New Project Staffing for Outsourced Call Centers with Global Service Level Agreements

Working Paper

Thomas R. Robbins • Terry P. Harrison
Department of Supply Chain and Information Systems , Smeal College of Business, Pennsylvania State University, University Park, PA

We consider the issue of new project staffing in an outsourced call center required to meet a monthly Service Level Agreement. We present empirical analysis generated during a field study with a provider of outsourced call center services to illustrate the unique issues related to staffing a new project. Our work shows that during the start up phase of a project, agents experience significant improvements in productivity which reduce the staffing requirements over time. We also find that turnover, which is typically high in a call center environment, may be even higher for a project launch dominated by newly hired agents. These factors interact with the uncertainty of call volumes and talk time to create a difficult hiring challenge. We develop a model that finds the level of hiring with the lowest total expected cost of operation while meeting service level commitments.

Working Paper

1 Introduction

Call centers are a critical component of the worldwide services infrastructure and are often tightly linked with other large scale services. Many outsourcing arrangements, for example, contain some level of call center support, often delivered from offshore locations. A call center is a facility designed to support the delivery of some interactive service via telephone communications; typically an office space with multiple workstations manned by agents who place and receive calls (Gans, Koole *et al.* 2003). Call centers are a large and growing component of the U.S. and world economy and by 2008 the United States will employ an estimated 2.1 million call center agents (Aksin, Armony *et al.* 2007). Large scale call centers are technically and managerially sophisticated operations and have been the subject of substantial academic research. Call center applications include telemarketing, customer service, help desk support, and emergency dispatch. Staffing is a critical issue in call center management as direct labor costs often account for 60-80% of the total operating budget of a call center (Aksin, Armony *et al.* 2007).

While there has been significant research on issues related to scheduling of call center agents, less attention has been focused on the appropriate number of agents to hire. In particular we are interested in the number of agents to hire prior to the launch of a new project. Our research is based on a field study with a provider of outsourced technical support. The new project hiring decision is especially troublesome in this environment for several reasons. Standard staffing model require estimates of call volumes, arrival patterns, and talk times, all of which are uncertain prior to launch. In addition new agents are likely to experience significant learning effects during their first few weeks on the job dramatically altering the required staffing profile. Additionally, call center attrition is high in general, but we found it is especially high among new hires.

Working Paper

2 Literature

Call centers have been the focus of significant academic research. A detailed review of the call center oriented literature is provided in (Gans, Koole *et al.* 2003). More recent work is summarized in (Aksin, Armony *et al.* 2007). Empirical analysis of call center data is provided in (Brown, Gans *et al.* 2005).

A large body of literature addresses the issue of call center staffing requirements and scheduling. (Halfin and Whitt 1981) develop a popular staffing heuristic known as the s*quare root safety staffing rule*, for systems with non abandonment. This heuristic is extended to systems with abandonment in (Garnett, Mandelbaum *et al.* 2002). These models use queuing theory to find the number of agents required to achieve some performance measure under steady state conditions. Extension to the server sizing problem are addressed in a number of papers (Borst, Mandelbaum *et al.* 2004; Whitt 2005; Whitt 2006a; Whitt 2006b). In systems with time varying arrivals a piecewise approach is often used to develop a required staffing level of each 15 or 20 minute period (Green, Kolesar *et al.* 2001; Green, Kolesar *et al.* 2005).

The general problem of stetting aggregate staffing levels is also widely studied in the operations research and is often referred to as the manpower planning problem. The manpower planning literature is in general divided into two complementary approaches. In one approach the evolution of the workforce is modeled as a stochastic process that evolves over time (Bartholomew and Forbes 1979; Bartholomew 1982). This approach explicitly models the stochastic nature of hiring, turnover, skills acquisition and demand. An alternate approach is based on an optimization paradigm in which the objective is to make a set of control decisions over time that optimize some measure of system performance, such as total cost, deviation from staffing plan, or expected profit (Holt, Modigliani *et al.* 1960). More recent work has attempted to integrate uncertainty and optimization.

Stochastic models of manpower systems focus on the uncertainty inherent in the system. Bartholomew provides a general review of the application of stochastic modeling to social systems in (Bartholomew 1982), and a more specific application of these principals to the manpower planning problem in (Bartholomew and Forbes 1979). A basic model incorporates a

Working Paper

number of discrete manpower grades and models the system as a Discrete Time Markov Chain (DTMC). Many papers have built on this simple Markov model to analyze manpower systems, introducing various control objectives into the process. Grinold develops a stochastic model motivated by the demand for naval aviators (Grinold 1976). The environment evolves as a Markov process and the demand for aviators therefore has a definable probability distribution. The control objective is then to find the optimal accession policy that governs new entrants into the system, and the continuation policy that governs movement through the system. In (Anderson 2001) demand is driven by a continuous nonstationary seasonal process meant to approximate a business cycle. The model explicitly assumes employees progress at differential rates, unlike the deterministic rates in Grinold. The objective trades off the discounted cost of meeting demand requirements with a penalty term for abrupt changes in the employee stock. Based on this objective Anderson uses a dynamic programming approach to define optimal control policies. A similar model that focuses on cohort analysis is developed in (Gaimon and Thompson 1984). The Gaimon and Thompson model postulates that the effectiveness of an individual can be defined exogenously as a function of organizational age and grade. Effectiveness may be defined to increase throughout the individual's career, or it may be specified to peak at some organizational age and begin to decline in an environment with rapid technological change.

A number of other papers examine the strategic staffing problem using a stochastic setting. (Gans and Zhou 2002) develop a model with learning curve and stochastic turnover issues. (Gaimon 1997) examines manpower planning in the context of knowledge intensive IT workers. (Bordoloi and Matsuo 2001) also examine a knowledge intensive work environment with stochastic turnover.

An alternative approach to manpower planning is based on optimization theory. The theoretical foundations of the optimization approach to manpower were developed in Holt *et al.* (Holt, Modigliani *et al.* 1960) Holt evaluates manpower as a component of the productive capacity of a manufacturing enterprise, evaluating staffing decisions in an aggregate planning context. Holt develops a quadratic cost model that includes both the costs of maintaining a workforce and the cost of changing the workforce. Holt's quadratic cost model is converted to a linear cost model in (Hanssmann and Hess 1960) and solved via linear programming. The Holt model is also extended in (Ebert 1976) with the inclusion of time varying productivity. Ebert uses the quadratic

Working Paper

cost model directly from Holt, but allows productivity to vary over time as learning takes place. Ebert solves this non-linear program using a search heuristic. An alternative formulation that also includes learning curve effects is presented in (Harrison and Ketz 1989). This model is non-linear but is solved via successive linear programming.

The two approaches to manpower planning outlined above emphasize different aspects of the system and as such have different applications. The stochastic models are generally high level abstractions useful for identifying system phenomenon or developing general policies. The optimization models on the other hand are often crafted to identify specific management actions but tend to ignore the variability in the system. Variable parameters are typically modeled with their expected values yielding what is known as the *mean value problem* which may result in solutions that are far from optimal (Birge and Louveaux 1997). Modeling variability in optimization problems is likely to yield solutions that are superior to the deterministic counterparts, but solutions to these stochastic programs are difficult to find.

3 Empirical Analysis

3.1 Background

This work is directly related to a research project with a provider of outsourced technical support delivered via globally distributed call centers. The company provides both tier 1 (help desk) and tier 2 (desk-side) support. The bulk of their business, and the focus of this research, is on the inbound call center operation. This operation involves providing help desk support to large corporate and government entities. While the scope of services varies from account to account, many accounts are 24 x 7 support and virtually all accounts are subject to some form of Service Level Agreement (SLA). There are multiple types of SLAs, but the most common specifies a minimum level of the Telephone Service Factor (TSF). A TSF SLA specifies the proportion of calls that must be answered within a specified time. For example, an 80/120 SLA specifies that 80% of calls must be answered within 120 seconds. A very important point is that the service level applies to an extended period, typically a month. The SLA does not define requirements for a day or an hour. So the desk is often staffed so that at some times the service level is underachieved, sometimes overachieved, and is on target for the entire month.

Working Paper

3.2 Call Volume

The key challenge involved with staffing this call center is meeting a fixed SLA with a variable and uncertain arrival rate pattern. The number of calls presented in any ½ hour period is highly variable with multiple sources of uncertainty. In the following figure we see daily call volume for a typical project shown over a three month period.

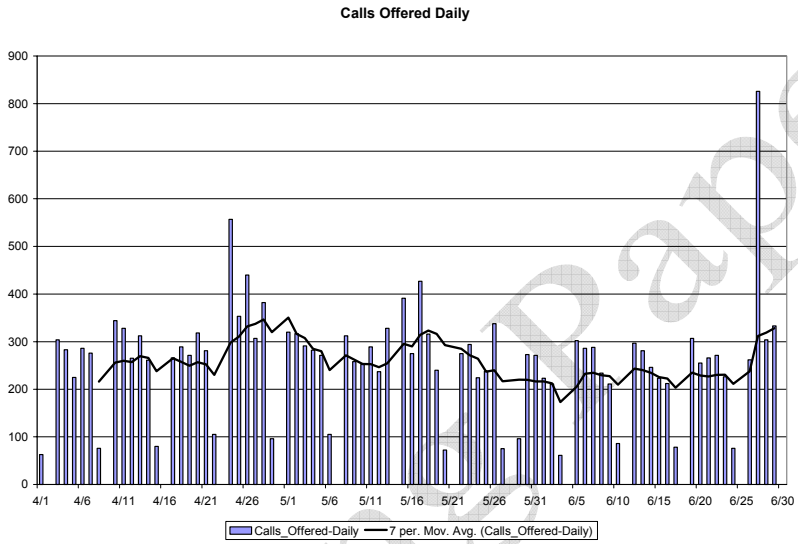

Figure 3-1 Sample Daily Arrival Pattern

This graph shows strong "seasonal" variation over the course of a week. Mondays tend to be the highest volume days with volumes dropping off over the course of the week. Call volume on Saturday is a small fraction of the weekday volume, and this particular desk is closed on Sundays. The graph also reveals significant stochastic variability. Tuesdays are, for example, often higher volume than Wednesdays but this is not always the case. During the weeks of 4/26 and 5/16 we see larger volumes on Wednesday than Tuesday. We also see the issue of unanticipated spikes in demand, often referred to as *significant events*. This is an extremely common event in support desk operations. A downed server, for example, will generate a large call volume. While some contracts provide SLA relief in the case of significant events, in general the desk must meet the SLA even when significant events occur. The large volume of calls during a significant event not only results in poor performance, but also represents a large portion of total calls received making it more difficult to achieve the overall SLA.

Working Paper

In addition to day of week seasonality these call centers also experience very significant time of day seasonality. Volume tends to dip down around the lunch break, but a second peak occurs in the afternoon; though the afternoon peak is typically lower volume than the morning peak. While this basic arrival pattern exists on most business days, there is significant stochastic variability in the call pattern from day to day. The following graph shows call volume over an eight week period for a particular project. The inner region represents the minimum volume presented in each period, while the overall envelope is the maximum volume presented in each period. The outer region then represents the variability over this eight week period.

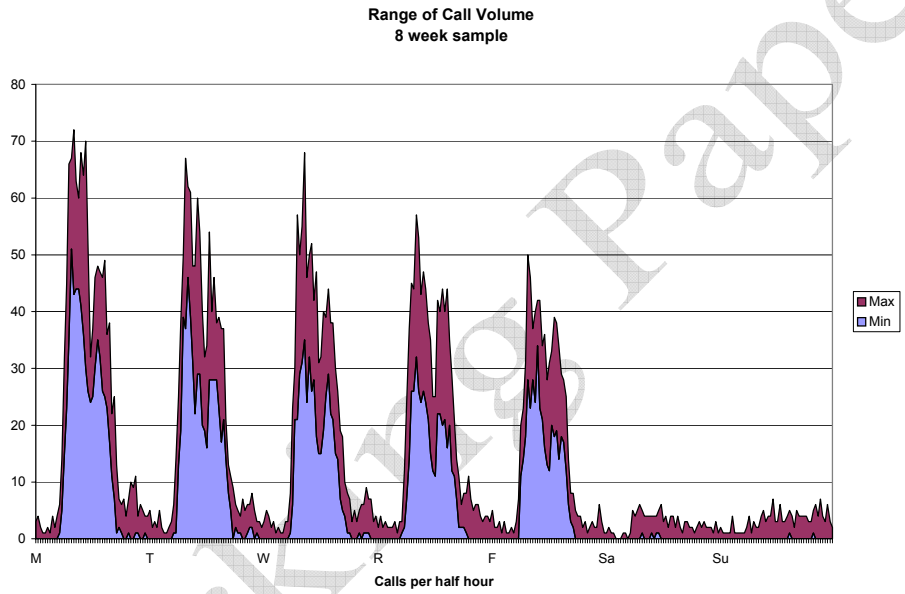

Figure 3-2 Range of Call Volume

This particular desk operates 24x7 and we see that the volume during the overnight hours is quite low. Volume ramps up sharply in the morning with a major surge of calls between 7 and 11 AM. Volume tends to dip down around the lunch break, but a second peak occurs in the afternoon,; though the afternoon peak is typically lower volume than the morning peak. The staffing challenge in this call center is to find a minimal cost staffing plan that achieves a global service level target with a high probability. The schedule must obviously be locked in before arrival rate uncertainty is revealed.

Working Paper

3.3 Model Projects

Throughout this analysis we evaluate the models using three test problems based on specific outsourcing projects. Project J is a corporate help desk for a large industrial company averaging about 750 calls a day where the volatility of call volume is relatively low. Project S is help desk that provides support to workers in a large national retail chain. Call volume on this desk is about 2,000 calls a day. Because this desk supports users in retail stores, as opposed to a corporate office, the daily seasonality of call volume is quite different. This company is making major changes in its IT infrastructure and as such call volume is very volatile and difficult to forecast. Project O is a help desk that provides support to corporate and retail site users of another retail chain. This is a smaller desk with about 500 calls a day, where call volume is fairly volatile and shocks are relatively common.

3.4 New Project Launch

One of the key challenges in this business model is the new project launch process. A significant problem is determining the appropriate number of agents to hire and train. Because of the substantial (project specific) training investment required for new hires, management is reluctant to hire extra workers. Standard operating procedures call for hiring to the projected steady state level based on expected call volumes. As in the case of the short term scheduling problem the decision is complicated by uncertainty. Attrition levels are again uncertain, as is demand. The level of demand uncertainty is very high prior to the *go live* event because accurate counts of current call volumes are often extremely difficult to obtain. Business process changes involved with the transition, such as call center consolidation, changes in hours of operation, or changes in the type of support provided, often make previous data of limited value, even if known. Another major complicating factor is the evolving level of productivity due to learning curve effects. Talk times tend to decline during a launch as agents become more familiar with the environment and the project knowledge base becomes better populated. Variability in talk time is subject to institutional learning curve effects, individual learning curve effects, and stochastic variability. A final complicating factor is the lead time required to add new capacity. Recruiting new hires can take time, but the biggest factor is training time. Since agents must provide detailed technical support they require extensive training before they can be deployed on the help desk. Training times are project dependent and vary from two weeks to three months.

Working Paper

The following graph shows the average talk time over the first 3 months of a major launch that occurred in 2005.

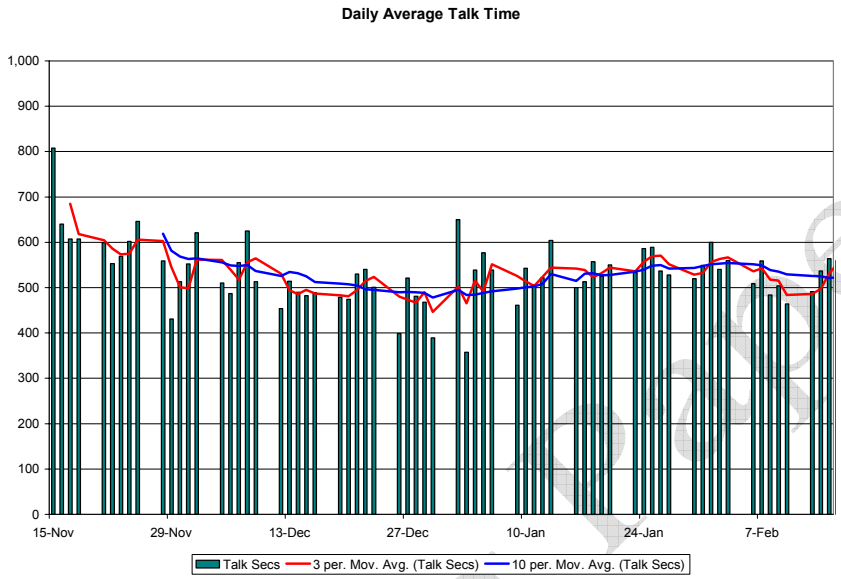

Figure 3-3 Talk Time Evolution during Startup

This graph reveals a general decline in talk time (increase in productivity) during the first several weeks of the launch, followed by a leveling off and a slight increase in the third month. The increase in talk time in January is due, at least in part, to the addition of new hires made to replace resigning workers[1]. This particular project involved a phased deployment where large groups of agents were added at various time through the extended launch period. If we plot the average talk time over a longer period, we can clearly see the impact of new agents and the learning curve effect.

Working Paper

[1] Six new hires were made in December for a project with a total headcount at that time of about 35. These individuals began taking calls in early January.

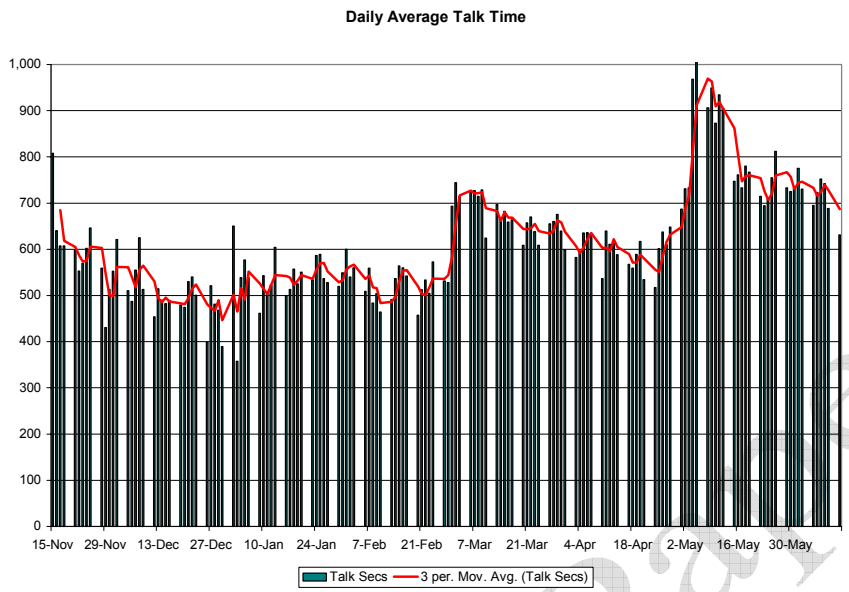

Figure 3-4 Talk Time Shocks

The startup problem has is illustrated by a recent launch of the company's single largest customer. Based on the scope of this launch the decision was made to conduct a phased launch effort, adding new users every few weeks over an extended period. Unfortunately this created multiple forecasting challenges. As the following graph shows, the inability to ramp up capacity along with demand led to extremely poor quality of service over an extended period of time.

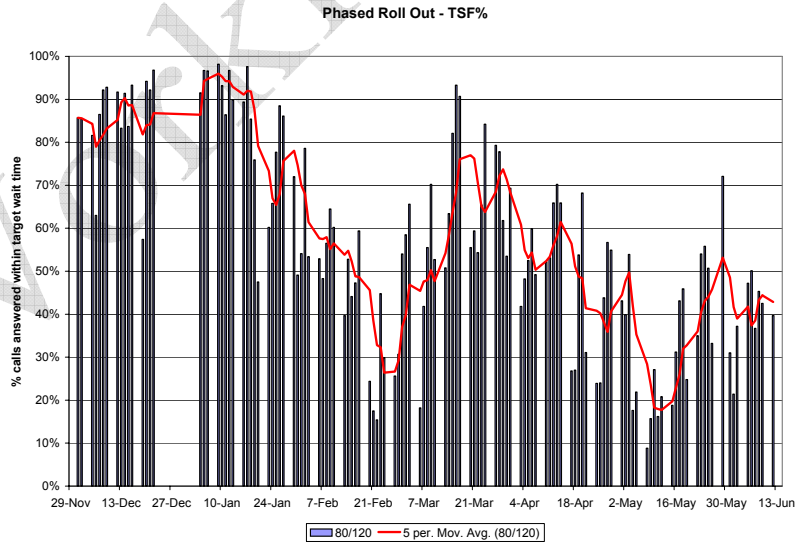

Figure 3-5 Service Levels During rollout

Working Paper

The challenge of forecasting the new demand with each subsequent roll out, coupled with learning curve issues as new agents were added with each rollout, created acute mismatches in capacity and demand. The service level target for this project was 80%, but as the graphic shows actual service levels were well below this target for extended periods with several periods of extremely poor performance[2].

3.4.1 A Statistical Model of Learning Curve Effects

To develop a statistical model of learning curve effects we collected individual agent scorecards from a sample project. These scorecards are prepared for each front line worker each month and assess the worker on a number of key operating metrics including talk time, "wrap" time, first tier closure rate, inside call volume, and monitoring scores[3]. The data set included scorecards for 2004 through November 2006. The data was pulled from the individual scorecards and arranged by month of service. We included only agents where we had at least three months of contiguous data. The resulting data set had measure for 53 agents with length of service ranging from three months to 19 months.

As a proxy for agent productivity we examined talk time, first tier closure rate, and inside call volume. Our hypothesis is that as agents learn they will resolve more problems, in less time, with less help from other agents.

Working Paper

[2] The problems associated with this launch have provided the motivation for the company to rethink its launch process.

[3] Talk time is the average time the agent spends on the phone per call, while wrap time is the post call time spent processing data from the call. First tier closure rate is the proportion of calls resolved directly by the agents, as opposed to escalation to a tier 2 agent. Inside call volume are calls placed by the agent to other agents seeking help to resolve difficult problems. Monitoring score are the scores given by QA personnel who anonymously monitor a portion of calls and grade agents against a broad list of subjective performance measures.

The following graph shows the reported monthly average talk time for each agent as a function of their month of service.

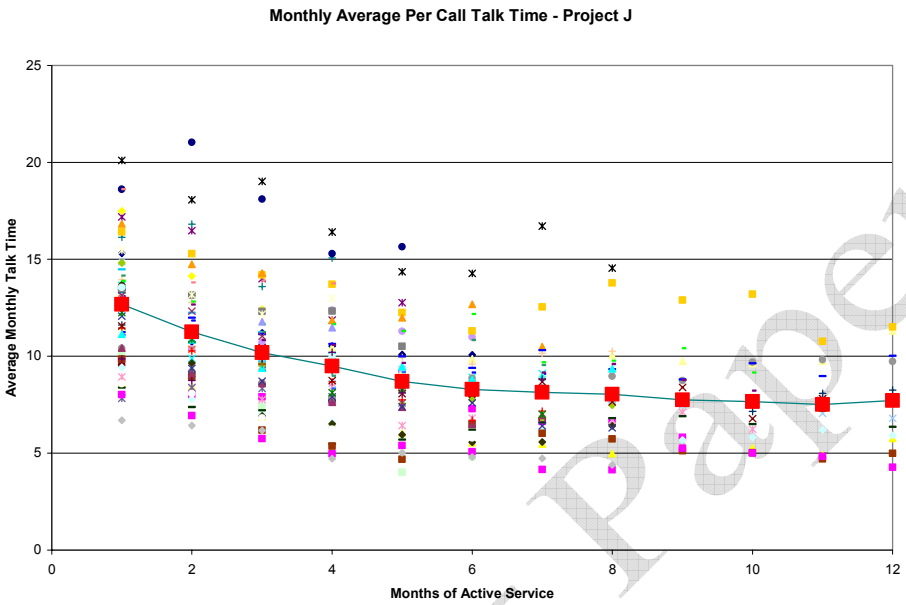

Figure 3-6 Monthly Average Talk Time

Each point on this graph represents an individual agent employed for a month. The data reveals a general decline in average talk time over the first several months of service as expected. Talk time declines from an average of 12.7 mins in the first month, to 8.3 mins in the 6th month. The standard deviation of talk time ranges between 3.0 and 2.5 over this period. However, when we examined first tier closure rate and inside call volume the picture became a little more complicated.

Working Paper

The following graph shows that First Tier Closure rate decreases over the first few months of service.

Figure 3-7 Monthly Average First Tier Closure Rate

This unexpected result, the decrease in First Tier Closure rate, is partially explained by the Inside Call Volume statistic. Shown in the following graph:

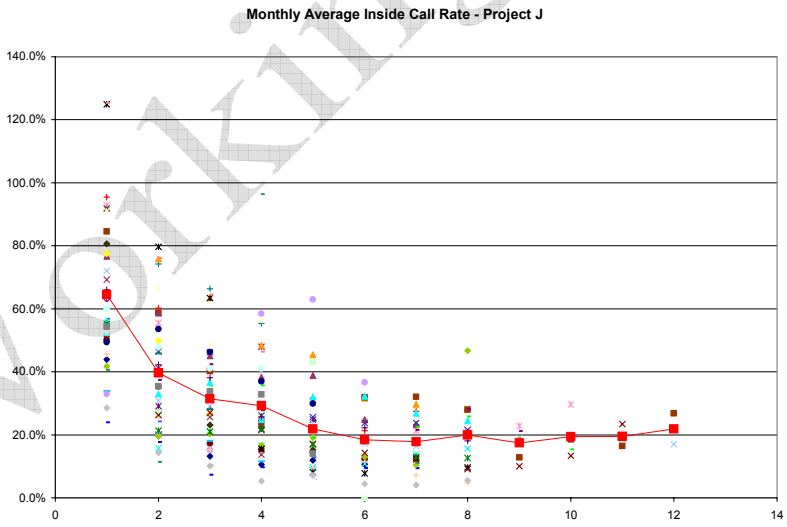

Figure 3-8 Monthly Average Inside Call Rate

During the first few months when agent experience is low they draw upon more experienced agents to help them solve a large portion of their calls and because of this they are able to close a

Working Paper

relatively high percentage of calls without escalation.[4] So as agents progress they are able to resolve calls faster with less outside support. As a first preliminary measure of productivity we use the talk time measure as a surrogate of productivity.

We first compare the average talk time for all agents in their n^{th} month of service to the average of the $n\text{-}1^{th}$ month of service. Using a standard T-test we find the reduction is statistically significant at the 0.1 level through the first five months of service. Average talk time continues to decline through the first 11 months of service, but the month to month changes are statistically significant only at the 0.5 level. If we evaluate the two month improvement the improvement is significant at the 0.01 level through the sixth month.

This analysis shows us that average talk time for more experienced agents is lower than talk time for less experienced agents; but does not give us conclusive evidence as to why talk time decreases. Two explanations are possible; first agents learn and become more productive and second slower agents are removed from the system. To verify that individual agents become more productive we examine the one month difference for individual agents.

The data is summarized in the following graph

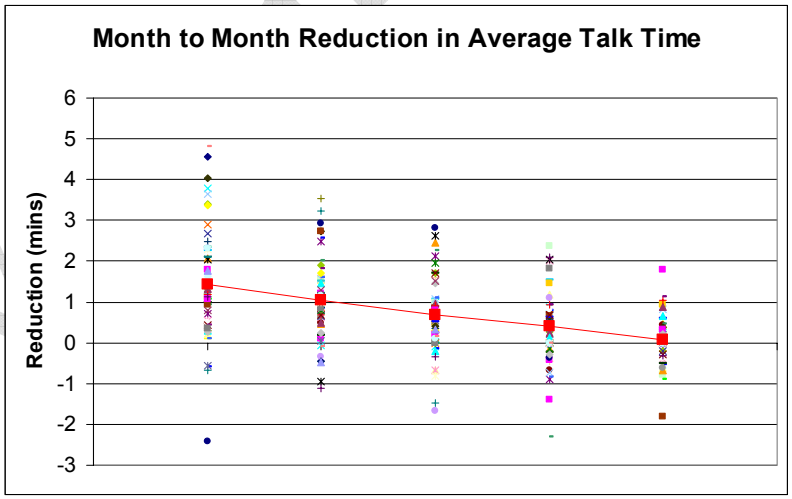

Figure 3-9 Monthly Reduction in Talk Time

[4] Management policy prevents agents on this project from escalating tickets without concurrence of another more experienced agent until they earn their escalation rights, typically sometime in the 2nd or 3rd month of service.

The graph shows positive reductions (improvements) in average talk time through the first five months. The improvement in month six is not statistically significant.

All of this data suggests that a standard learning curve model is appropriate. Given the data we have, we developed a learning curve model based on months of service rather than cumulative call volume. We fit a model of the form

$$T_t = T_0\left(1+e^{-\alpha n}\right), n \geq 1 \tag{3.1}$$

where T_t is the average talk time in period t, T_0 is the average talk time for an experienced agent (6 months +), n is the month of service, and α is the learning curve rate parameter. We fit a curve to the total average talk time measure[5] and calculated an α value of 0.4605.

The curve gives a reasonably good fit as the following graph illustrates:

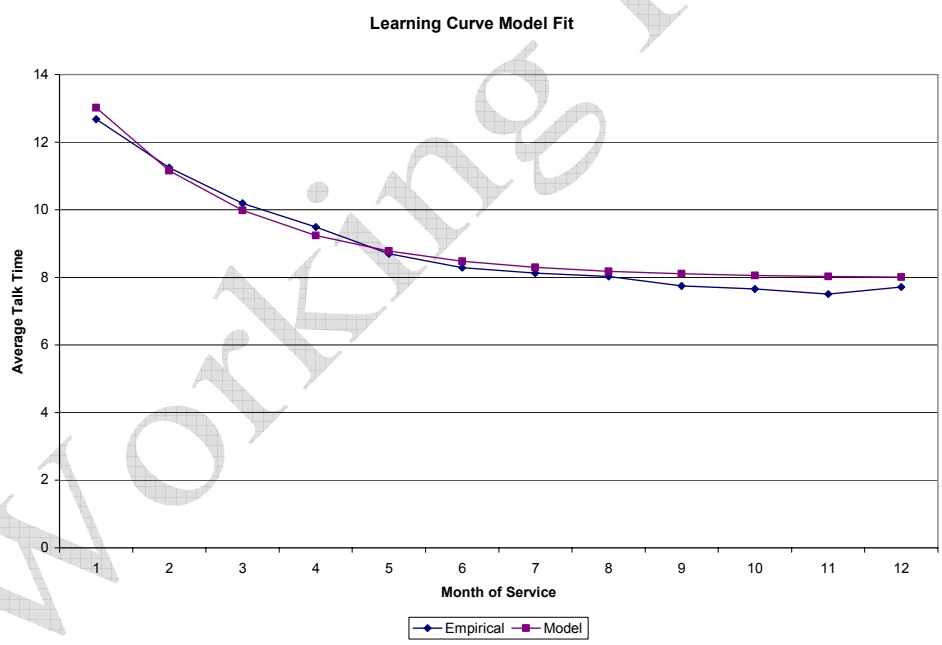

Figure 3-10 Learning Curve Model Fit

We were able to collect similar data for a second project and performed a similar analysis. We obtained slightly different, but similar results. On this project the improvement is statistically

[5] We fit the curve using Excel's solver to find the value at minimizes the sum of squared errors.

Working Paper

significant at the .2 level only through the 4th month of service, indicating a somewhat faster learning process. The corresponding alpha value for this project is .703.

The learning curve model expressed in (3.1) is quite flexible and can be used to represent a wide range of learning curve effects as the following graph illustrates.

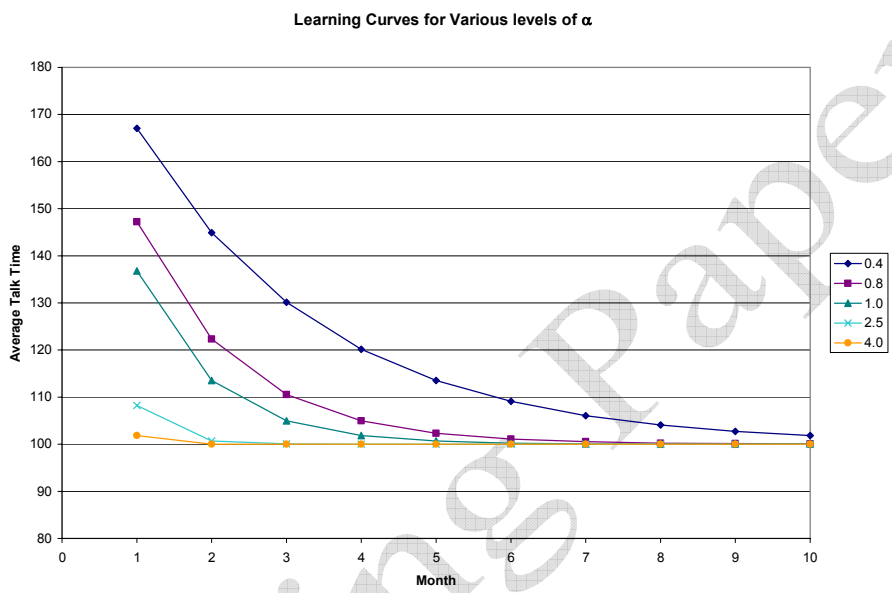

Figure 3-11 Family of Learning Curves

The curve is valid for any $\alpha \in (0, \infty)$. For large values of α the curve is relatively flat; new agents perform nearly as well as experienced agents and any gap is quickly closed. As the value of α decreases, the curve becomes steeper and the learning effect is more pronounced. The limitation of this functional form is that α must be strictly greater then zero and the initial effort can be no more than twice the long run effort. This limitation is easily overcome by adding a second scaling parameter to (3.1), but that is not necessary to fit our data.

In Figure 3-11 we show the impact of learning on talk time. An alternative way to think of learning is the impact on relative productivity or capacity. As learning occurs agents become more productive and able to handle more calls and increase their effective capacity. Based on equation (3.1) average talk time will settle in at the value of T_0. If we equate a relative

Working Paper

productivity level of one with a talk time of T_0 then we can define the relative productivity index ρ as the ratio of the average talk time of the inexperience agent with that of the experience agent $\left(T_0 / T_t\right)$ or

$$\rho_t = \frac{1}{1+e^{-\alpha n}} \tag{3.2}$$

The relative productivity will evolve as shown in the following graph

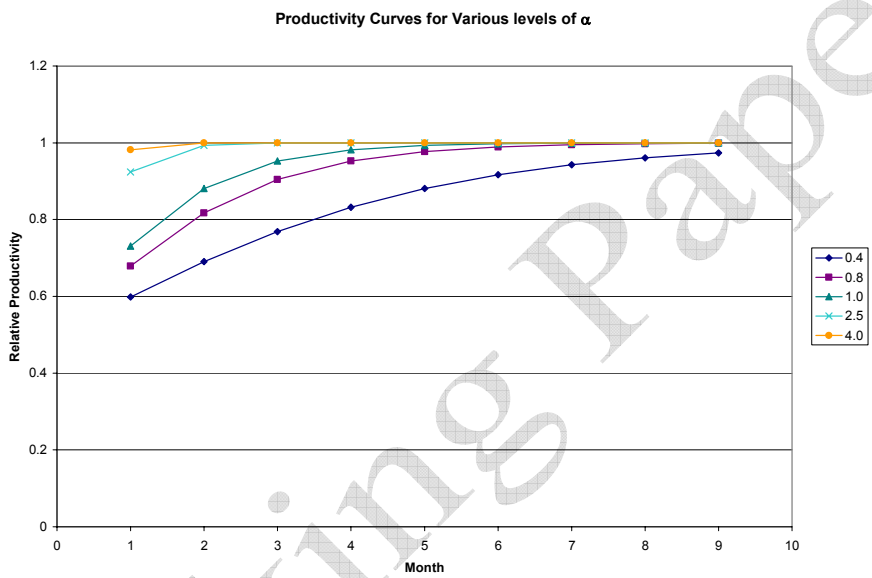

Figure 3-12 Relative Productivity Curves

This graph shows that with an α value of .8 a new agent can handle approximately 68% as many calls as a fully experienced agent, based simply on talk time. A more complete analysis would take a broader measure of agent productivity then just talk time. Based on the inside call volume statistic shown in Figure 2-16 , new agents place a high burden on experienced agents by asking them questions. While we have a measure of the number of inside calls made we have no data on the duration of calls. Qualitatively, the data indicates that the burden placed on experienced agents decreases with time and we can conclude that the productivity curves of Figure 3-12 moderately understate the productivity improvement associated with learning.

Working Paper

3.5 Turnover Issues

As is often the case in call center environments, turnover in this company is a significant issue. The following graph shows the month by month annualized turnover rate over an approximately 28 month period.

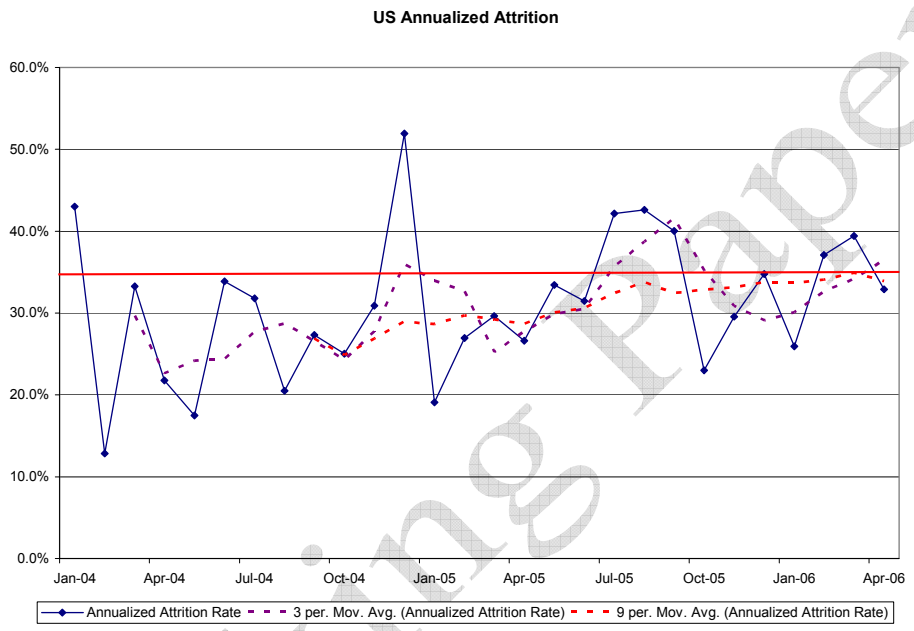

Figure 3-13 Annualized Attrition Rates by Month

We see that turnover varies significantly from month to month, with the 9 month moving average in the range of 25-35% per year. A widely used model for employee attrition estimates the attrition probability as a function of length of service. (See for example (Bartholomew and Forbes 1979)).

We collected detailed termination data on the 1,400 terminations (voluntary and involuntary) that occurred between January 2001, and May 2006. We used this data to estimate a hazard rate function for the probability of quitting.

Working Paper

The data was fit to a Weibull distribution with shape parameter equal to 0.918 and scale parameter equal to 0.0309. The hazard rate function derived from that distribution is shown in the following graph:

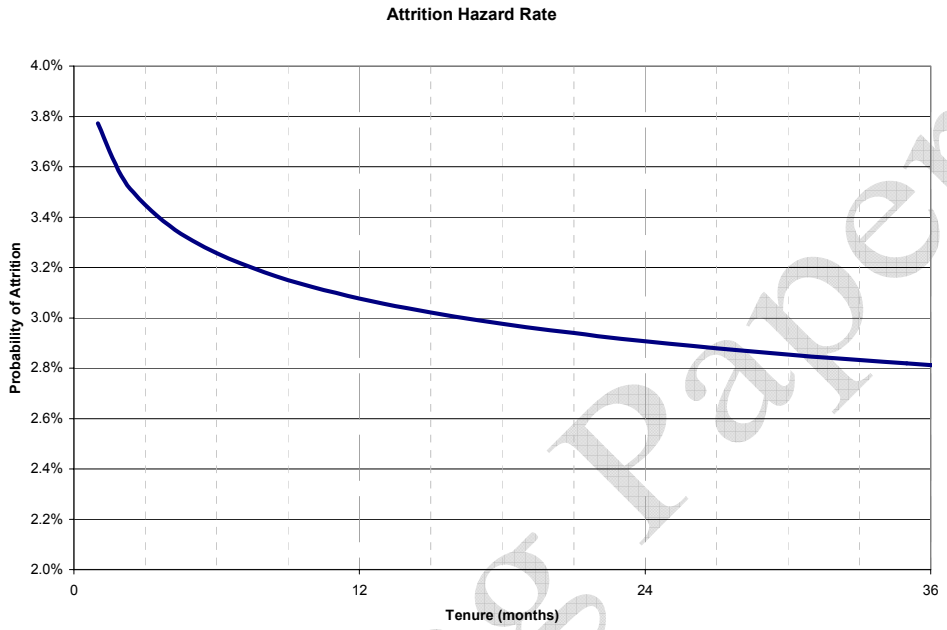

Figure 3-14 Attrition Hazard Rate

The analysis reveals a decreasing failure rate function; that is the probability of quitting declines with length of service[6]. This is consistent with summary data that shows that over this period approximately 15% of new hires quit within the first three month of employment. We termed this the *washout rate*. The observations that are relevant for the analysis in this paper are the following. First, attrition rates of 3.0%-3.5% per month are realistic, and second the probability of quitting declines with length of service. These observations, and parameter values, will become important when we analyze new project start ups. The high attrition rate, especially among new hires, implies that we must consider attrition when staff planning for a new project.[7]

[6] A Weibull distribution will have a decreasing failure rate if the shape parameter α is less than 1 and an increasing failure rate if the shape parameter is greater than 1. If the shape parameter is equal to 1 the Weibull is equal to the exponential distribution and the failure rate is constant.

[7] Interestingly, the company currently does not factor attrition levels into new project start ups. Hiring is currently capped at the number of agents specified in the long run cost model for the project.

Working Paper

4 The Hiring Model

4.1 Overview

The objective of this model is to address capacity management in the 0-3 month start up phase of a new project launch. Since the services provided are highly technical a significant training investment is required for new hires. Significant learning occurs during the start up phase and productivity increases rapidly. The outsourcing contract typically specifies a global service level agreement, but the SLA is often not strictly enforced until the third month of the launch. The management problem we address is the development of a staffing plan for a new project in the face of uncertain demand and productivity. Over hiring results in training expenses that can not be recouped on other projects, while under hiring results in poor customer service and may make it impossible to achieve the service level commitment. We seek to develop a model that finds the optimal level of hiring; that is the level of hiring with the lowest total expected cost of operation.

The level of pre-launch hiring is a critical decision in the new project launch process. The following graph outlines the basic timeline of the process.

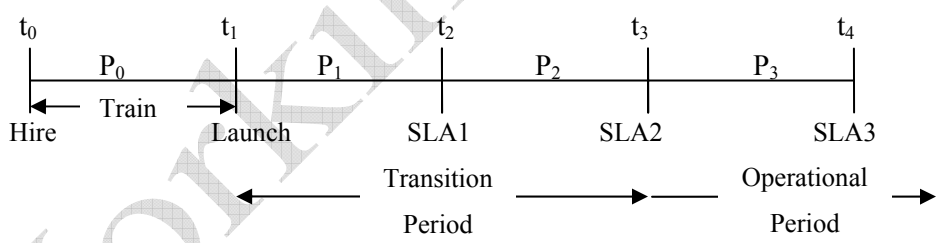

Figure 4-1 Project Launch Timeline

At time t_0 a set of resources are hired and entered into a training program. The company makes a hiring decision with an uncertain call volume and talk time. Training occurs in period zero and at time t_1 the project launches and the uncertainty in average call volume and initial talk time are revealed. In periods one through three the company provides service, measuring service levels which are reported at times t_2 through t_4. Periods one and two are considered transition periods, significant learning occurs during this period and service level agreements are often not strictly

Working Paper

enforced. Period three is typically the first month in which service levels are contractually enforced. Turnover and learning occur throughout the launch process.

Learning occurs through individual and institutional processes. Individual learning occurs as agents become more familiar with the systems and become more productive in solving customer problems. Institutional learning occurs as the project knowledge base, the repository of known problems and solutions, is enhanced. Both types of learning result in increased agent productivity. Individual learning is lost when turnover occurs while institutional learning remains in the knowledge base. In this model we assume turnover occurs only at discrete points in time, t_1 through t_4. We assume hiring is instantaneous, but that training delays the deployment of replacement hires by one period. Average volume is revealed at time t_1, but call volumes are subject to stochastic variability in all periods.

This model is developed as a multistage decision problem. The initial decision on hiring occurs at time t_0. This decision occurs before average call volumes or learning curve effects are revealed. Recourse decisions are made at times t_1, t_2, and t_3 that include additional hiring and/or termination. The management objective is to minimize the overall expected cost of staffing such that the service level agreement is satisfied by period three.

Working Paper

4.2 Model Formulation

We formulate the model as a multi stage stochastic program with the following definitions:

Sets

T: time periods
K: scenarios
H: linear segments of service level curve

Deterministic Parameters

w: wage rate
h_t: hiring and training cost
f: termination cost
g: SLA target
q_t: SLA penalty rate in period t
μ: minimum expected SLA
r: expected SLA shortfall penalty rate

Stochastic Parameters

m_{tkh}: SLA slope in period t of scenario k
b_{tkh}: SLA intercept in period t of scenario k
a_{tk}: attrition rate in period t of scenario k
γ_{tk}: institutional productivity in period t of scenario k
ρ_{jtk}: individual productivity in period t of scenario k for resources hired in period j

Variables

X_{tk}: resources in period t of scenario k
H_{tk}: hires in period t of scenario k
Y_{tk}: SLA shortfall in period t of scenario k
F_{tk}: terminations in period t of scenario k
C_{tk}: effective capacity in period t of scenario k
S_{tk}: SLA in period t of scenario k
E_t: Expected SLA penalty

Probabilities

p_k: probability of scenario k

$$Min \sum_{t \in T} \sum_{k \in K} p_k (wX_{tk} + h_t H_{gk} + fF_{tk} + q_t Y_{tk} + rE_t) \tag{4.1}$$

subject to

$$X_{tk} = X_{t-1,k} + H_{tk} - F_{tk} - a_{tk} X_{t-1,k} \qquad \forall t \in T, k \in K \tag{4.2}$$

$$C_{tk} = \gamma_{tk} \left(\sum_{j \in T, j<t} \rho_{jtk} \left(H_{jk} - F_{jk} - a_{tk} X_{t-1,k} \right) \right) \qquad \forall t \in T, k \in K \tag{4.3}$$

$$S_{tk} \le m_{tkh} C_{tkh} + b_{tkh} \qquad \forall t \in T, k \in K, h \in H \tag{4.4}$$

$$Y_{tk} \ge g - S_{tk} \qquad \forall t \in T, k \in K \tag{4.5}$$

$$E_t \ge u - \sum_{k \in K} p_k Y_{tk} \qquad \forall t \in T \tag{4.6}$$

$$X_{tk}, H_{tk}, Y_{tk}, F_{tk}, C_{tk}, S_{tk} \ge 0 \qquad \forall t \in T, k \in K \tag{4.7}$$

$$X_{0k} \in \mathbb{Z} \qquad \forall k \in K \tag{4.8}$$

Working Paper

The objective function (4.1) seeks to minimize the expected cost of staffing plus the penalty cost associated with failing to meet the SLA target; a penalty is applied for any scenario that does not achieve the period's SLA target. A second, large penalty is assessed of the expected service level in any period it is below some minimal threshold. This condition ensures that some minimal service level target is enforced based on expected volume. Constraint (4.2) is the staff balance constraint; it defines the staff in a period to equal the prior period staff plus new hires, less attrition and terminations. Constraint (4.3) defines the effective capacity of the current staff. For each hiring cohort the factor β_{thk} specifies the individual productivity component. Effective capacity is further adjusted by the institutional capacity factor α_{tk}. Constraint (4.4) defines the SLA achieved in period t based on the stochastic demand. The model is formulated so that demand is expressed in terms of the slope and intercept of the linear approximation of the TSF curve. Constraint (4.5) defines the SLA shortfall, the degree to which the realized SLA is below the target level g. Constraint (4.6) calculates the expected SLA shortfall, the degree to which the expected service level falls short of the minimum target. Constraint (4.7) defines non-negativity and conditions and constraint (4.8) forces the period 0 hiring decisions to be integral valued.

The most significant uncertainty in demand is in the first period where the overall level of demand is uncertain. After the general level of demand is revealed, period to period volume varies stochastically.

4.2.1 Detailed Decision Process Timing

The model (4.1) - (4.6) implements the decision process outlined in Figure 4-1. To further clarify how this process works, the following steps present the process in more detail.

1. At time t_0 the firm hires an initial group of agents. Those agents are trained during period P_0. During this period the agents are paid a salary and the firm makes an additional investment in training.
2. At time t_1 the project goes live and begins accepting calls. Calls are received throughout period P_1 and overall call volumes and call patterns are revealed. Throughout period P_1 agents may resign reducing the capacity of the project team.

Working Paper

3. At time t_2 the first period SLA is calculated and any shortfall penalty is assessed. At this time the firm may choose to hire additional agents or terminate existing agents. The firm incurs a severance cost for all terminated agents.
4. Newly hired agents are trained during period P_2 and are unavailable to take calls. A training cost is incurred for these agents and they are paid a salary. Call volume during period P_2 is handled by the original set of agents who are now more productive due to learning.
5. At time t_3 the second period SLA is calculated and any shortfall penalty is assessed. At this time the firm may again make a hire/termination decision.
6. During period P_3 agents hired at time t_3 are trained and paid a salary. Call volume is handled by the remaining agents hired at times t_0 through t_2.
7. At time t_4 the third period SLA is calculated and any shortfall penalty is assessed. At this time the firm may again make a hire/termination decision.

Working Paper

The detailed decision process is illustrated in the following diagram:

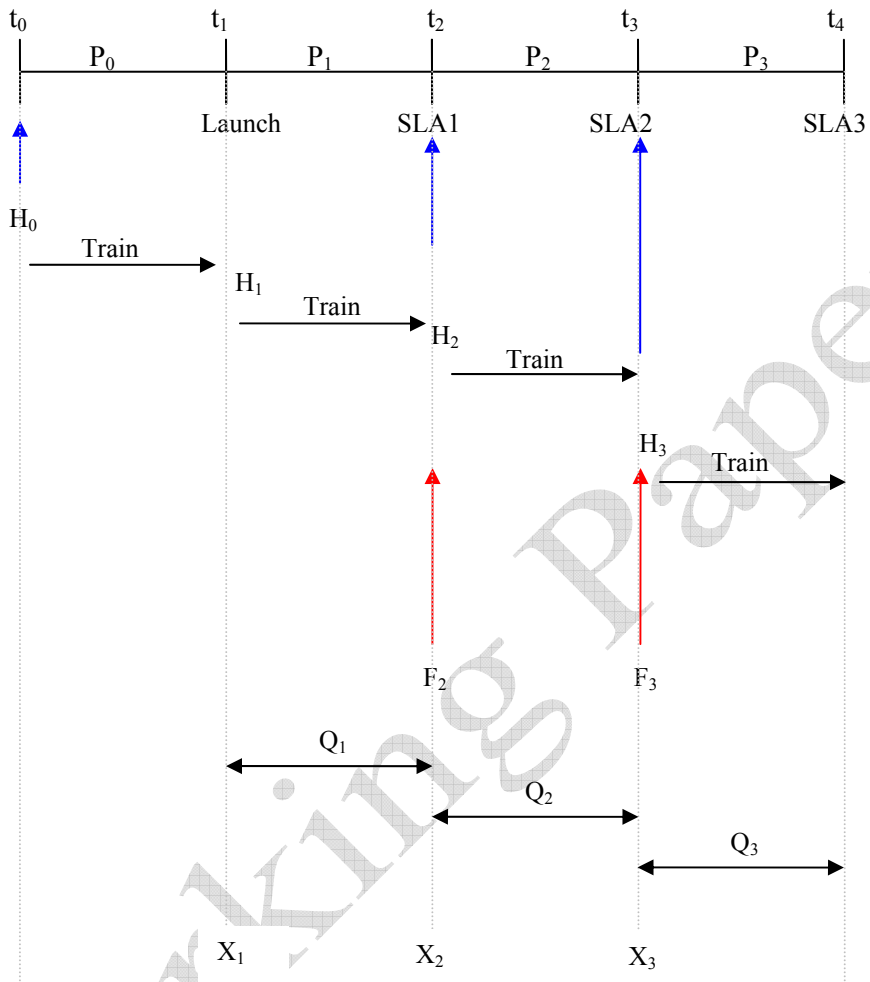

Figure 4-2 Detailed Timeline

4.2.2 Effective Capacity

An important consideration in this model is the capacity available to service calls in any time period. We define the *base capacity* as the total number of agents available in the period. The *effective capacity* is the capacity of the equivalent number of experienced agents; that is the base capacity adjusted for training and deflated by the relative productivity of the agent base.

The base capacity is impacted by hiring, firing and attrition. Net capacity is impacted by agents held for training along with learning curve issue. In any time period average agent productivity will vary based on length of service. Throughout this analysis we assume that all terminations,

Working Paper

voluntary or involuntary, come from the initial hiring class. This is a conservative assumption; these will be the longest tenured and most productive agents. We also assume that *quits* are distributed evenly through the time period but firing occurs at the beginning of the period. Furthermore, as illustrated in figure 5-2, we assume that no firing occurs at the start of period one. In this model firing occurs only to adjust capacity and there is no reason to adjust capacity prior to launch. In reality firing may also occur because of performance issues. This is a random event and for the purposes of this model such firing can be included in the attrition parameter.

The base and effective capacity available in each time period is summarized in the following table:

Period	Base Capacity	Effective Capacity
1	$H_0 + H_1 - .5Q_1$	$\rho_{01}\left(H_0 - .5Q_1\right)$
2	$H_0 + H_1 + H_2 - .Q_1 - .5Q_2 - F_2$	$\rho_{02}\left(H_0 - .Q_1 - .5Q_2 - F_2\right) + \rho_{12}H_1$
3	$H_0 + H_1 + H_2 + H_3 - .Q_1 - Q_2 - .5Q_3 - F_1 - F_2$	$\rho_{03}\left(H_0 - .Q_1 - Q_2 - .5Q_3 - F_2 - F_3\right) + \rho_{13}H_1 + \rho_{23}H_2$

Table 4-1 Start Up Capacity by Period

4.3 Characterizing Uncertainty

In (Robbins and Harrison 2007b) they develop a scheduling model that considered variability in arrival rates. In this model the variability is stochastic in the sense that outcomes are random realizations from known probability distributions. In the start up phase we face the additional challenge of parameter uncertainty. In most cases the decision maker does not have hard data on key system parameters but must instead make subjective prior estimates.

For the sake of this analysis we will assume uncertainty in the following parameters:

- **Volume**: overall average weekly call volume.
- **Arrival Rate Variability**: the level of variability in day of week and time of time call variability.
- **Talk Time**: average service time for experienced agents.
- **Learning Curve**: the learning curve coefficients both at the individual and institutional level.

Working Paper

In each case we assume that the true parameter value is unknown and is drawn from some prior probability distribution. In addition to parameter uncertainty the system under analysis here faces stochastic variability in key parameters; specifically:

- **Realized Volume**: actual call volume presented by day of week and time of day.
- **Attrition**: the number of employees who resign in any time period.

For the sake of this analysis, all other parameters (e.g. hiring cost, firing cost and SLA penalties) are considered known[8]. These parameters may however vary over time; for example hiring may be less expensive in the initial period when training costs can be amortized over a large number of hires. In this model we assume that parameter uncertainty is effectively eliminated during the first operational period. During P1 managers have the opportunity to observe four weeks of data and make informed estimates of model parameters. In subsequent periods model parameters exhibit only stochastic variability.

The process of generating scenarios than proceeds as follows. For a given number of realizations at each stage, we calculate the total number of scenarios, information bundles, and nonanticipativity constraints at stages two and three. For each realization at stage one a set of parameter values are sampled that hold for that branch of the tree. In each stage stochastic variability is added to the service level curves by adding an error term to the intercept of the service level curves. Attrition rates are estimated by calculating a random binomial variable based on the period average attrition and an average estimate of staffing. Individual productivity is calculated by calculating the average productivity rate from the learning curve and adding a stochastic error term.

4.3.1 Service Level Approximations

A key consideration in the practical application of this program is the development of a set of service level approximation curves; the curves whose coefficients create the piecewise linear

[8] We also fix the institutional productivity factor to 1 for this analysis and consider only the impact of individual productivity. While the distinction between individual and institutional productivity is theoretically appealing, we lack the data to make independent estimates. It's also apparent from the analysis that follows that the level of post launch hiring is small enough so that the distinction has no practical impact on the results.

Working Paper

approximation of the service level achieved for various staffing level decisions. These curves are represented in the problem formulation as the coefficients m_{tkh} and b_{tkh} in equation (4.4).

To develop the service level approximation curves we utilize the procedure described below. Let N be the average weekly volume, T the average talk time, v_d be the daily variability scale factor, and v_t be the time period variability scale factor.

1. Identify a template project profile that has the approximate seasonality pattern of the new project.
2. Define prior probability distributions for stochastic parameters, N, T, v_d, v_t
3. Generate a uniform design for four factors and 10 design points.
4. Define S different staffing levels.
5. For each design point, set the total volume and scale the variability of arrivals appropriately. Generate five batches of 25 scenarios each.
6. For each staff level find the associated service level by solving problem.
7. Calculate a slope and intercept for each adjacent pair of staffing levels using the average of the five batches.

Figure 4-3 Service Level Approximation Process

In step six of the process outlined in Figure 4-3we must find the service level for a sample call pattern for a fixed staffing level. The scheduling algorithm in (Robbins and Harrison 2007a) solves a related problem - finding the minimum cost schedule to achieve a desired service level. We can modify this program to find the maximum expected service level possible for a given staffing level. We call this program the service level maximization problem. Using the notation from (Robbins and Harrison 2007a) the model can be expressed as

$$\max \sum_{k \in K} p_k \frac{\sum_{i \in I} y_{ik}}{\sum_{i \in I} n_{ik}} \tag{4.9}$$

subject to

$$y_{ik} \le m_{ikh} \sum_{j \in J} a_{ij} x_j + b_{ikh} \qquad \forall i \in I, k \in K, h \in H \tag{4.10}$$

$$y_{ik} \le n_{ik} \qquad \forall i \in I, k \in K \tag{4.11}$$

Working Paper

$$\sum_{j \in J} a_{ij} x_j \geq \mu_i \qquad \forall i \in I \qquad (4.12)$$

$$x_j \leq m_j \qquad \forall j \in J \qquad (4.13)$$

$$\sum_{j \in J} x_j \leq N \qquad (4.14)$$

$$x_j \in \mathbb{Z}^+, y_{ik} \in \mathbb{R}^+ \qquad \forall i \in I, j \in J, k \in K \qquad (4.15)$$

The objective function (4.9) seeks to maximize the expected service level; the ratio of calls answered within service level to the total number of calls. Constraints (4.10) and (4.11) create a piecewise linear approximation of the service level curve. Equation (4.12) creates a lower bound on the total number of agents scheduled in each period and this coefficient is set to achieve at least a 50% expected service level and guarantee that at least two agents are always staffed. Constraint (4.13) sets an upper limit on the total number of agents that can be scheduled to each shift, and constraint (4.15) enforces non-negativity and integrality conditions. Constraint (4.14) specifies the allowable number of agents. The curve will be generated by solving this program for a series of agent levels.

4.3.2 The Base Case Example

To illustrate the analysis process use an example. Assume that we are planning a launch of a new corporate support project that operates 24x7. The project is subject to an 80/60 SLA. We first pick a similar project profile, which in this case is Project J. In many cases detailed call volume data is not available, for example if multiple help desks are being consolidated. In this case assume that the best estimate is that call volume will average 5,000 calls per week, and that we are reasonably sure volume will be at least 4,000 and no more than 7,000 calls per week. With limited data a common prior distribution is the triangular distribution (Law 2007). So for planning purposes we assume that the true expected average weekly volume has a triangular distribution with parameters (4000, 5000, 7000). Talk time for corporate projects tends to be in the range of 9 to 14 minutes. Without empirical data we will assume that the true average talk time is drawn from a uniform distribution on this range. Finally we must develop a prior estimate for the variability of arrivals, relative to project J. We will assume that the scaling factor is uniformly distributed on [0.75, 1.25].

Working Paper

Using these distributions I use a 10 point uniform design in four factors to generate 10 design points. The Uniform Design is summarized in the following table:

DP	Volume	Talk Time	Daily Variability	Time Period Variability
1	4,387	10.75	0.88	0.98
2	5,183	11.75	1.08	0.78
3	5,775	9.75	0.83	0.83
4	5,025	12.25	0.78	1.18
5	5,357	10.25	1.23	1.03
6	5,551	13.75	0.93	1.08
7	6,452	12.75	0.98	0.93
8	6,051	11.25	1.13	1.23
9	4,671	13.25	1.18	0.88
10	4,866	9.25	1.03	1.13

Table 4-2 Uniform Design for Service Level Approximations

The Uniform Design approach ensures that the 10 points effectively fill the four dimensional design space (Fang, Lin *et al.* 2000; Santner, Williams *et al.* 2003).

At each design point we generate 25 scenarios. The service level maximization problem (4.9) - (4.15) is solved for 8 staffing levels (15,20,25,30,35,40,50,65).

Working Paper

Executing this process results in the service level curves shown in the following figure:

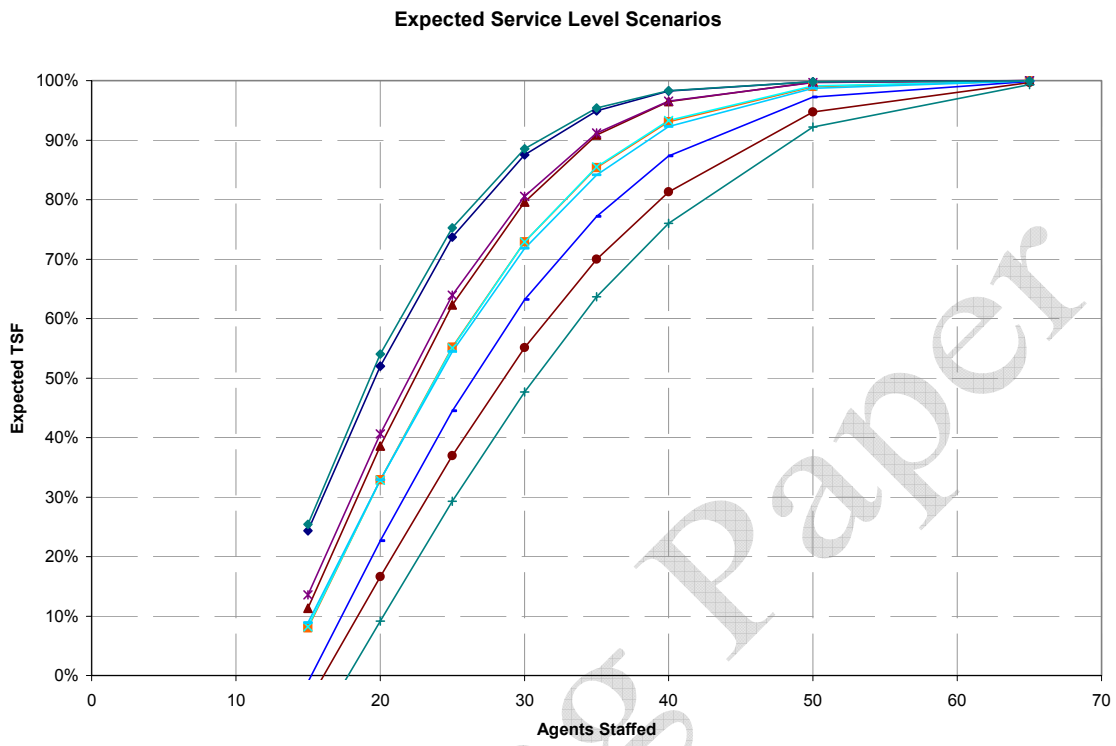

Figure 4-4 Estimated Service Level Curves

Each of the 10 lines in this figure represents, based on the assumed priors, an equally likely aggregate service level curve.

4.3.3 SAA Based Algorithm

The results of the previous section suggest that finding a precise solution from a single optimization run may be quite expensive. They also suggest however, that on average, even low scenario solutions provide good estimates. This suggests a batch solution algorithm that is a variation of the Sample Average Approximation approach and has three main steps

- **Identify Candidate Solutions**: solve a batch of sample path problems to identify one or more candidate solutions.
- **Evaluate Candidates**: calculate the expected outcome for each candidate against a reference set of scenarios. Select the candidate with the lowest expected cost.
- **Calculate Bounds**: calculate statistical bounds on the outcome and optimality gap.

Working Paper

4.4 Numerical Analysis

4.4.1 Screening Analysis

In this section we perform a series of computational experiments to estimate the optimal level of hiring and expected cost for various project conditions. As an initial objective we seek to determine what factors have the biggest impact on the optimal level of excess hiring. To accomplish this goal we conduct a preliminary screening experiment. We use a fractional factorial experiment of resolution IV. This experiment will allow the unconfounded estimation of all main effects.

In this initial screening we consider the following seven factors:

1. **Hiring Cost**: the cost to hire new agents in period 0.
2. **Termination Cost**: the cost to terminate an employee.
3. **SLA Target**: the contractual SLA target.
4. **Operational Penalty Rate**: the penalty rate per point of SLA shortfall assigned beginning in period three.
5. **Hiring Cost Differential**: the incremental cost, relative to the period zero hiring cost, to hire agents in periods 1-3.
6. **Period Two Penalty Rate**: the penalty rate per point of SLA shortfall assigned in period two.
7. **Minimum Launch Expected SL**: the minimal expected service level allowable in the first period.

The first four factors essentially address base operating characteristics of the system, such as the cost to hire and fire along with the financial constraints placed on SLA attainment. Factors five through seven are focused on the transition phase. Factor five estimates the inefficiency from last minute hiring, while factors six and seven indicate how poor service level performance can be during the transition period; factor 6 specifies how severely shortfalls are penalized in the second month of operation and factor seven specifies the required expected performance in the first month of launch.

Working Paper

The experimental design is a 16 run 2_{IV}^{7-3} experiment and allows the unconfounded estimation of all main (single factor) effects.

	A	B	C	D	E	F	G
1	-	-	-	-	-	-	-
2	+	-	-	-	+	-	+
3	-	+	-	-	+	+	-
4	+	+	-	-	-	+	+
5	-	-	+	-	+	+	+
6	+	-	+	-	-	+	-
7	-	+	+	-	-	-	+
8	+	+	+	-	+	-	-
9	-	-	-	+	-	+	+
10	+	-	-	+	+	+	-
11	-	+	-	+	+	-	+
12	+	+	-	+	-	-	-
13	-	-	+	+	+	-	-
14	+	-	+	+	-	-	+
15	-	+	+	+	-	+	-
16	+	+	+	+	+	+	+

Factor Definitions		-	+
A	Hiring Cost	500	1500
B	Hiring Cost Differential	0%	50%
C	Termination Cost	0	3200
D	SLA Target	80%	90%
E	Operational Penalty Rate	50,000	320,000
F	Period Two Penalty Rate	0%	100%
G	Min Launch Expected SL	50%	75%

Table 4-3 Screening Analysis Design of Experiment

At each design point we first solve the Mean Value problem. We then solve the stochastic problem using the process outlined in 5-13. We solve 15 instances of the problem at each design point using a 30/10/10 realization pattern for 3,000 scenarios. We then run an evaluation comparing the average hiring level, rounded to the nearest integer, with the four closest neighboring solutions. The comparison is run against a set of 9,000 scenarios generated from a 60/15/10 realization pattern. The solution with the best expected outcome is selected and the results from the 9,000 scenario run are used to calculate the statistical properties of the outcome. The mean value solution is evaluated against the same set of scenarios to estimate its expected outcome. The Value of the Stochastic Solution (VSS) is then calculated as the difference between the expected outcome of implementing the mean value solution and implementing the stochastic solution.

The results of this analysis are summarized in the following table. The table lists the coded value of each factor along with the hiring level determined by the mean value program and the hiring level selected through the evaluation process outlined above. The table also lists the VSS calculated.

Working Paper

DP	A	B	C	D	E	F	G	MV Hire	Best Hire	Best Outcome	VSS	VSS %
1	-	-	-	-	-	-	-	38	38	266,460	0	0.0%
2	+	-	-	-	+	-	+	50	49	356,154	3,408	0.9%
3	-	+	-	-	+	+	-	50	52	340,442	3,360	1.0%
4	+	+	-	-	-	+	+	50	49	375,309	600	0.2%
5	-	-	+	-	+	+	+	50	50	367,397	0	0.0%
6	+	-	+	-	-	+	-	43	45	390,142	2,393	0.6%
7	-	+	+	-	-	-	+	50	49	328,474	5,435	1.6%
8	+	+	+	-	+	-	-	40	38	318,460	7,777	2.4%
9	-	-	-	+	-	+	+	54	56	387,236	900	0.2%
10	+	-	-	+	+	+	-	56	60	469,510	6,769	1.4%
11	-	+	-	+	+	-	+	50	49	338,152	2,162	0.6%
12	+	+	-	+	-	-	-	39	38	346,832	1,633	0.5%
13	-	-	+	+	+	-	-	38	38	309,384	0	0.0%
14	+	-	+	+	-	-	+	50	49	388,464	0	0.0%
15	-	+	+	+	-	+	-	54	52	401,208	623	0.2%
16	+	+	+	+	+	+	+	56	58	491,623	1,136	0.2%

Table 4-4 Experimental Results

A few important observations are apparent from this data. First and foremost, the mean value problem often provides very good results. In several cases the mean value problem finds the same hiring level as the stochastic problem and hence the VSS is zero. In cases where the stochastic model finds a different solution, the Value of the Stochastic Solution is relatively small. In the best case, the VSS represents about a 2.4% improvement over the mean value solution. It is apparent from this analysis that the main benefit from the stochastic model will not be in improving the objective, but rather in understanding the statistical distributions of the outcomes. We return to this topic later, but first analyze the impact of the experimental factors on the outcomes.

The following table summarizes the Main Factor Effects of each experimental factor on two response variables, the stage zero hiring level and the expected cost of operation over the start-up period.

Working Paper

Factor Definitions	Main Effects Hiring	Main Effects Objective
A Hiring Cost	0.06	12,429 *
B Hiring Cost Differential	0.00	180
C Termination Cost	-0.38	3,596 *
D SLA Target	0.94	12,174 *
E Operational Penalty Rate	0.56	3,344
F Transition Penalty Rate	2.31 *	17,828 *
G Min Launch Expected SL	1.50 *	5,949 *
Average	48.13	367,203

* Indicates significance at the 95% level

Table 4-5 Main Effects on Hiring and Objective

The data in Table 4-4 shows that for this particular volume estimates the average optimal hiring level is just under 49, but varies from 38 to 56. Table 4-5 decomposes this variability into the effect that results from each variable. The most significant factors to impact the stage zero hiring decision relate to the degree that the SLA must be met during the transition phase; factors F and G are in fact the only statistically significant factors. In a tight start up, where the transition penalty and minimum staffing levels are both high, the optimal hiring is on average higher by about 7.6 individuals, an increase of nearly 16% from the mean.

Other factors have a much smaller impact on initial hiring, and are not statistically significant. Raising the steady state service level requirement only increases hiring by about 1.8 on average. A tighter service level requirement, expressed as a higher operational penalty rate, increase hiring by only about one full time equivalent. Conversely, making it more expensive to terminate employees depresses initial hiring by only about one FTE. Finally, the cost of hiring has very little impact on the optimal number to hire as these costs are dominated by other costs in the decision process.

The last column of Table 4-5 provides additional information about how these factors drive the expected cost of operation. Again, the service level requirements during transition have a practical and statistically significant impact on the expected outcome. Tight transition requirements add about 12.9% to the start up cost. The steady state service level requirement and penalty rate also add significantly to the overall cost.

Working Paper

Termination costs on the other hand have a more significant impact on the total cost than they do on the stage zero hiring decision. While a high cost of terminating employees adds to the total cost of operation, it has a relatively limited impact on the initial decision. The ability to downsize the staff once uncertainty is revealed is a valuable recourse option, even if the cost is high. Since only a few agents will be terminated the increased cost of termination does little to lower the initial hiring level.

Similarly, the cost of hiring has a significant impact on the cost of operation but almost no impact on the hiring decision; the hiring cost shifts the cost by 6.8% but the hiring decision by only 0.4% on average. The rationale is that while the cost to hire has a major impact on the cost to start up the project, there are no other options in this model.

4.4.2 Distribution of Outcomes

The results of the previous section indicate that solving the stochastic model may lead to moderately reduced cost launches as compared to solving the mean valued solution. However, in addition to decreasing the expected cost of operation, the stochastic model provides insight to decision makers on the statistical distribution of outcomes. The mean value model provides a single point-estimate of model parameters while the stochastic model allows us to estimate the statistical distribution of outcomes.

In the solution process outlined in Figure 5-13 we estimate the outcome of the candidate solution by evaluating the solution against a set of evaluation scenarios. The expected outcome for any random parameter is the average result over all the evaluations scenarios. If we examine the scenario results in detail we can estimate the distribution.

As an example, the following figure plots a histogram of the objective value generated for DP1. Recall from Table 4-4 that the expected outcome of this startup is $266,458.

Working Paper

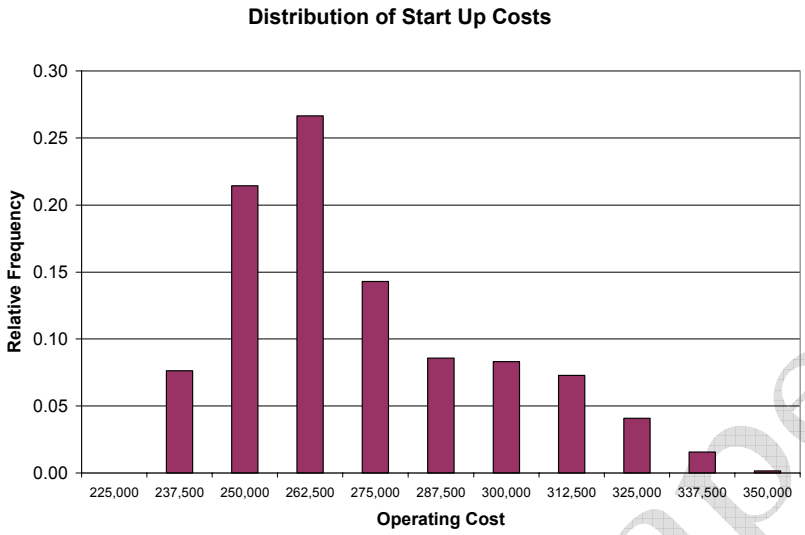

Figure 4-5 Distribution of Start Up Costs – DP1

The graph shows that the outcome is positively skewed with outcomes as much as $83,000 (31%) above the mean possible. In this particular case there is a 21% probability that start up costs will exceed $300,000. A similar analysis generates the following histogram of the ending staff level.

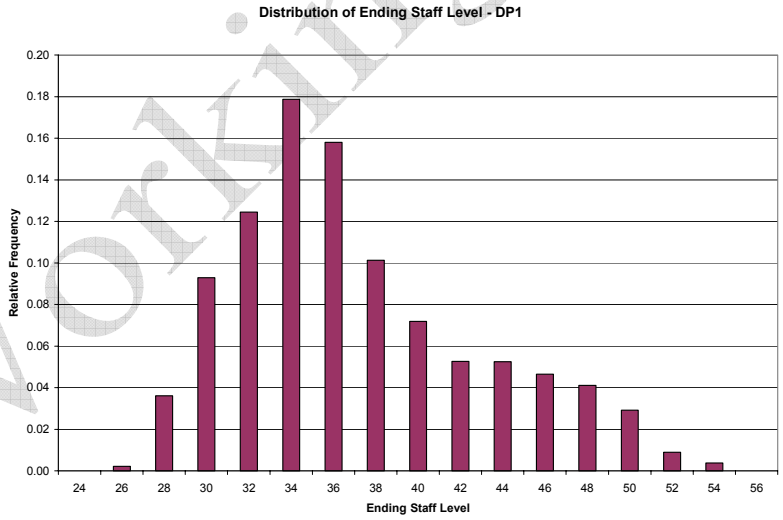

Figure 4-6 Distribution of Ending Staff Level – DP1

The mode of this distribution is 34, while the initial hiring level was 38; indicating that on average the staff level will decrease over the course of the start-up. In this particular case there is in fact an approximately 60% probability that the ending staff level will be less than the number

Working Paper

originally hired. The staff level reduction occurs because of attrition, but also because of terminations. In this particular case the expected number of post launch terminations is 1.6. The logic is fairly straightforward. Given uncertainty and learning curve issues the optimal policy calls for acquiring some spare capacity prior to launch.

The following figure plots the distribution for the number of agents hired and fired post launch. It may seem odd that the expected number of hire and fires are both positive, but the specific action will depend on how demand is realized. In rare cases the model may call for post launch hiring and firing in the same scenario.

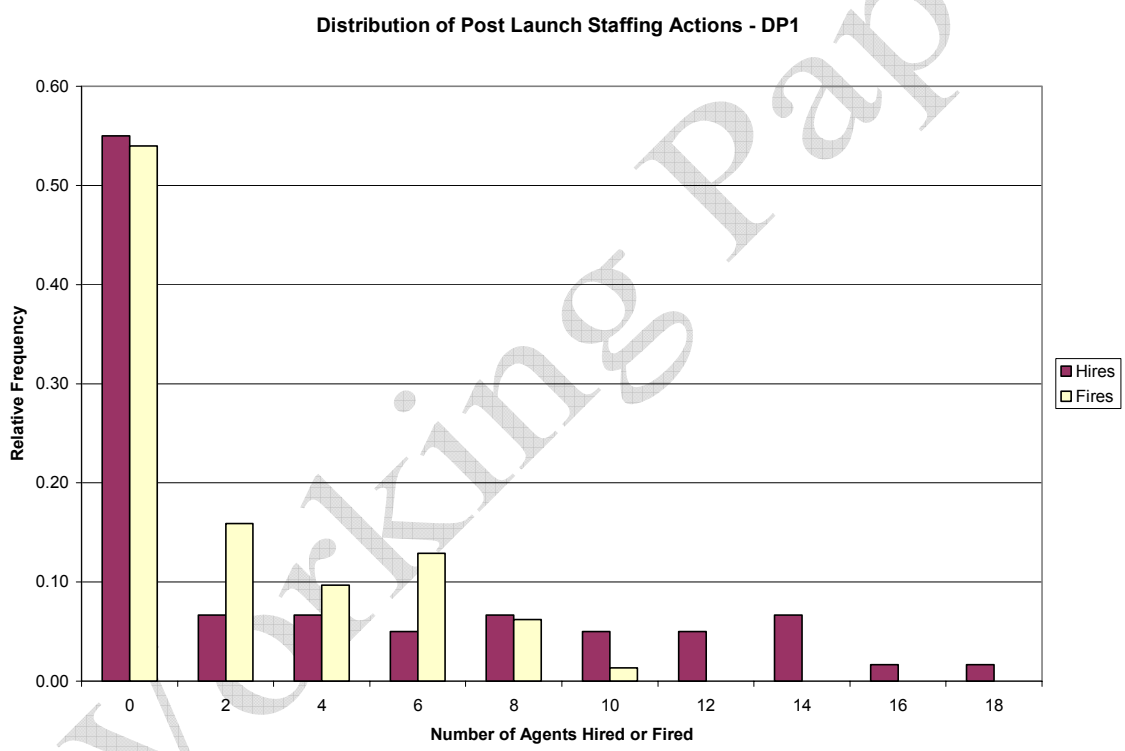

Figure 4-7 Distribution of Post Launch Hiring and Firing – DP1

The above graphs show the distribution of key outcome graphically for a single design point. To get a sense of how these distributions may vary with the model's control factors we list some summary statistics in the following tables. The first table summarizes the distribution of the objective and ending hiring level, while the second summarizes the post launch hiring and firing.

Working Paper

	Cost					Ending Staff				
DP	Avg	SD	Min	Max	Range	Avg	SD	Min	Max	Range
1	266,460	24,496	227,229	345,676	118,447	36.1	5.8	24.9	53.6	28.7
2	356,154	19,052	322,936	445,399	122,463	36.6	5.2	26.1	52.8	26.7
3	340,442	59,465	284,468	592,640	308,172	36.7	5.3	26.1	52.1	26.0
4	375,309	49,156	322,376	551,197	228,821	36.1	4.9	26.0	49.0	23.0
5	367,397	62,719	313,761	635,479	321,718	45.3	2.2	36.8	52.8	16.0
6	390,142	60,797	327,385	599,291	271,906	40.9	2.0	33.1	46.7	13.7
7	328,474	7,180	308,240	368,134	59,894	44.3	2.2	36.0	49.7	13.7
8	318,460	33,559	276,458	444,065	167,607	38.0	5.1	28.6	55.6	27.0
9	387,236	56,808	318,638	567,845	249,207	43.4	5.2	31.2	56.0	24.8
10	469,510	73,557	395,926	747,177	351,252	44.8	6.1	31.6	61.5	29.9
11	338,152	31,519	289,148	449,185	160,037	44.5	7.1	31.1	67.2	36.1
12	346,832	38,252	282,628	453,562	170,934	36.2	2.7	28.6	46.5	18.0
13	309,384	35,323	247,620	422,780	175,160	46.5	8.4	30.6	70.9	40.3
14	388,464	18,862	357,240	462,380	105,141	44.3	2.2	36.0	49.0	13.0
15	401,208	63,593	331,165	605,114	273,949	47.0	2.3	38.2	52.0	13.8
16	491,623	74,325	421,963	778,523	356,561	52.5	2.6	42.6	59.1	16.4

Table 4-6 Summary of Cost and Ending Staff

	Post Launch Hiring					Post Launch Firing				
DP	Avg	SD	Min	Max	Range	Avg	SD	Min	Max	Range
1	3.41	4.91	0.00	16.72	16.72	1.56	2.34	0.00	8.84	8.84
2	0.21	0.80	0.00	4.53	4.53	8.31	5.50	0.00	19.84	19.84
3	0.02	0.16	0.00	1.23	1.23	10.90	6.02	0.00	22.83	22.83
4	0.00	0.00	0.00	0.00	0.00	8.63	5.58	0.00	19.84	19.84
5	0.14	0.62	0.00	3.85	3.85	0.00	0.00	0.00	0.00	0.00
6	0.20	0.66	0.00	3.19	3.19	0.00	0.00	0.00	0.00	0.00
7	0.05	0.29	0.00	2.17	2.17	0.00	0.00	0.00	0.00	0.00
8	3.92	5.40	0.00	17.79	17.79	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	7.63	5.91	0.00	20.53	20.53
10	0.08	0.49	0.00	3.58	3.58	10.06	6.76	0.00	23.53	23.53
11	3.22	5.34	0.00	18.65	18.65	2.99	3.56	0.00	12.53	12.53
12	2.01	2.67	0.00	9.95	9.95	0.08	0.33	0.00	2.53	2.53
13	12.99	9.08	0.00	32.86	32.86	0.00	0.00	0.00	0.00	0.00
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	0.02	0.12	0.00	0.84	0.84	0.00	0.00	0.00	0.00	0.00
16	0.08	0.53	0.00	4.00	4.00	0.00	0.00	0.00	0.00	0.00

Table 4-7 Summary of Post Launch Staffing Actions

The data shows the wide range of outcomes possible. In many cases the cost of operation varies by more than $250,000, and final staffing varies by as much as 40.

Working Paper

4.5 Summary and Conclusions

4.5.1 Summary

This model examines the issue of how to staff a new call center outsourcing project in the face of uncertainty about demand. We use a version of a stochastic scheduling model developed in (Robbins and Harrison 2007a) to generate an estimated aggregate service level curve for a range of possible demand outcomes. Given those estimates we develop a multistage model of the project start-up process that accounts for agent learning, and attrition.

The model is developed as a multistage stochastic problem, with an integer constrained decision in the first stage. The analysis shows that an approach, based on Sample Average Approximations, provides tractable solutions to this problem, exploiting the fact that the stage 0 decision is scalar valued.

Detailed numerical analysis shows that the stochastic formulation provides a moderate benefit in terms of lowering the cost of the launch. It does however provide a significant qualitative benefit in terms of contingency planning by providing estimated distributions of key parameters such as total cost, hiring, or firing. Given these distributions, managers can make more informed decisions regarding pricing as well as staffing contingencies.

A key insight in this model is the importance of the transition phase of the start-up. Given the nature of the learning curve, it is extremely difficult, and expensive, to achieve targeted service levels in the first few month of the launch. Managers have responded by attempting to lower expectations for the service level over the first few months of launch. Our analysis supports that strategy, but also helps to quantify the costs associated with attempting to meet service level commitments in the first few moths. The model also helps to quantify the degree to which the project should be overstaffed at start-up, a practice currently not employed at the company we analyzed. The analysis indicates the combined effects of learning and attrition provide strong incentive to err on the side of over hiring.

The analysis presented in this paper identifies several issues with important managerial implications. The most important implications arise from issues related to the learning curve. At

Working Paper

launch time all agents are inexperienced and subject to rapid productivity improvement. Failure to account for this learning when planning a startup, as is the case in the company we studied, will often lead to significant start up challenges. A second key implication is the need to plan for staffing flexibility. The analysis shows that the optimal staffing level at start-up time is not likely to be the optimal staffing level once the transition phase has been completed. Managers must maintain the flexibility to add additional staff, or if necessary to remove staff once uncertainty is revealed and learning has occurred. Lastly, is the issue of the quality of service during the transition phase. In projects where learning is significant rapid attainment of service level objectives is possible, but very expensive.

Several extensions to this model are possible including the analysis of alternative utility finctions, incorporating the effects of investments in learning and issues related to phased start ups.

- **Alternative Utility Functions**: the model has implicitly assumed a risk neutral decision maker seeking to minimize expected cost. The stochastic model formulation allows for other utility models. An approach such as minimax that seeks to minimize the maximum cost can be implemented quite easily.
- **Investment in Learning**: the learning curve has a significant impact on the project start-up. Throughout this analysis we assumed that the learning rate was exogenous. One potential extension of this research is to examine the benefit that would accrue from investments that increase the rate of learning, for example higher levels of investment in training or knowledgebase development.
- **Phased Start-Up**: the analysis presented here is based on a single cut over of support services. The model is however motivated in part by the service level collapse observed during a phased roll out process. Each time a new cutover occurred, things became worse as the project got further and further behind. Extending this model to look at a phased rollout would be difficult, but beneficial.

Working Paper

5 References

Aksin, Z., M. Armony and V. Mehrotra 2007. The Modern Call-Center: A Multi-Disciplinary Perspective on Operations Management Research. Working Paper 61p.

Anderson, E. G., Jr. 2001. The Nonstationary Staff-Planning Problem with Business Cycle and Learning Effects. *Management Science* **47**(6) p. 817-832.

Bartholomew, D. J. 1982. *Stochastic models for social processes*, Wiley. Chichester [England]; New York.

Bartholomew, D. J. and A. F. Forbes 1979. *Statistical techniques for manpower planning*, Wiley. Chichester [Eng.]; New York.

Birge, J. R. and F. Louveaux 1997. *Introduction to Stochastic Programming*, Springer. New York.

Bordoloi, S. K. and H. Matsuo 2001. Human resource planning in knowledge-intensive operations: A model for learning with stochastic turnover. *European Journal of Operational Research* **130**(1) p. 169.

Borst, S., A. Mandelbaum and M. I. Reiman 2004. Dimensioning Large Call Centers. *Operations Research* **52**(1) p. 17-35.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, S. Haipeng, S. Zeltyn and L. Zhao 2005. Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective. *Journal of the American Statistical Association* **100**(469) p. 36-50.

Ebert, R. J. 1976. Aggregate Planning with Learning Curve Productivity. *Management Science* **23**(2) p. 171-182.

Fang, K.-T., D. K. J. Lin, P. Winker and Y. Zhang 2000. Uniform Design: Theory and Application. *Technometrics* **42**(3) p. 237-248.

Gaimon, C. 1997. Planning Information Technology-Knowledge Worker Systems. *Management Science* **43**(9) p. 1308-1328.

Gaimon, C. and G. L. Thompson 1984. A Distributed Parameter Cohort Personnel Planning Model that Uses Cross-Sectional Data. *Management Science* **30**(6) p. 750-764.

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

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

Garnett, O., A. Mandelbaum and M. I. Reiman 2002. Designing a Call Center with impatient customers. *Manufacturing & Service Operations Management* **4**(3) p. 208-227.

Green, L. V., P. J. Kolesar and J. Soares 2001. Improving the SIPP Approach for Staffing Service Systems That Have Cyclic Demands. *Operations Research* **49**(4) p. 549-564.

Green, L. V., P. J. Kolesar and W. Whitt 2005. Coping with Time-Varying Demand when Setting Staffing Requirements for a Service System. Working Paper 58p.

Grinold, R. C. 1976. Manpower Planning with Uncertain Requirements. *Operations Research* **24**(3) p. 387-399.

Working Paper

Halfin, S. and W. Whitt 1981. Heavy-Traffic Limits for Queues with Many Exponential Servers. *Operations Research* **29**(3) p. 567-588.

Hanssmann, F. and S. W. Hess 1960. A Linear Programming Approach to Production and Employment Scheduling. *Management Technology* **1**(1) p. 46-51.

Harrison, T. P. and J. E. Ketz 1989. Modeling Learning Effects via Successive Linear Programming. *European Journal of Operational Research* **40**(1) p. 78-84.

Holt, C. C., F. Modigliani, J. F. Muth and H. A. Simon 1960. *Planning production, inventories, and work force*, Prentice-Hall. Englewood Cliffs, N.Y.

Robbins, T. R. and T. P. Harrison 2007a. A Stochastic Programming Model for Scheduling Call Centers with Global Service Level Agreements. Working Paper 34p.

Robbins, T. R. and T. P. Harrison 2007b. A Stochastic Programming Model for Scheduling Call Centers with Uncertain Arrivals. *Eighteenth Annual Conference of POMS*, Dallas, TX.

Santner, T. J., B. J. Williams and W. Notz 2003. *The design and analysis of computer experiments*, Springer. New York.

Whitt, W. 2005. Engineering Solution of a Basic Call-Center Model. *Management Science* **51**(2) p. 221-235.

Whitt, W. 2006a. The Impact of Increased Employee Retention Upon Performance in a Customer Contact Center. Working Paper 35p.

Whitt, W. 2006b. Sensitivity of Performance in the Erlang A Model to Changes in the Model Parameters. *Operations Research* **54**(2) p. 247-260.

Working Paper