Interfaces
Vol. 36, No. 5, September–October 2006, pp. 433–445
ISSN 0092-2102 | EISSN 1526-551X | 06 | 3605 | 0433

informs

DOI 10.1287/inte.1060.0219

Improving Customer Service Operations at Amazon.com

Matthew F. Keblis
Mays Business School, Texas A&M University, College Station, Texas 77843, keblis@tamu.edu

Maomao Chen
Amazon.com, 605 5th Avenue South, Seattle, Washington 98104, mchen@amazon.com

The success of the Internet retailer Amazon.com depends on its providing high-quality customer service. Amazon.com's customer service operations consist of internally and externally managed contact centers. Amazon.com must size its contact centers appropriately, deciding about hiring and training at internally managed centers, and the volume of voice calls and e-mail messages to allocate to external service providers. We developed an approach based on mathematical programming that Amazon.com uses in planning capacity, reducing the average cost of handling a customer contact, and increasing the service level provided customers.

Key words: organizational studies: manpower planning; programming: integer.
History: This paper was refereed.

Amazon.com, Inc. started in 1995 as an Internet retailer of books. Scarcely a year after opening its virtual doors, Amazon was rumored to have achieved annualized revenues of $17 million (Reid 1997, p. 50). Since its inception, the firm has grown rapidly, and it is now a Fortune 500 company with sales in fiscal year 2004 of approximately $7 billion (Amazon.com 2005, p. 25). In less than a decade, Amazon has evolved from just an online bookstore, admittedly with "Earth's biggest selection" (Amazon.com 2003, p. 1), to an Internet retailer that offers new, used, and refurbished items in a number of categories, including music, food, apparel, kitchenware, and consumer electronics.

Making available such a broad array of products reflects Amazon's desire to be the place "where customers can find and discover anything they may want to buy online" (Amazon.com 2003, p. 1). The American Customer Satisfaction Index (ACSI) shows that it has succeeded; in 2001, 2002, and 2003, it received the highest score ever recorded by the ACSI in any service industry. Its success can be attributed partly to the strength of Amazon's customer service operations (CSO). As stated in a recent annual report, "We believe that our ability to establish and maintain long-term relationships with customers and to encourage repeat visits and purchases depends on the strength of customer service operations" (Amazon.com 2003, p. 4).

CSO provides service to customers via internally and externally managed contact centers and features on the company Web site. These features allow customers to perform various activities, including tracking orders and shipments, reviewing estimated delivery dates, and cancelling unshipped items. Customers who cannot resolve their inquiries using the Web site features can call or e-mail customer service representatives (CSRs) available in the contact centers 24 hours a day.

To handle growing sales and their inherent seasonality (the traditional retail variety and that due to Internet usage, which generally declines during the summer), Amazon must size appropriately the capacity of its contact centers (processing network). It must make decisions about hiring and training at internally managed centers and about the volume of voice calls and e-mail messages to allocate to external service providers (*cosourcers*).

Problem Setting and Previous Work

Customers place orders and follow up on orders on the company Web site. Customers who cannot resolve issues using features on the Web site can either call the company's 800 number or send e-mail messages to customer service.

Customer calls and e-mail messages are fielded by CSRs located in internally managed contact centers or in centers operated by vendors with which Amazon has cosourcing agreements. The company-managed contact centers are located in North America (Tacoma, Washington; Grand Forks, North Dakota; Huntington, West Virginia), in Europe (Slough, the United Kingdom; Regensburg, Germany), and Asia (Sapporo, Japan). The cosourcers are spread throughout the world. We focus here on sizing that portion of the processing network that consists of cosourcers and internally managed contact centers located in the United States. From an operational perspective, we can view them as a single virtual contact center.

The e-mail messages and voice calls (customer contacts) number in the millions annually with the peak just before and after Christmas and the nadir in midsummer (Figure 1). The handling time for voice calls and e-mail exchanges depends on such contact attributes as product type, customer type, and purchase type. Amazon uses these attributes to categorize contacts. Most are classified as primary, while the remainder fall into seven speciality categories: hard lines (consumer electronics, home improvement, and kitchen stores), digital (downloads from the Web site, such as software and e-books), wireless (cell phones), institutional buying (corporate accounts), community helper (posting reviews, listmania, and so forth on the Web site), community specialty (quality assurance vis-a-vis community-helper activities), and gift certificates (Figure 2).

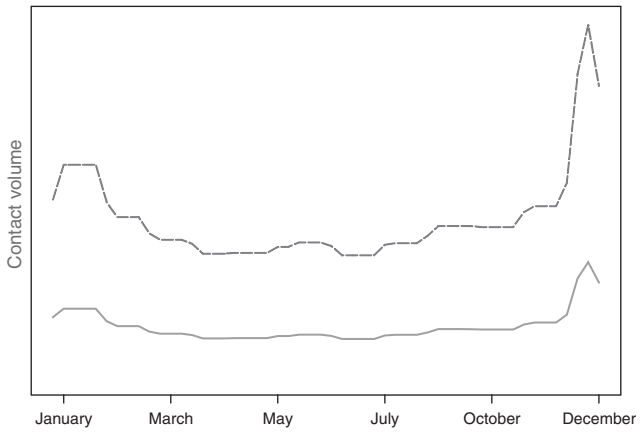

Figure 1: The weekly volume of voice (solid line) and e-mail (broken line) customer contacts shows the typical peak around Christmas.

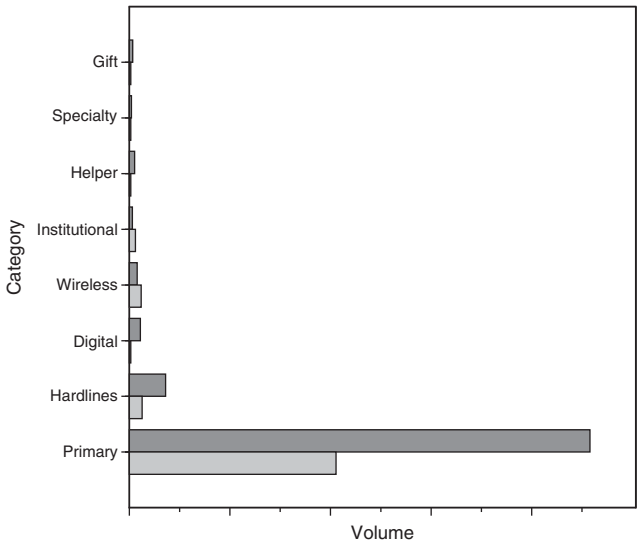

Figure 2: The voice (lower bar) and e-mail (upper bar) contact volume for the primary category outweighs that for the other categories on a typical day.

Amazon classified contacts into categories to reflect the skill sets needed to resolve different issues. It created eight planning groups (PGs) dedicated to processing the contacts in the eight categories. CSRs at internally managed contact centers are assigned to specific PGs and trained to handle both voice and e-mail contacts. All new representatives begin with several weeks of training in the primary PG. Those hired into the other, specialty PGs transfer from the primary PG and undergo additional training. The firm divides the CSRs in each PG into teams, based on their location (contact center).

The CSO's objective is to handle contacts at target service levels. For each of the eight categories, it sets service-level targets for both types of contacts. For voice contacts, the objective is that a specific percentage of callers wait no more than a certain amount of time before speaking with a CSR. For e-mail contacts, the objective is that a specific percentage of all

e-mail messages receive a response within some time. Achieving service-level objectives is a function of the processing network's capacity.

Previous Planning Approach

Amazon began by forecasting demand by product line, for example, for apparel, music, or kitchenware, by week over a time horizon of a year. It then transformed this product-line forecast into a forecast of orders, using weekly time buckets over a horizon of one year. It then converted the point forecasts developed in this fashion into weekly forecasts of e-mail and voice contacts for the eight categories over the planning horizon.

After the development of these contact forecasts, the capacity-planning team in CSO assessed the contact-handling capacity of each PG for each week of the planning horizon. Beginning with week one, it computed the capacity of each speciality PG for handling voice calls and compared it with the corresponding voice forecast. The team addressed capacity shortfalls for speciality PGs by planning to transfer primary PG CSRs to the speciality PGs. When the capacity in a speciality PG exceeded the forecast, it converted the excess capacity, calculated in terms of handling voice calls, into capacity for handling e-mail messages. It compared the value for each speciality PG with the corresponding e-mail forecast and addressed shortfalls by planning to transfer primary PG CSRs to the specialty PG. Any capacity not consumed in handling speciality e-mail then became capacity available for handling primary e-mail contacts.

Once planners had sized the specialty PGs, albeit for only week one of the planning horizon, they focused on primary voice contacts. First, they allocated some forecast voice contacts to cosourcers for handling. Then, they compared the unallocated volume remaining with the capacity in the primary PG for handling voice calls and planned to hire externally to make up any shortfall or to convert excess voice capacity into capacity for handling e-mail. They combined this capacity in the primary PG for handling e-mail with any excess capacity in the speciality PGs and compared the result with the forecast of primary e-mail contacts less some portion allocated to cosourcers. If the capacity was less than the forecast, they planned to hire externally. After planning for the first week, they repeated the steps for the remaining weeks of the horizon to develop a complete capacity plan.

The company planned in this way every week of the year. Although planners used a spreadsheet for the calculations, close to a day was still required to investigate a single scenario. CSO managers recognized this shortcoming and the lack of rigor in evaluating important trade-offs. They asked us to help them strengthen the capacity-planning process, specifying that any new approach had to address three important issues.

Three Issues

CSO managers thought that three important issues were not adequately considered in the existing planning process: how they added CSRs to teams, differences in contracts with cosources, and staffing and service levels. CSO managers added CSRs to teams when they brought on external hires or transformed primary PG CSRs to specialty PGs. Traditionally they added or removed CSRs from teams to maintain the existing proportion of PG members on the various teams (at each contact center); for example, if 20 percent of the CSRs of a PG were located at a particular contact center, then the managers would hire and make transfers for the entire PG so that 20 percent of the CSRs of the PG would continue to be located at that center. They ignored the fact that average productivity varied across teams within a PG and that the average wage differed among centers.

Second, contract terms differed across cosourcers. For some cosourcer contracts, Amazon incurred costs per contact handled. For other cosourcer contracts, Amazon incurred a fixed charge if the volume allocated to the cosourcer fell below a minimum threshold; otherwise, it followed an all-units discount price schedule. Furthermore, some of these contracts had ceilings on the volume of contacts. If the volume of contacts allocated to the cosourcer exceeded some maximum amount in a time period, the minimum threshold for future time periods would ratchet upwards. Amazon allocated contacts to cosourcers

with such contracts to meet any minimum thresholds. For cosourcers with contracts based on the number of contacts handled, Amazon allocated contacts to ensure that it restricted the percentage of primary contacts handled outside of an internally managed contact center. It sought to minimize the risks of relying on cosourcers.

Regarding staffing and service levels, Amazon established the capacity for handling both types of contacts at the minimum levels needed to maintain system stability. It set the number of CSR hours available in a week for handling voice calls to a quantity just barely exceeding the expected number of hours of voice-call-related work that a PG would need to perform. It established capacities for e-mail in a similar manner. Using this approach, it ignored the random behavior of contact arrival rates and handling times. Even so, it achieved service-level objectives for both types of contacts regularly. For e-mail messages, the company set response-time targets that allowed CSRs to postpone e-mail work. For voice calls, however, CSRs could not postpone responding. Although the spreadsheet-based approach sized voice and e-mail capacities independently for a PG, the operational reality is that CSRs handle both voice and e-mail requests and interrupt their processing of e-mail messages to handle voice calls as they arrive. Because most contacts are e-mail messages, the centers regularly achieved voice target service levels despite shortcomings in planning. Nevertheless, the spreadsheet-based approach possessed no lever that allowed CSO managers to specify a service-level objective and see its impact on staffing levels.

Literature

Management science analysts have only recently considered the problem of determining the capacity required to serve customer classes differentiated by response-time requirements, where customer arrival rates are time dependent. Gans et al. (2003) provide a comprehensive summary of the state of call-center research pertaining to capacity management.

Whitt (1999) examined the determination of capacity in a setting with two customer classes, one requiring immediate response and the other, response within a day. To determine the capacity required for the highest priority class, he employed an $M/G/\infty$ model and normal approximation with a target probability that a service request will be delayed before service begins. For less-time-sensitive customers, he used a normal approximation alone with another target probability that all daily demand will be met. He showed that the capacity the service provider needs is the maximum of the two previously defined requirements.

Armony and Maglaras (2004) considered a call center in which customers, assuming that their calls are not answered immediately, can choose to hold for service (class 1), indicate their desire to be called back (class 2), or simply balk, making the choice after being informed of the expected delay. The authors modeled the dynamics of this environment as an $M/M/N$ multiclass system and performed an asymptotic analysis to choose the minimum number of agents to guarantee performance measures, such as a bound on the expected waiting time of class 1 customers and bounds on the probability that the waiting time exceeds some threshold.

Chen and Henderson (2001) examined a call-center setting with two or more classes where the objective is that, for each class, a class-specific percentage of calls are answered within a class-specific time frame. For the highest priority class, the authors leveraged transform methods to determine the probability that a call will be delayed longer than a certain period of time (the tail probability), while for other classes they used Markov's inequality to obtain a bound on waiting-time performance. To establish the required staffing level, they increased the number of agents until the tail probability was as small as desired and each Markov inequality was satisfied.

Harrison and Zeevi (2005) considered centers with more than two customer classes (and more than one pool of agents) where the objective is to minimize the sum of staffing costs and expected abandonment penalties for the various classes. They assumed time-dependent arrival rates that can vary stochastically. They used stochastic fluid models to reduce the staffing problem to a multidimensional newsvendor problem, which they then solved numerically with a combination of linear-programming and simulation methods.

Gans and Zhou (2002) examined the problem of determining the number of employees of different speed or skill levels to staff, where workers gain in speed or skill and become capable of handling more classes of customers or kinds of work. They employed a Markov decision process model to determine hiring and promotion policies that minimize hiring, compensation, and other operational costs. Gans and Zhou (2004) focused on a situation where there are two classes of customers (high and low value), and the problem is to determine the staffing level at an outsourcer handling the low-value customers. They examined and compared three approaches for determining the outsourcer's staffing levels.

Like Whitt (1999) and Chen and Henderson (2001), we show how to apply queuing-related concepts in setting staffing levels in contact centers with more than one customer class. However, whereas they focused on determining staffing levels to attain specific service-level objectives irrespective of cost, we focused on meeting such objectives as inexpensively as possible given a global processing network with differing economics throughout its parts. Such a perspective might have led us to consider call-routing issues, like Armony and Maglaras (2004), Gans and Zhou (2004), and Harrison and Zeevi (2005), but we chose not to investigate such matters when we worked on our problem given the added complexity of call routing and our desire to quickly improve capacity planning at Amazon. Gans and Zhou (2002) allowed stochastic turnover and considered outsourcing as we do; however, they considered a firm operating only a single internal call center. We applied existing methods, with some modification, to planning the capacity of a firm with multiple internal contact centers and multiple outsourcing options where the objective is to minimize total costs subject to service-level targets.

Solution Approach

From the outset, we thought that we could represent most of the essential elements of the capacity-planning problem CSOs faced, with one notable exception, naturally within an optimization framework. The exception was the third issue concerning staffing and service levels; we thought we would need to apply some concepts from queueing theory. We developed a two-stage solution approach. In the first stage, we adjusted contact forecasts previously generated using concepts from queueing to take into account different sources of uncertainty and service-level objectives. In the second stage, we solved an optimization model, using as input the adjusted forecasts and other relevant data, to determine the best allocation of contacts across all centers and the staffing levels at internal ones.

We began our optimization-based approach with a collection of contact forecasts adjusted to account for the randomness inherent in contact arrival rates and handling times, and the existence of service-level objectives. Our adjustment procedure was shaped by our observation that for those categories with a large volume of e-mail contacts, CSO's voice service levels regularly met targeted objectives. We take into account the e-mail forecast when generating the corresponding adjusted voice-call forecast.

We will simplify our explanation of the adjustment procedure by focusing on an individual contact category and a single week of the planning horizon. The task thus becomes, for the week of interest, to produce a pair of adjusted forecasts, one for e-mail and one for voice. The information we have to work with in computing these numbers includes hourly forecasts of e-mail and voice contacts for the week concerned, an average CSR handling time for each type of contact, and service-level objectives for both contact types. We denote the forecast of e-mail (voice) in hour h of the week as $\lambda_{e,h}$ ($\lambda_{v,h}$). We denote the average rate at which CSRs handle e-mail (voice) contacts per hour as μ_e (μ_v). Finally, service-level objectives are of the telephone-service-factor variety, that is, at least x percent of contacts answered within y time units.

Adjustment Procedure

The adjustment procedure consists of five steps.

Step 1

We determine the minimum number of CSRs needed to prevent the number of unprocessed contacts from growing to infinity. We perform this calculation for both types of contacts for each hour of the week, and it amounts to dividing each hourly forecast by the relevant service rate. In the case of e-mail, the resulting

value $\lambda_{e,h}/\mu_e$ for each hour h of the week is denoted as $\rho_{e,h}$. Similarly for voice, $\rho_{v,h} = \lambda_{v,h}/\mu_v$. We perform the calculations in the first step without regard to service-level objectives.

Step 2
We determine the minimum number of CSRs needed to achieve the specified service-level objective for voice contacts using the Erlang C formula to perform the calculation for each hour of the week, using as inputs $\lambda_{v,h}$, μ_v, and the specified target service level. For each hour h of the week, we denote the resulting value as $\tilde{\rho}_{v,h}$.

Step 3
Because our optimization model requires weekly forecasts and the data that we are working with is hourly, we aggregate this hourly information. We perform an aggregation for each day of the week for each of the above collections of data, producing three values for each day d of the week: θ_d which is a summation of $\rho_{v,h}$ for a given day, ϕ_d which is a summation of $\rho_{e,h}$ for a given day, and $\tilde{\theta}_d$ which is a summation of $\tilde{\rho}_{v,h}$ for a given day.

Step 4
We establish the weekly forecast for voice contacts to use in the optimization model. We arrive at this weekly value by first assessing the capacity needed for each day of the week. We do this by evaluating the following inequality for each day d of the week: $\theta_d + \phi_d > \tilde{\theta}_d$. When this inequality is true, the forecast amount of postponable work for the day (given in terms of CSRs by ϕ_d) is sufficient to buffer against voice-contact-related variability. We set the voice-contact forecast for the day equal to $\theta_d \cdot \mu_v$, which we denote as γ_d. If the inequality evaluates to false, then the e-mail volume is not sufficient to buffer against voice-contact-related variability and γ_d is set equal to $\tilde{\theta}_d \cdot \mu_v$. By summing over γ_d for a week, we produce the weekly forecast for voice contacts, which we denote as V_t^k, where k indicates the contact category and t the week of interest.

Step 5
We establish the weekly forecast for e-mail contacts, which we denote as E_t^k, where k indicates the contact category and t the week of interest. We arrive at it by summing over $\phi_d \cdot \mu_e$ for the week, which completes our task of producing an adjusted forecast for e-mail and an adjusted forecast for voice for the week concerned. We then apply the adjustment procedure to the voice and e-mail contact forecasts for all the remaining categories and weeks of the planning horizon. This collection of adjusted forecasts becomes input to the optimization model.

This adjustment procedure will generate aggregate CSR requirements and ultimately forecasts that are identical for different call-volume scenarios; for example, a scenario where the call-volume pattern dictates the need for 10 CSRs per hour over a 10-hour day will generate the same aggregate requirement as a scenario where the need is for 100 CSRs in one hour and none in any other time period. Nonetheless, the adjustment procedure recurrently generates output that is meaningful for two reasons: (1) While the call-volume pattern Amazon faces over a workday is certainly not stationary, it is also not anywhere near as lumpy as depicted in the latter, second scenario. (2) Although we can expect the call volume to be much higher in some hours than it is in others, Amazon does not necessarily have to increase staffing at its internal contact centers at such times because the cosourcing agreements it has allow it to look to cosourcers to provide capacity when it provides enough advance notice. Put another way, the flexibility afforded by the cosourcing agreements allows Amazon to plan to handle a baseline load internally and to push to cosourcers any excess volume. Aksin et al. (2004) discuss the economic rationale for this type of agreement.

Optimization Model

The optimization model we developed is a mixed-integer program (appendix). The program outputs a minimum-cost capacity plan for processing the contacts forecast for a given finite planning horizon, detailing for each week decisions regarding hiring and training CSRs and the volume of contacts to allocate to each cosourcer.

Objective Function
The terms of the objective (cost) function fall into two categories: those pertinent to internally managed contact centers, and those related to cosourcers. We identified four cost drivers as relevant for each week t and

each team (i, j), where i denotes the PG and j the contact-center location: the number of normal-time hours (n_t^{ij}) CSRs work, the number of overtime hours (o_t^{ij}) CSRs work, the number of new CSRs hired (h_t^{ij}), and the number of CSRs transferred (s_t^{ij}) from the primary PG at a contact center to one of the speciality PGs at that center. In our mathematical program, we model the costs associated with these drivers using linear expressions (appendix). For the first two drivers (variables), the coefficients are normal and overtime hourly wages, respectively, while for the third and fourth variables, the coefficients capture training and hiring/or transfer expenses in addition to wages paid during the training period. Because many CSRs are contract employees, costs for decreasing the size of the workforce are minimal and hence ignored.

The remaining terms of the objective function concern the cost of contracting with cosourcers to handle some primary voice and e-mail contacts. Amazon employs two kinds of agreements with its cosourcers: a take-or-pay contract with an all-units discount price schedule, and a per-contact contract.

Under a take-or-pay contract, Amazon guarantees a cosourcer a contract-specific minimum weekly payment regardless of the volume of contacts it allocates to the cosourcer (Figure 3). The fifth term of the objective function indicates that every week t Amazon makes a payment of F_t^i to each cosourcer i. The matter of a minimum payment becomes irrelevant, however, if the volume of contacts allocated exceeds a contract-specific minimum threshold, at which point the payment made becomes a function of the number of contacts the cosourcer handles: Amazon then pays only a fee per contact handled, with the fee depending on the actual volume allocated and becoming progressively lower as the volume allocated rises.

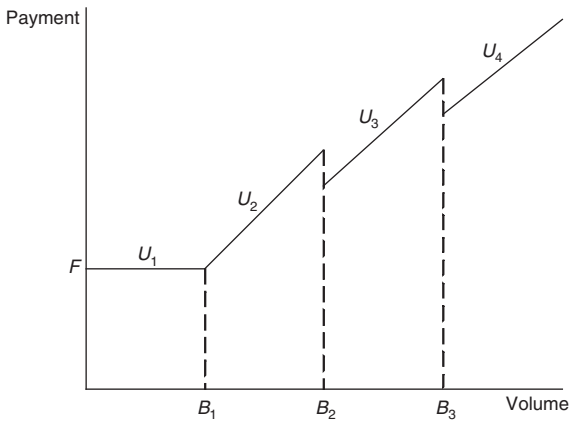

Figure 3: In a take-or-pay contract, when the volume allocated is less than or equal to B_1, the minimum threshold, the volume pushed to the cosourcer falls within the first range of the price schedule and the cosourcer receives a minimum payment F. The fee per contact in the first range, U_1, is equal to the slope of the payment function in that range. When the volume allocated is between B_1 and B_2, the volume pushed to the cosourcer falls into the second range of the price schedule, with a fee per contact of U_2, and so forth.

In our objective function, the sixth and seventh terms adjust the payment made when contact volumes exceed the minimum threshold. The sixth term offsets, when the volume allocated exceeds the minimum threshold, the minimum payment made to a cosourcer per the fifth term. We accomplish this by setting the negative of F_t^i as the coefficient of the binary variable $y_{k,t}^i$, which takes the value 1 when the volume of contacts allocated to cosourcer i in week t falls into range k. Because we seek an offsetting effect only when the volume allocated exceeds the minimum threshold, we include such a term in the objective function only when the subscript k of the variable $y_{k,t}^i$ is greater than one. The seventh and last term captures per-contact handling charges. It contains the variable $x_{k,t}^i$, which indicates the number of contacts handled by cosourcer i in week t if the total volume processed falls into range k of the price schedule. For a given week t and cosourcer i, one such variable exists for each range in the price schedule of the cosourcer. Of this collection of $\{x_{k,t}^i\}$, only one will ever be greater than zero in a given week t for cosourcer i, and that variable will correspond to the same range of the price schedule as the $y_{k,t}^i$ that takes the value 1. Because the coefficient of each $x_{k,t}^i$ is the relevant fee per contact ($U_{k,t}^i$), it captures the payment due to handling charges for each week t and cosourcer i.

A per-contact contract is just a special case of take-or-pay. In a per-contact contract, Amazon does not guarantee a minimum weekly payment; hence the value of F_t^i is equal to zero for each cosourcer i under a per-contact contract for every week t. That makes the fifth and sixth terms of the objective function irrelevant under a per-contact contract; the only meaningful term therefore is the last involving the variable $x_{k,t}^i$. With a per-contact contract, the fee per contact

does not vary with the volume of contacts allocated; hence the price schedule has only a single range.

Constraints

We can divide the constraints largely into two categories, those for internally managed contact centers, and those for cosourcers. The first two constraints we discuss, however, represent a point of intersection. Defining $v_t^{ij,k}$ as the number of category k (k equal to 1 denotes primary) voice contacts allocated to team (i, j) (i denotes the PG and j the contact center location) and c_t^i as the number of primary contacts allocated to cosourcer i, constraint 1 indicates that the sum of primary voice contacts allocated over all teams and over all cosourcers that handle voice contacts must be at least as large as V_t^1, the voice-call forecast. Defining $e_t^{ij,k}$ as the number of category k e-mail contacts allocated to team (i, j), the second constraint indicates that the sum of primary e-mail contacts allocated over all teams and over all cosourcers that handle e-mail contacts must be at least as large as E_t^1, the e-mail forecast.

Constraint 3 resembles constraint (1); for each category k of voice contacts (with the exception of primary), it establishes that Amazon must allocate contacts to each team (i, j), given by $v_t^{ij,k}$, when summed over all teams, at least as large as V_t^k, the voice-call forecast. Unlike the first constraint, the third contains no cosourcer-related term. The fourth constraint resembles the second. The remaining constraints follow from the first four in some manner.

Constraint 5 requires that the number of normal (n_t^{ij}) and overtime (o_t^{ij}) hours each team (i, j) works (adjusted by a shrinkage factor that captures that not all hours a CSR spends at work are spent productively) must be at least as many as the number of hours team (i, j) allocates to handling contacts. We arrive at this latter amount by adding the number of hours allocated by team (i, j) to handling voice contacts to the number of hours team (i, j) allocates to handling e-mail. We find the number of hours team (i, j) allocates to handling voice by multiplying $v_t^{ij,k}$ (each PG handles only its own voice calls so the value of k is equal to the value of i) by the average handling time of a voice call by team (i, j). Each team will handle its own e-mail, and speciality PGs may also handle primary e-mail. Hence, we find the number of hours team (i, j) allocates to handling e-mail by multiplying $e_t^{ij,k}$ by the average handling time of a category k (k equal to i) e-mail message by team (i, j) and adding that to $e_t^{ij,1}$ multiplied by the average handling time of a primary e-mail message by team (i, j).

Constraint 6 specifies that the number of overtime hours (o_t^{ij}) that each team (i, j) can work is bounded by a percentage of the normal hours (n_t^{ij}) each team (i, j) works, while the number of normal hours (n_t^{ij}) each team (i, j) works is by constraint 7 bounded by w_t^{ij}, the number of CSRs on team (i, j), multiplied by the number of normal hours in a standard work week. Two constraints capture the number of CSRs on a team. For a team that is part of the primary PG, constraint 8 sets w_t^{ij}, the number of CSRs on team (i, j) in week t, equal to the number available the previous week (w_{t-1}^{ij}) (adjusted by an attrition rate reflecting occasional voluntary departures), less any involuntary separations (d_t^{ij}), less the planned transfer of CSRs to any speciality PG (s_t^{ij}; the superscript denoting the destination team), but augmented by any new outside hires (h_t^{ij}). For each team that is a member of a speciality PG, constraint (9) performs a similar function, capturing planned in-bound transfers, that is, from the primary PG, the only way of increasing the number of CSRs in a speciality PG; there are no outside hires.

Constraints 10 through 15 concern risk mitigation. The first two concern teams in internally managed contact centers. Constraint 10 indicates that for each category k of voice contacts, the number allocated to each team (i, j), given by $v_t^{ij,k}$, must be less than some percentage of V_t^k, the voice-call forecast. Constraint 11 holds similarly for e-mail. Constraints 12 through 15 concern managing cosourcer-related risk. Constraint 12 indicates that the number of primary voice contacts allocated to each cosourcer must be less than some percentage of the voice-call forecast, while constraint 14 limits the number of primary voice contacts allocated to all cosourcers combined to less than some percentage of the number of voice calls forecast. Constraints 13 and 15 are equivalent constraints for e-mail.

The remaining constraints, except those that indicate whether a variable is continuous or integer, concern cosourcers and fall into two categories: contract cost and contract smoothing. We use the

contract-cost constraints (16 through 19) to ensure that Amazon obtains the most attractive prices of the take-or-pay contract only when they meet the required volume minimums. We use the contract-smoothing constraints (20 through 25) to constrain variation in the number of contacts allocated week to week to each cosourcer. We do this by establishing thresholds. If Amazon pushes more volume to a cosourcer than a threshold (monitored by 20 and 21) or less (monitored by 22 and 23), then new thresholds become established and the volume pushed to that cosourcer henceforth is not allowed to cross the newly established thresholds for a fixed amount of time (enforced by 24 and 25).

Results

Each week, the capacity-planning team in CSO employs our solution approach. When the planning horizon is 52 weeks, as it is at the beginning of a calendar year, the optimization model consists of approximately 134,000 constraints and almost 16,000 variables, where a little over 1,000 of these are both binary and integer. The model is encoded as an AMPL program and is solved using CPLEX on an HP 9000 Superdome server with a 1.1 GHz processor. Each run of the model requires slightly less than five minutes of computing time. A planner can investigate a single scenario (inputs adjusted, model executed, and output analyzed) in less than an hour, a process that formerly consumed an entire day. Now the capacity-planning team can examine a larger number of scenarios and consider uncertainty by performing sensitivity analysis on the inputs to the planning process. After analyzing the output for a set of scenarios, the planners pass on information for the time horizon of interest to three groups. They inform Amazon Human Resources of the number of new CSRs Amazon will need to hire, CSO managers of the transfers needed into and out of their PGs, and cosourcers of future contact volumes.

The new approach saves time and therefore enables additional scenario analysis and, most important, brings optimization to bear directly on the planning process. Planners previously considered cost trade-offs by analyzing the outputs of the spreadsheet model. Our optimization model captures these trade-offs explicitly and greatly increases annual operational cost savings.

Managers thought that three important issues did not receive due consideration with spreadsheet-based planning: First was adding new CSRs to PGs without regard to productivity and wage differences. The model revealed that Amazon should stop processing e-mail at one internally managed center or change its process or provide further training to CSRs. Second was allocating contacts to cosourcers. The model revealed that some cosourcers were more expensive for processing voice calls than some internally managed centers. We discovered this by forcing the model to allocate contacts according to existing practice and then allowing it to allocate contacts as it deemed optimal. We found that Amazon could save over one million dollars by handling more calls internally. We attributed the savings largely to smoothing constraints in the cosourcers' contracts that established new long-lasting thresholds when an existing threshold was exceeded. Third was a lack of consideration between service objectives and staffing costs. With the forecast adjustment procedure we incorporated into our approach, planners can evaluate the cost effects of changing service parameters, such as target response times or limits on customers' waiting times (Figure 4). Kim Rachmeler, Amazon.com's vice president of worldwide customer service, said "These advancements in planning our capacity and optimizing our contact allocation plans have significantly improved our ability both to respond to customers quickly, which improves customer experience, and also to lower our costs, which increases corporate flexibility" (personal communication, 2003).

Although we developed our approach with the weekly planning process in mind, the benefits extend to contract negotiations with cosourcers. Periodically, Amazon revisits the terms of its existing agreements with each of its cosourcers. Previously, the tools available for investigating cosourcer relationships were limited and time consuming. Our optimization model yielded insights concerning the costs to Amazon of the parameters (volume thresholds, and the length of time that volume allocated to a cosourcer is required to remain between a pair of newly established thresholds after the breaching of previously established ones) of the contract-smoothing pieces of contracts.

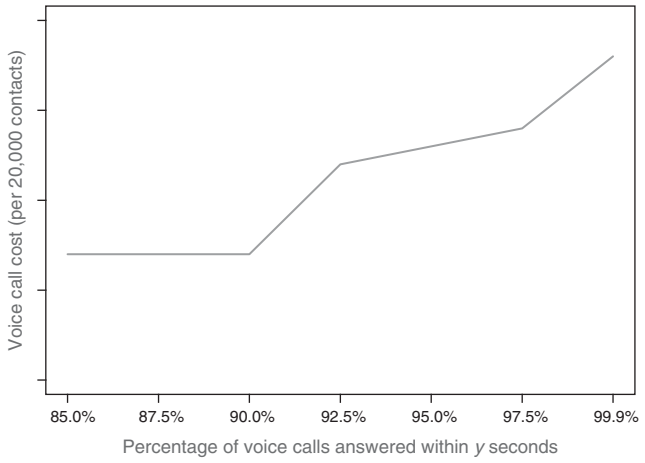

Figure 4: In the flat range of the curve it is possible to attain successively higher voice service levels without incurring additional cost because the number of "e-mail handlers" is sufficiently large and each of these CSRs can switch to processing voice calls without any operational delay. Beyond the flat range, the combined number of e-mail and voice-call handlers is smaller than that required to achieve the desired service levels, making it necessary to add resources essentially dedicated to processing voice calls, which causes the curve to rise with a trajectory that depends on the mix of cosourcers and internal hires employed.

CSO managers are now able to understand when contract-smoothing parameters are actually constraining operational flexibility in the Amazon processing network (and hence raising its cost of operation) versus when they appear to be, but actually are not. This is information CSO managers find useful when negotiating new contracts with cosourcers as they assess whether to make specific concessions.

Appendix

Parameters

$\mathcal{C} = \{1, \ldots, C\}$ is the set of contact categories where 1 denotes primary and 2 through C the speciality categories.

$\mathcal{L} = \{1, \ldots, L\}$ is the set of contact-center locations.

$\mathcal{P} = \{1, \ldots, P\}$ is the set of planning groups (PGs), where 1 denotes the primary PG and 2 through P, the speciality PGs.

$\mathcal{G} = \{(i, j) \mid i = 1, \ldots, P,\ j = 1, \ldots, L\}$ is the set of teams.

$\mathcal{Q} = \{1, \ldots, Q\}$ is the set of cosourcers.

$\mathcal{Q}^v$ = subset of $\mathcal{Q}$ that handles voice calls.

$\mathcal{Q}^e$ = subset of $\mathcal{Q}$ that handles e-mail.

T = number of weeks in the planning horizon.

V_t^k = number of category k voice contacts forecast for week t.

E_t^k = number of category k e-mail contacts forecast for week t.

$\mu_{ij,k}^{-1}$ = average handling time (in hours) of a category k voice contact by team (i, j).

$\hat{\mu}_{ij,k}^{-1}$ = average handling time (in hours) of a category k e-mail contact by team (i, j).

N_t^{ij} = normal time wage for a CSR on team (i, j) in week t.

O_t^{ij} = overtime wage for a CSR on team (i, j) in week t.

H_t^{ij} = cost to hire and train a new CSR for team (i, j) in week t.

S_t^{ij} = costs related to switching a CSR to team (i, j) in week t.

W^{ij} = number of normal hours in the work week of a CSR on team (i, j).

γ_t^{ij} = upper bound (expressed as a proportion of normal hours) on number of overtime hours that may be worked in week t by team (i, j).

δ^{ij} = shrinkage factor (proportion of a CSR's time on team (i, j) lost to things like breaks, absenteeism, and ongoing training).

α^{it} = attrition factor (proportion of CSRs on team (i, j) that voluntarily leave the firm).

τ = number of weeks before a newly hired CSR becomes a productive worker.

$\hat{\tau}$ = number of weeks before a CSR that transfers from the primary PG to a speciality PG becomes productive as a specialist.

β_t^{ij} = upper bound (expressed as a proportion of forecast voice contacts) on number of voice contacts that may be handled by team (i, j) in week t.

$\hat{\beta}_t^{ij}$ = upper bound (expressed as a proportion of forecast e-mail contacts) on number of e-mail contacts that may be handled by team (i, j) in week t.

ξ_t^i = upper bound (expressed as a proportion of forecast voice or e-mail contacts) on number of contacts that may be handled by cosourcer i in week t.

$\hat{\xi}_t^v$ = upper bound (expressed as a proportion of forecast voice contacts) on number of voice contacts that may be handled by all cosourcers combined in week t.

$\hat{\xi}_t^e$ = upper bound (expressed as a proportion of forecast e-mail contacts) on number of e-mail contacts that may be handled by all cosourcers combined in week t.

A^i = number of break points in the price schedule of cosourcer i; equals 0 (zero) when the price schedule does not involve quantity discounts.

$\mathcal{B}^i = \{B_1^i, \ldots, B_{A_i}^i\}$ is the set of volume break points in the price schedule of cosourcer i, where $0 < B_1^i < B_2^i < B_3^i \cdots$.

R^i = number of distinct ranges in the price schedule of cosourcer i, where assuming $\mathcal{B}^i \neq \varnothing$ the first range is $[0, B_1^i]$; note that $R^i = A^i + 1$.

F_t^i = fixed payment made to cosourcer i unless the total volume of contacts processed by the cosourcer in week t exceeds a specific threshold.

$U_{k,t}^i$ = per-contact handling fee at cosourcer i in week t when the total volume of contacts processed falls into range k.

ζ^i = threshold expressed as a proportion of the number of contacts pushed to cosourcer i.

Ω^i = number of weeks the volume pushed to cosourcer i must remain within newly established limits (upper and lower thresholds) after crossing (exceeding or falling below) a previously existing threshold.

M = a very large number.

Variables

$v_t^{ij,k}$ = number of category k voice contacts handled by team (i, j) in week t.

$e_t^{ij,k}$ = number of category k e-mail contacts handled by team (i, j) in week t.

c_t^i = number of primary contacts handled by cosourcer i in week t.

n_t^{ij} = number of planned normal hours for team (i, j) in week t.

o_t^{ij} = number of planned overtime hours for team (i, j) in week t.

w_t^{ij} = number of CSRs needed on team (i, j) in week t.

h_t^{ij} = number of planned outside hires for team (i, j) in week t.

s_t^{ij} = number of planned CSR transfers to speciality team (i, j) from the colocated primary team in week t.

d_t^{ij} = number of involuntary departures from team (i, j) in week t.

$x_{k,t}^i$ = number of contacts handled by cosourcer i in week t if the total volume processed falls into range k of its price schedule; 0 otherwise.

$y_{k,t}^i$ = 1 if the number of contacts handled by cosourcer i in week t falls into range k of its price schedule; 0 otherwise.

z_t^i = 1 if the proportional increase in the number of contacts pushed to cosourcer i in week t is greater than ζ^i, when compared to the week prior; 0 otherwise.

$\hat{z}_t^i$ = 1 if the proportional decrease in the number of contacts pushed to cosourcer i in week t is greater than ζ^i, when compared to the week prior; 0 otherwise.

Formulation

$$\min \sum_{t=1}^{T} \sum_{(i,j)\in\mathcal{G}} \left(N_t^{ij} n_t^{ij} + O_t^{ij} o_t^{ij}\right) + \sum_{t=1}^{T} \sum_{\{(i,j)\in\mathcal{G} \mid i=1\}} H_t^{ij} h_t^{ij}$$
$$+ \sum_{t=1}^{T} \sum_{\{(i,j)\in\mathcal{G} \mid i\neq 1\}} S_t^{ij} s_t^{ij} + \sum_{t=1}^{T} \sum_{i\in\mathcal{C}} F_t^i$$
$$- \sum_{t=1}^{T} \sum_{i\in\mathcal{C}} \sum_{k=2}^{R^i} F_t^i y_{k,t}^i + \sum_{t=1}^{T} \sum_{i\in\mathcal{C}} \sum_{k=1}^{R^i} U_{k,t}^i x_{k,t}^i$$

$$\text{s.t.} \sum_{(i,j)\in\mathcal{G}} v_t^{ij,1} + \sum_{i\in\mathcal{C}^v} c_t^i \geq V_t^1, \quad t = 1, \ldots, T, \tag{1}$$

$$\sum_{(i,j)\in\mathcal{G}} e_t^{ij,1} + \sum_{i\in\mathcal{C}^e} c_t^i \geq E_t^1, \quad t = 1, \ldots, T, \tag{2}$$

$$\sum_{\{(i,j)\in\mathcal{G} \mid i=k\}} v_t^{ij,k} \geq V_t^k$$
$$\forall\, k \in \mathcal{C},\ k \neq 1,\ t = 1, \ldots, T, \tag{3}$$

$$\sum_{\{(i,j)\in\mathcal{G} \mid i=k\}} e_t^{ij,k} \geq E_t^k$$
$$\forall\, k \in \mathcal{C},\ k \neq 1,\ t = 1, \ldots, T, \tag{4}$$

$$\mu_{ij,i}^{-1} v_t^{ij,i} + \sum_k \hat{\mu}_{ij,k}^{-1} e_t^{ij,k} \leq (1 - \delta^{ij})\left(n_t^{ij} + o_t^{ij}\right)$$
$$\forall\, (i, j) \in \mathcal{G},\ t = 1, \ldots, T, \tag{5}$$

$$o_t^{ij} \leq \gamma_t^{ij} n_t^{ij} \quad \forall\, (i, j) \in \mathcal{G},\ t = 1, \ldots, T, \tag{6}$$

$$W^{ij} w_t^{ij} \geq n_t^{ij} \quad \forall\, (i, j) \in \mathcal{G},\ t = 1, \ldots, T, \tag{7}$$

$$w_{t-1}^{1j}(1-\alpha^{1j}) - d_t^{1j} - \sum_{\{i\in\mathcal{P}\,|\,i\neq 1\}} s_t^{ij} + h_{t-\tau}^{1j} = w_t^{1j}$$
$$\forall\, j\in\mathcal{L},\ t=1,\ldots,T, \quad (8)$$

$$w_{t-1}^{ij}(1-\alpha^{ij}) - d_t^{ij} + s_{t-\hat{\tau}}^{ij} = w_t^{ij}$$
$$\forall\,(i,j)\in\mathcal{G},\ i\neq 1,\ t=1,\ldots,T, \quad (9)$$

$$v_t^{ij,k} \le \beta_t^{ij,k} V_t^k$$
$$\forall\,(i,j)\in\mathcal{G},\ \forall k\in\mathcal{C},\ t=1,\ldots,T, \quad (10)$$

$$e_t^{ij,k} \le \hat{\beta}_t^{ij,k} E_t^k$$
$$\forall\,(i,j)\in\mathcal{G},\ \forall k\in\mathcal{C},\ t=1,\ldots,T, \quad (11)$$

$$c_t^i \le \xi_t^i V_t^1 \quad \forall\, i\in\mathcal{Q}^v,\ t=1,\ldots,T, \quad (12)$$

$$c_t^i \le \xi_t^i E_t^1 \quad \forall\, i\in\mathcal{Q}^e,\ t=1,\ldots,T, \quad (13)$$

$$\sum_{i\in\mathcal{Q}^v} c_t^i \le \hat{\xi}_t^v V_t^1, \quad t=1,\ldots,T, \quad (14)$$

$$\sum_{i\in\mathcal{Q}^e} c_t^i \le \hat{\xi}_t^e E_t^1, \quad t=1,\ldots,T, \quad (15)$$

$$x_{k,t}^i - B_k^i y_{k,t}^i \le 0$$
$$\forall\, i\in\mathcal{Q},\ k=1,\ldots,R^i-1,\ t=1,\ldots,T, \quad (16)$$

$$x_{k,t}^i - (B_{k-1}^i+1) y_{k,t}^i \ge 0$$
$$\forall\, i\in\mathcal{Q},\ k=2,\ldots,R^i,\ t=1,\ldots,T, \quad (17)$$

$$c_t^i = \sum_{k=1}^{R^i} x_{k,t}^i \quad \forall\, i\in\mathcal{Q},\ t=1,\ldots,T, \quad (18)$$

$$\sum_{k=1}^{R^i} y_{k,t}^i = 1 \quad \forall\, i\in\mathcal{Q},\ t=1,\ldots,T, \quad (19)$$

$$M(1-z_t^i) \ge (1+\zeta^i)c_{t-1}^i - c_t^i$$
$$\forall\, i\in\mathcal{Q},\ t=1,\ldots,T, \quad (20)$$

$$Mz_t^i \ge c_t^i - (1+\zeta^i)c_{t-1}^i \quad \forall\, i\in\mathcal{Q},\ t=1,\ldots,T, \quad (21)$$

$$M(1-\hat{z}_t^i) \ge c_t^i - (1-\zeta^i)c_{t-1}^i$$
$$\forall\, i\in\mathcal{Q},\ t=1,\ldots,T, \quad (22)$$

$$M\hat{z}_t^i \ge (1-\zeta^i)c_{t-1}^i - c_t^i \quad \forall\, i\in\mathcal{Q},\ t=1,\ldots,T, \quad (23)$$

$$M(1-z_t^i) \ge c_{t+\omega}^i - (1+\zeta^i)c_t^i$$
$$\forall\, i\in\mathcal{Q},\ t=-\Omega^i+1,\ldots,T,\ \omega=1,\ldots,\Omega^i, \quad (24)$$

$$-M(1-\hat{z}_t^i) \le c_{t+\omega}^i - (1-\zeta^i)c_t^i$$
$$\forall\, i\in\mathcal{Q},\ t=-\Omega^i+1,\ldots,T,\ \omega=1,\ldots,\Omega^i, \quad (25)$$

$$n_t^{ij},\ o_t^{ij},\ w_t^{ij},\ h_t^{ij},\ d_t^{ij} \ge 0$$
$$\forall\,(i,j)\in\mathcal{G},\ t=1,\ldots,T, \quad (26)$$

$$s_t^{ij} \ge 0 \quad \forall\,(i,j)\in\mathcal{G},\ i\neq 1,\ t=1,\ldots,T, \quad (27)$$

$$v_t^{ij,k},\ e_t^{ij,k} \ge 0$$
$$\forall\,(i,j)\in\mathcal{G},\ \forall k\in\mathcal{C},\ t=1,\ldots,T, \quad (28)$$

$$c_t^i \ge 0 \quad \forall\, i\in\mathcal{Q},\ t=1,\ldots,T, \quad (29)$$

$$x_{k,t}^i \ge 0 \quad \forall\, i\in\mathcal{Q},\ k=1,\ldots,R^i,\ t=1,\ldots,T, \quad (30)$$

$$y_{k,t}^i = 0 \text{ or } 1$$
$$\forall\, i\in\mathcal{Q},\ k=1,\ldots,R^i,\ t=1,\ldots,T, \quad (31)$$

$$z_t^i,\ \hat{z}_t^i = 0 \text{ or } 1 \quad \forall\, i\in\mathcal{Q},\ t=1,\ldots,T, \quad (32)$$

where w_0^{ij} is given $\forall\,(i,j)\in\mathcal{G}$, h_t^{ij} is given $\forall\,(i,j)\in\mathcal{G}$, $i=1$, $t=-\tau+1,\ldots,0$, s_t^{ij} is given $\forall\,(i,j)\in\mathcal{G}$, $i\neq 1$, $t=-\hat{\tau}+1,\ldots,0$, c_t^i is given $\forall\, i\in\mathcal{Q}$, $t=-\Omega^i+1,\ldots,0$, z_t^i is given $\forall\, i\in\mathcal{Q}$, $t=-\Omega^i+1,\ldots,0$, and $\hat{z}_t^i$ is given $\forall\, i\in\mathcal{Q}$, $t=-\Omega^i+1,\ldots,0$.

Acknowledgments

We thank the anonymous reviewers for their suggestions that helped us improve the paper. The first author also thanks Bill Stein for his many useful comments.

References

Aksin, O. Z., F. Vericourt, F. Karaesmen. 2004. Call center outsourcing contract design and choice. Working paper, Fuqua School of Business, Duke University, Durham, NC.

Amazon.com. 2003. 2002 Annual Report. Amazon.com, Seattle, WA.

Amazon.com. 2005. 2004 Annual Report. Amazon.com, Seattle, WA.

American Customer Satisfaction Index, The. www.theacsi.org.

Armony, M., C. Maglaras. 2004. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Oper. Res.* **52**(2) 271–292.

Chen, Bert P. K., S. G. Henderson. 2001. Two issues in setting call centre staffing levels. *Ann. Oper. Res.* **108**(1–4) 175–192.

Gans, N., Y-P. Zhou. 2002. Managing learning and turnover in employee staffing. *Oper. Res.* **50**(6) 991–1006.

Gans, N., Y-P. Zhou. 2004. Overflow routing for call-center outsourcing. Working paper, The Wharton School, University of Pennsylvania, Philadelphia, PA.

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

Harrison, J. M., A. Zeevi. 2005. A method for staffing large call centers based on stochastic fluid models. *Manufacturing Service Oper. Management* **7**(1) 20–36.

Reid, R. H. 1997. *Architects of the Web*. John Wiley and Sons, New York.

Whitt, W. 1999. Using different response-time requirements to smooth time-varying demand for service. *Oper. Res. Lett.* **24**(1–2) 1–10.

Raghu Sethuraman, Manager of Worldwide Customer Service Network, Amazon.com Inc., 605 5th Ave. S, Seattle, WA 98104, writes: "I am writing this letter to confirm that the planning and optimization model presented in this paper has been implemented at Amazon.com. The model has enabled us to optimize staffing and contact allocation across all global sites and media types to ensure worldclass timely experience for our customers.

"Furthermore I can tell you that, after implementation, it recently passed its toughest test with flying colors: our company's holiday season and high service level goals. The model allows more flexibility for business rules and "what-if" sensitivity analysis, helping us make high-level strategic decisions to optimize our global customer service network. In summary, the model has tremendously improved our planning process and is now one of our key decision support tools."