor's utility as her objective.

3.4 Time Series of Expertise and Waiting Time

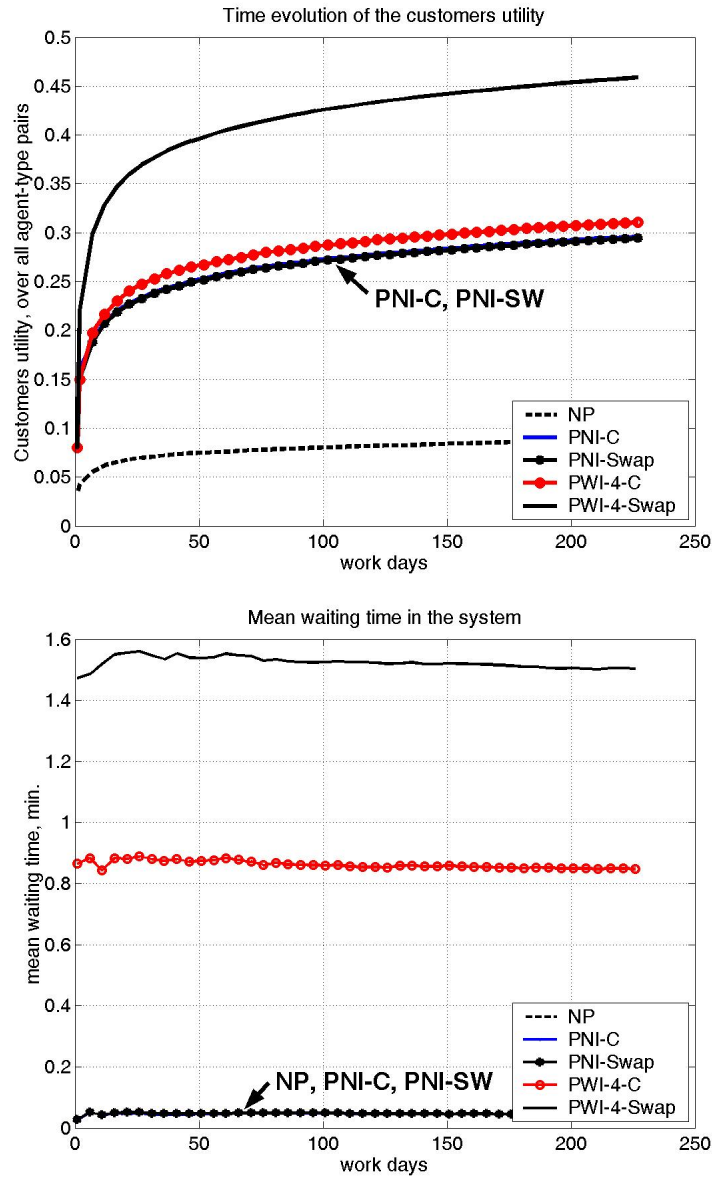


Figure 11: Examples of simulation time series. Top: evolution of the customer's objective $U_c(p_{ik})$ in the system. Priority-With-Idle-Swap performs best. Bottom: cumulative average waiting time in the system.

As an illustration of how the customer's objective $U_c(p_{ik})$ develops as a simulation experiment unfolds, Figure 11 provides two time series plots from one replication for quantities of interest in this case. The top of Figure 11 shows the evolution of $U_c(p_{ik})$ in the system, and the bottom shows

the mean waiting time for the five policy types. Routing rule PWI-4-Swap appears in the highest position both times—it provides the largest trade-off of waiting time for expected expertise; and policy NP gives the smallest trade-off, as we expect.

4. Conclusion

In this paper we analyzed policies to develop agent expertise in stochastic models of call center operations. We took a two-step approach of setting development targets with an optimization program, and exploring routing rules to implement those targets in a discrete event system model with randomized arrival and service times. Our approach allows contact center management to assess the impact of targeted call routing, which increases average system waiting time, in order to develop specialized expertise from on-the-job-learning over a period of weeks or months. Results suggest that a class of policies we call Priority-No-Idle provide most of the benefits of learning in return for a relatively small increase in customer waiting.

We foresee many interesting extensions to this work. First, we only present results for a few scenarios, out of the many configurations possible for modern contact centers. Different parameterizations of call types, starting agent skills, and learning rates based on empirical data from different industries could be explored using our experimental framework. Second, the impact of learning was measured as an improvement in handle time; but we could also consider improvements in service quality that, for example, would reduce the number of callbacks—and a discrete event modeling study would be the ideal means to explore this topic. Finally, more sophisticated routing rules could be devised to balance long-term learning objectives with short-term operational goals; how would such rules fare against the simpler rules we tested here?

Acknowledgments

We offer our sincere thanks to the ExtendSim Team for their generous support during the course of this research.

References

- S. Aguir, F. Karaesmen, O.Z. Aksin, and F. Chauvet. "The impact of retrials on call center performance". *OR Spektrum*, 26:353–376, 2004.
- Z. Aksin, M. Armony, and V. Mehrotra. "The modern call center: a multi-disciplinary perspective on operations management research.". *Production and Operations Management*, 16(6):665 – 688, December 2007.
- Z. Aksin, F. de Vericourt, and F. Karaesmen. "Call center outsourcing contract analysis and choice.". *Management Science*, 54(2):354–368, February 2008.
- O. Alban. "Enhancing workforce efficiencies: taking an enterprise view of quality assurance". Witness Systems Corporation web site, <http://www.witness.com>, November 2004.

- J. Artajelo and G. Falin. "Standard and retrial queueing systems: a comparative analysis". *Revista Matematica Complutense*, XV(1):101–129, 2002.
- J.R. Artalejo. "A queueing system with returning customers and waiting in line". *Operations Research Letters*, 17:191–199, 1995.
- A. Avramidis and P. L'Ecuyer. "Modeling and simulation of call centers". In *Proceedings of the 2005 Winter Simulation Conference*, pages 144 – 152, Orlando, FL, December 2005. IEEE Press.
- F. Baccelli and P. Bremaud. *Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences*, volume 26 of *Applications of Mathematics*. Springer-Verlag, Berlin, second edition, 2003.
- A. Badiru. "Computational survey of univariate and multivariate learning curve models". *IEEE Transactions on Engineering Management*, 39(2):176 – 188, May 1992.
- J. Baras, A. Dorsey, and A. Makowski. "Two competing queues with linear costs and geometric service requirements: the μ - c rule is often optimal". *Advances in Applied Probability*, 17: 186–209, March 1985.
- T. Basar and G.J. Olsder. *Dynamic Noncooperative Game Theory*. Academic Press Inc., New York, 1982. ISBN 0 12 080220 1.
- D.P. Bertsekas. *Convex Analysis and Optimization*. Athena Scientific, PO Box 391, Belmont, MA 02178, 2003.
- D.P. Bertsekas. *Dynamic Programming and Stochastic Control*. Mathematics in Science and Engineering. Academic Press, Inc., 111 Fifth Avenue, New York, NY 10003, first edition, 1976.
- D.P. Bertsekas. *Dynamic Programming: Deterministic and Stochastic Models*. Simon and Schuster, Englewood Cliffs, New Jersey, 07632, first edition, 1978.
- D.P. Bertsekas. *Dynamic Programming and Optimal Control*, volume I. Athena Scientific, PO Box 391, Belmont, MA 02178, 1995a.
- D.P. Bertsekas. *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, PO Box 391, Belmont, MA 02178, second edition, 1995b.
- P.P. Bocharov, C.D'Apice, A.V. Pechinkin, and S. Salerno. *Queueing Theory*. Modern Probability and Statistics. VSP, Brill Academic, The Netherlands, first edition, 2004.
- J. Bodreau, W. Hopp, J. McClain, and L. Thomas. "On the interface between operations and human resource management". *Manufacturing and Service Operations Management*, 5(3):179–202, Summer 2003.
- S.K. Bordoloi. "Agent recruitment planning in knowledge-intensive call centers". *Journal of Service Research*, 6(4):309 – 323, 2004.

- S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004. ISBN 0 521 83378 7.
- L. Breuer and D. Baum. *An Introduction to Queueing Theory and Matrix-Analytic Methods*. Springer, P.O. Box 17, 3300 AA Dordrecht, The Netherlands, first edition, 2005.
- L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. "Statistical analysis of a telephone call center: a queueing science perspective". Technical Report 03-12, Wharton Financial Institutions Center, November 9, 2002.
- F. Caro and J. Gallien. "Dynamic assortment with demand learning for seasonal consumer goods". *Management science*, 53(2):276–292, 2007.
- H. Chen and D. Yao. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*, volume 46 of *Applications of Mathematics*. Springer-Verlag, New York, first edition, 2001.
- B.D. Choi, K.B. Choi, and Y.W. Lee. "M/G/1 retrial queueing systems with two types of calls and finite capacity". *Queueing Systems*, 19:215–229, 1995.
- B. Cleveland and J. Mayben. *Call Center Management on Fast Forward: Succeeding in Today's Dynamic Inbound Environment*. Call Center Press, Annapolis, Maryland, first edition, 2000.
- T. Coleman and Y. Zhang. *Matlab Optimization Toolbox*. The MathWorks, Inc., 3 Apple Hill Drive, Natick, MA 01760 USA, 2009.
- F. de Vericourt and Y. Zhou. "Managing response time in a call-routing problem with service failure". *Operations Research*, 53(6):968 – 981, 2005.
- E.V. Denardo. *Dynamic Programming: Models and Applications*. Prentice-Hall, Inc., Englewood Cliffs, NJ 07632, first edition, 1982.
- C. Derman. *Finite State Markov Decision Processes*, volume 67 of *Mathematics in Science and Engineering*. Academic Press, Inc., 111 Fifth Avenue, New York, NY 10003, first edition, 1970.
- B. Dietrich and T. Harrison. "Serving the services: the emerging science of service management opens opportunities for operations research and management science". *OR/MS Today*, at <http://www.lionhrtpub.com/orms/orms-6-06/frservice.html>, June 2006.
- J. Dopkeen. "Academia Dissects the Service Sector, but Is It a Science?". *New York Times*, April 18 2006.
- R. Durrett. *Essentials of Stochastic Processes*. Springer Texts in Statistics. Springer-Verlag, New York, 1999.
- G. Eitzen, D. Pantou, and G. Mills. "Multi-skilled workforce optimization.". *Annals of Operations Research*, 127(1):359 – 372, 2004.
- S. Elaydi. *An Introduction to Difference Equations*. Springer, 233 Spring Street, New York, NY 10013, USA, third edition, 2005.

- G.I. Falin and J.G.C. Templeton. *Retrial Queues*. Monographs on Statistics and Applied Probability. Chapman and Hall, 2-6 Boundary Row, London, SE1 8HN, UK, first edition, 1997.
- W. Fischer and K. Meier-Hellstern. "The Markov-modulated Poisson process (MMPP) cookbook". *Performance Evaluation*, 18(2):149–171, September 1993.
- C. Froehle. "Service personnel, technology, and their interaction in influencing customer satisfaction". *Decision Sciences*, 37(1):5 – 38, 2006.
- N. Gans and Y. Zhou. "Managing learning and turnover in employee staffing". *Operations Research*, 50(6):991–1006, November-December 2002.
- N. Gans and Y. Zhou. "A call-routing problem with service-level constraints". *Operations Research*, 51(2):255 – 271, March 2003.
- N. Gans, G. Koole, and A. Mandelbaum. "Telephone call centers: tutorial, review, and research prospects". *Manufacturing and Service Operations Management*, 5:79–141, Spring 2003.
- J. George and J. Harrison. "Dynamic control of a queue with adjustable service rate". *Operations Research*, 49(5):720–731, November-December 2002.
- P. Ghemawat. "Building strategy on the experience curve". *Harvard Business Review*, pages 143–149, March-April 1985.
- S. Globerson and N. Levin. "Incorporating forgetting into learning curves". *International Journal of Operations and Production Management*, 7(4):80 – 94, 1987.
- W. Grassmann. "The use of eigenvalues for finding equilibrium probabilities of certain Markovian two-dimensional queueing problems". *INFORMS Journal on Computing*, 15(4):412–421, December 2003.
- D. Gross and C. Harris. *Fundamentals of Queueing Theory*. Wiley Series in Probability and Statistics. Wiley Interscience, New York, third edition, 1998.
- H.W. Gustafson. *Force-loss cost analysis*, chapter "Employee Turnover: Causes, Consequences, and Control". Addison-Wesley, Reading, MA, first edition, 1982. W.H. Mobley, ed.
- S. Halfin and W. Whitt. "Heavy traffic limits for queues with many exponential servers". *Operations research*, 29(3):567 – 588, May-June 1981.
- R. Hampshire, M. Harchol-Balter, and W. Massey. "Fluid and diffusion limits for transient sojourn times of processor sharing queues with time-varying rates". *Queueing Systems*, 53(1-2):19–30, 2006.
- J. Harrison. *Brownian Motion and Stochastic Flow Systems*. Probability and Mathematical Statistics. John Wiley and Sons, New York, first edition, 1985.
- J. Harrison. "A priority queue with discounted linear costs". *Operations Research*, 23(2):270–282, March-April 1975.

- S. Hasija, E. Pinker, and R. Shumsky. "Staffing and routing in a two-tier call center". *Int. J. Operational Research*, 1(1/2):8 – 29, 2005.
- S. Hasija, E. Pinker, and R. Shumsky. "Call center outsourcing contracts under information asymmetry". *Management Science*, 54(4):793 – 807, April 2008.
- C. Heitz, G. Ryder, and K. Ross. "Knowledge Management in Call Centers: How Routing Rules Influence Expertise and Service Quality". In *MSOM Conference Proceedings*, pages 1–7, University of Maryland, College Park, MD, June 5-6 2008.
- J. Higgins. *Introduction to Modern Nonparametric Statistics*. Thomson Brooks/Cole, 511 Forest Lodge Road, Pacific Grove, CA 93950, 2004. ISBN 0-534-38775-6.
- F. Hillier and G. Lieberman. *Introduction to Operations Research*. McGraw Hill Inc., New York, eighth edition, 2005.
- W. Hopp, S. Iravani, and G. Yuen. "Operations systems with discretionary task completion". *Management Science*, 53(1):61–77, January 2007.
- A. Hordijk and G. Koole. "On the optimality of LEPT and μ -c rules for parallel processors and dependent arrival processes". *Advances in Applied Probability*, 25:979–996, 1993.
- S. Howick and C. Eden. "Learning in disrupted projects: on the nature of corporate and personal learning". *International Journal of Production Research*, 45(12):2775 – 2797, 2007.
- S. Iravani, B. Kolfal, and M. Oyen. "Call-center labor cross-training: it's a small world after all". *Management Science*, 53(7):1102 – 1112, July 2007.
- E. Ishay. "Fitting Phase-Type Distributions to Data from a Telephone Call Center". Master's thesis, Technion - Israel Institute of Technology, Heshvan 5763, Haifa, Israel, October 2002.
- N. Jamison. "Speech analytics in the contact center". Witness Systems Corporation web site, <http://www.witness.com>, 2005.
- E.T. Jaynes. "Information theory and statistical mechanics". *Physical Review*, 106(4):620–630, May 1957.
- I. Karatzas and S. Shreve. *Brownian motion and stochastic calculus*, volume 52 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, first edition, 1991.
- S.M. Kay. *Intuitive Probability and Random Processes Using MATLAB*. Springer-Verlag, New York, first edition, 2006.
- S.M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*, volume I. Prentice-Hall, Upper Saddle River, New Jersey, 07458, 1993.
- S.M. Kay. *Fundamentals of Statistical Signal Processing: Detection Theory*, volume II. Prentice-Hall, Upper Saddle River, New Jersey, 07458, 1998.
- ed. Keith Dawson. "Call center industry statistics related to human resources". Incoming Customer Management Institute, at <http://www.incoming.com/statistics/hr.aspx>, June 22 2006.

- F. Kelly. "Charging and rate control for elastic traffic". *European Transactions on Telecommunications*, 8:33 – 37, 1997.
- F. Klebaner. *Introduction to Stochastic Calculus with Applications*. Imperial College Press, 57 Shelton Street, Covent Garden, London WC2H 9HE, UK, second edition, 2005.
- G. Koole. "Assigning a single server to inhomogeneous queues with switching costs". *Theoretical Computer Science*, 182(1-2):203–216, August 1997.
- H. Kushner. *Heavy Traffic Analysis of Controlled Queueing and Communication*, volume 47 of *Applications of Mathematics*. Springer-Verlag, New York, first edition, 2001.
- D. Lay. *Linear Algebra and Its Applications*. Addison-Wesley, New York, first edition, 1994.
- P. L'Ecuyer and E. Buist. "Variance Reduction in the Simulation of Call Centers". In *Proceedings of the 2006 Winter Simulation Conference*, pages 604 – 613, Monterey, CA, Dec. 3-6 2006.
- Alberto Leon-Garcia. *Probability and Random Processes for Electrical Engineering*. Addison-Wesley, Reading, Massachusetts, 2 edition, 1994.
- N. Madras. *Lectures on Monte Carlo Methods*. Fields Institute Monographs. American Mathematical Society, Providence, Rhode Island, 2002.
- A. Mandelbaum. "Call center research bibliography with abstracts". <http://ie.technion.ac.il/serveng>, December 23 2004.
- A. Mandelbaum and A. Stolyar. "Scheduling flexible servers with convex delay costs: heavy traffic optimality of the generalized c-mu rule". *Operations Research*, 52(6):836–855, November-December 2004.
- A. Mandelbaum, W.A. Massey, M.I. Reiman, and A.L. Stolyar. "Waiting time asymptotics for time varying multiserver queues with abandonment and retrials". In *Proceedings of the Annual Allerton Conference on Communication, Control, and Computing*, volume 37, pages 1095 – 1104, 1999.
- W. Martinez and A. Martinez. *Computational Statistics Handbook with MATLAB*. Chapman and Hall/CRC, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487, 2008. ISBN 978-1-58488-566-5.
- J. Mazzola and K. McCardle. "The stochastic learning curve: optimal production in the presence of learning-curve uncertainty". *Operations research*, 45(3):440–450, May-June 1997.
- V. Mehrotra and J. Fama. "Call center simulation modeling: methods, challenges, and opportunities". In *Proceedings of the 2003 Winter Simulation Conference*, volume 1, pages 135 – 143, New Orleans, Louisiana, December 2003. IEEE Press.
- K. Meier-Hellstern. "A fitting algorithm for Markov-modulated Poisson processes having two arrival rates". *European Journal of Operational Research*, 29(3):370–377, 1987.

- G. Mishne, D. Carmel, R. Hoory, A. Roytman, and A. Soffer. "Automatic analysis of call center conversations". In *ACM Conference on Information and Knowledge Management*, Bremen, Germany, October 31 - November 5 2005.
- S. Misra, E. Pinker, and R. Shumsky. "Salesforce design with experience-based learning". *IIE Transactions*, 36(10):941–952, 2004.
- S.G. Nash and A. Sofer. *Linear and Nonlinear Programming*. McGraw-Hill Series in Industrial Engineering and Management Science. McGraw-Hill Inc., New York, first edition, 1996.
- D. Nembhard and N. Osothsilp. "An empirical comparison of forgetting models". *IEEE Transactions on Engineering Management*, 48(3):283 – 291, 2001.
- J.R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York, NY, first edition, 1997.
- R. Nunez-Queija. "A queueing model with varying service rate for ABR". In *Proceedings of the Tenth International Conference on Computer Performance Evaluation: Modelling Techniques and Tools*, pages 93–104. Springer-Verlag, 1998.
- G. Oxton. "The evolution of support: funnel or a cloud?". TIM Seminar, UCSC Silicon Valley Center, Mountain View, CA, April 19 2006.
- A. Parasuraman, V. Zeithaml, and L. Berry. "SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality". *Journal of Retailing*, 64(1):12 – 40, 1988.
- E. Pinker and R. Shumsky. "The efficiency-quality trade-off of cross-trained workers". *Manufacturing and Service Operations Management*, 2(1):32–48, Winter 2000.
- R. Núñez Queija. "A queueing model with varying service rate for ABR". In *Computer Performance Evaluation (Tools)*, pages 93–104, 1998.
- B. Read. "The struggling revolution". Call center magazine web site, <http://www.callcentermagazine.com>, December 2002.
- D.A. Reis. "Learning curves in food services". *Journal of the Operational Research Society*, 42 (8):623–629, 1991.
- Z. Ren and Y. Zhou. "Call center outsourcing: coordinating staffing level and service quality". *Management Science*, 54(2):369 – 383, February 2008.
- A. Ridley, M. Fu, and W. Massey. "Fluid approximations for a priority call center with time-varying arrivals". In *Proceedings of the 2003 Winter Simulation Conference*, New Orleans, Louisiana, December 7-10 2003.
- P. Robert. *Stochastic Networks and Queues*, volume 52 of *Applications of Mathematics*. Springer-Verlag, Berlin, first edition, 2003.
- S. Ross. *Simulation*. Elsevier, San Diego, CA, 2002a.

- S. Ross. *Simulation*. Academic Press, Elsevier, San Diego, third edition, 2002b.
- M. Rossiter. "A switched Poisson model for data traffic". *Australian Telecommunication Research*, 21(1):53–57, 1987.
- G. Ryder and K. Ross. "A Probability Collectives Approach to Weighted Clustering Algorithms for Ad hoc Networks". In *IASTED CCN*, pages 94 – 99, Marina Del Rey, CA, USA, Oct. 24 - 26 2005.
- G. Ryder, K. Ross, and J. Musacchio. "Optimal service policies under learning effects". *Int. J. Services and Operations Management*, 4(6):631–651, 2008.
- V. Rykov and E. Lember. "Optimal dynamic priorities in single-line queueing systems". *Engineering Cybernetics*, 5(1):21–30, 1967.
- S. Salas, E. Hille, and J. Anderson. *Calculus: One and Several Variables, With Analytic Geometry*. John Wiley and Sons, Inc., New York, fifth edition, 1986.
- S. Sayin and S. Karabati. "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, February 2007.
- M. Schilling, P. Vidal, R. Ployhart, and A. Marangoni. "Learning by doing something else: variation, relatedness, and the learning curve". *Management Science*, 49(1):39 – 56, January 2003.
- L. Schrage and L. Miller. "The queue M/G/1 with the shortest remaining processing time discipline". *Operations Research*, 14(4):670–684, July-August 1966.
- S. Shafer, D. Nembhard, and M. Uzumeri. "The effects of worker learning, forgetting, and heterogeneity on assembly line productivity". *Management Science*, 47(12):1639–1653, December 2001.
- O.P. Sharma. *Markovian Queues*. Mathematics and its Applications. Ellis Horwood, 111 Fifth Avenue, New York, NY 10003, first edition, 1990.
- G. Shorack. *Probability for Statisticians*. Springer Texts in Statistics. Springer-Verlag, New York, first edition, 2000.
- A. Shumsky and E. Pinker. "Gatekeepers and referrals in services". *Management Science*, 49(7): 839 – 856, July 2003.
- S. Sikstrom and M. Jaber. "The power integration diffusion model for production breaks". *Journal of Experimental Psychology: Applied*, 8(2):118 – 126, 2002.
- D. Stirzaker. *Elementary Probability*. Cambridge University Press, New York, second edition, 2003.
- T. Tezcan. "State Space Collapse in Many-Server Diffusion Limits of Parallel Server Systems and Applications". PhD thesis, Georgia Institute of Technology, August 2006.

- P. Thompson. "How much did the Liberty shipbuilders forget?". *Management Science*, 53(6): 908–918, June 2007.
- A. Tucker, I. Nembhard, and A. Edmondson. "Implementing new practices: an empirical study of organizational learning in hospital intensive care units". *Management Science*, 53(6):894–907, June 2007.
- R. Wallace and W. Whitt. "A staffing algorithm for call centers with skill-based routing". *Manufacturing and Service Operations Management*, 7(5):276–294, Fall 2005.
- W. Whitt. "The impact of increased employee retention on performance in a customer contact center". *Manufacturing and Service Operations Management*, 8(3):235–252, 2006.
- W. Whitt. "Planning Queueing Simulations". *Management Science*, 35(11):1341–1366, November 1989.
- P. Whittle. *Optimization Over Time: Dynamic Programming and Optimal Control*, volume I. John Wiley and Sons, Ltd., New York, first edition, 1982.
- D.H. Wolpert and S.R. Bieniawski. "Distributed control by Lagrangian steepest descent". *IEEE Conference on Decision and Control*, December 14-17 2004.
- D.H. Wolpert and W.G. MacReady. "No free lunch theorems for optimization". *IEEE Transactions on Evolutionary Computation*, 1(1), April 1997.
- G. Yin and Q. Zhang. *Discrete-Time Markov Chains: Two-Time-Scale Methods and Applications*, volume 55 of *Applications of Mathematics*. Springer-Verlag, New York, first edition, 2005.
- J. Zamiska, M. Jaber, and H. Kher. "Worker deployment in dual resource constrained systems with a task type factor". *European Journal of Operational Research*, 177(3):1507–1519, March 2007.
- E. Zohar, A. Mandelbaum, and N. Shimkin. "Adaptive behavior of impatient customers in tele-queues: theory and empirical support". *Management Science*, 48(4):566–583, April 2002.