This article was downloaded by: [155.246.103.35] On: 25 March 2017, At: 23:16
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA

Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

OM Practice—Work Expands to Fill the Time Available: Capacity Estimation and Staffing Under Parkinson's Law

Sameer Hasija, Edieal Pinker, Robert A. Shumsky,

To cite this article:
Sameer Hasija, Edieal Pinker, Robert A. Shumsky, (2010) OM Practice—Work Expands to Fill the Time Available: Capacity Estimation and Staffing Under Parkinson's Law. Manufacturing & Service Operations Management 12(1):1-18. http://dx.doi.org/10.1287/msom.1080.0250

Full terms and conditions of use: http://pubsonline.informs.org/page/terms-and-conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2010, INFORMS

Please scroll down for article—it is on subsequent pages

INFORMS is the largest professional society in the world for professionals in the fields of operations research, management science, and analytics.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 12, No. 1, Winter 2010, pp. 1–18
ISSN 1523-4614 | EISSN 1526-5498 | 10 | 1201 | 0001

informs®
DOI 10.1287/msom.1080.0250
© 2010 INFORMS

Downloaded from informs.org by [155.246.103.35] on 25 March 2017, at 23:16 . For personal use only, all rights reserved.

OM Practice

Work Expands to Fill the Time Available: Capacity Estimation and Staffing Under Parkinson's Law

Sameer Hasija
INSEAD, 138676, Singapore, sameer.hasija@insead.edu

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

Robert A. Shumsky
Tuck School of Business Administration, Dartmouth College, Hanover, New Hampshire 03755, robert.shumsky@dartmouth.edu

We develop a method to estimate the capacity of agents who answer e-mail in a contact center, given aggregate historical data that have been distorted both by constraints on work availability and by internal incentives to slow down when true capacity exceeds demand. We use the capacity estimate to find a contact center's optimal daily staffing levels. The implementation results, from an actual contact center, demonstrate that the method provides accurate staffing recommendations. We also examine and test models in which agents exhibit speed-up behavior and in which capacity varies over time. Finally, we use the capacity estimates to examine the implications of solving the staffing problem with two different model formulations, the service-level constraint formulation used by the contact center and an alternate profit-maximization formulation.

Key words: capacity planning; service operations; empirical research; OM-human resources interface
History: Received: September 26, 2007; accepted: November 15, 2008. Published online in *Articles in Advance* April 8, 2009.

1. Introduction

In this paper we describe capacity estimation and staffing algorithms for an e-mail contact center that provides customer support for a large client. The staffing problem is a standard and well-studied problem, but virtually all models in the published literature assume that a key parameter, the agents' service rate, is known or can be inferred directly from historical data. The same assumption is made in most commercial staffing software. In their survey of the call center staffing literature, for example, Gans et al. (2003, p. 96) state that service rates are usually found via simple calculations using "grand averages" of historical data.

In practice, historical data often cannot be taken at face value. In this paper we distinguish between the observed productivity (or simply productivity) of a group of servers and their actual capacity, where capacity is defined as maximum productivity attained while still satisfying quality constraints. Productivity can differ from capacity in a variety of ways, for a variety of reasons. For example, Brown et al. (2005, p. 39) observe service times of less than 10 seconds from one call center and find that these short times were "primarily caused by agents who simply hung up on customers to obtain extra rest time." They add, "The phenomenon of agents 'abandoning' customers is not uncommon; it is often due to distorted incentive schemes, especially those that overemphasize short average talk-time or, equivalently, the total number of calls handled by an agent."

For our contact center, the productivity data were distorted by a combination of external work limits and internal incentives. First, the client provided the contact center with daily *upper bounds* on its productivity.

RIGHTSLINK

Downloaded from informs.org by [155.246.103.35] on 25 March 2017, at 23:16 . For personal use only, all rights reserved.

Second, the line managers in our contact center were rewarded for high agent utilization, as measured by the percentage of time the servers spent replying to e-mails. As a result, it is likely that the managers encouraged the agents to stretch out their service times whenever it became apparent that the facility would reach the upper limit on work before the end of the day (thus fulfilling Parkinson's Law, the aphorism that work expands to fill the time available for its completion; Parkinson 1955). This combination of limited demand and internal incentives produces an interaction between observed individual productivity and the facility's workload. Our estimation methods take this interaction into account.

The particular mix of work rules and incentives in our facility may be unusual, but the estimation and staffing methods developed in this paper apply to many other production environments. In general, our methods apply when (i) work arrives in batches at the start of each work period, (ii) the amount of work required for each unit in the batch is difficult to measure directly, and (iii) the batch size varies and can sometimes be *less than* available capacity. Workers may respond in many ways when demand is less than capacity, and our methods are relevant when the worker's response is either to slow down (as in our facility) or to work at the maximum rate and "quit early" at a time that is difficult for the firm to observe accurately.

Examples of environments with attributes (i)–(iii) may be found in the large class of service factories that process information, such as firms that conduct insurance claims servicing, enter and adjust financial account information, prepare taxes, and perform general billing services. Because such services are often sent offshore, demand may be generated in one time zone and served halfway around the world in another. Therefore, the bulk of daily demand for the offshore service provider is generated "overnight" and is available as a variable-sized batch in the morning (attributes (i) and (iii) above). As for attribute (ii), in these facilities, unit-by-unit service times may be expensive to measure, requiring close observation or manual time stamping. When workers manage multiple jobs in parallel (such as agents working on multiple e-mails simultaneously), it may be virtually impossible to disentangle how much time was spent on an individual unit of work. Finally, the content of the services—and therefore the service rates—evolves over time. Automated capacity estimation and staffing procedures, such as those described here, are particularly useful in such facilities.

In §2 we will provide more detail about the facility's environment and existing staffing algorithms. Section 3 reviews the relevant literature and discusses this paper's contribution. Sections 4 and 5 describe the staffing model and estimation schemes, respectively. Section 6 describes the application of the model to the contact center's staffing problem. Postimplementation results demonstrate excellent performance in the center. In §7 we use our data to determine whether the contact center makes significantly suboptimal decisions when using a service-level constraint formulation for the staffing problem rather than a profit-maximization formulation. Finally, §8 summarizes our results and describes areas for further research.

2. Business Environment and Original Staffing Method

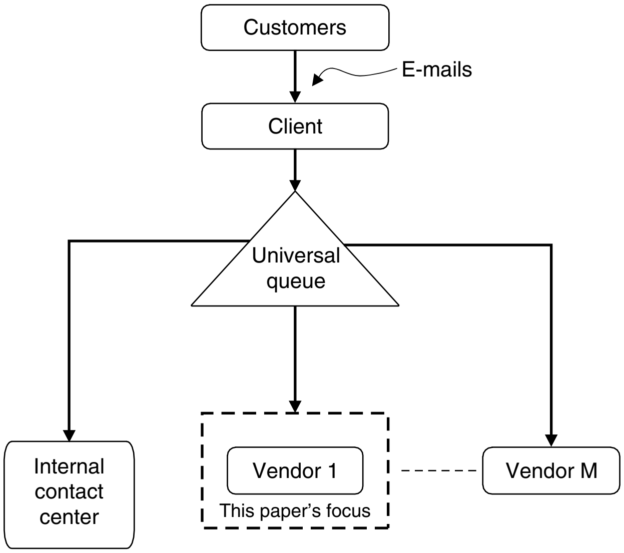
The client serves a consumer market and receives e-mail requests from its customers for sales or service. The client collects these e-mails in a "universal queue" that is shared by its own contact center as well as by multiple vendors that have been hired by the client to provide customer support (see Figure 1). All the facilities pull e-mails from the universal queue and send replies to the client's customers. The focus of this paper is exclusively on one vendor, labeled "Vendor 1" in the figure. For those interested in the client's problem, see Keblis and Chen (2006), who describe a similar environment along with a proposed solution for the client's internal staffing and call allocation problem.

For this vendor, the client sets a target number of e-mails (T_j) that the vendor should serve on day j. The seven daily targets for a particular week are sent by the client to the vendor at least one week in advance. The service-level agreement (SLA) in the contract between the client and the vendor specifies that the vendor should serve at least a proportion L and not exceed a proportion U of each daily target (for our vendor, the client specified $L = 0.9$ and $U = 1.1$). The

RIGHTSLINK

Downloaded from informs.org by [155.246.103.35] on 25 March 2017, at 23:16 . For personal use only, all rights reserved.

Figure 1 **Flow of E-Mails**



vendor is paid per e-mail served. The vendor also pays significant explicit and implicit penalties if the lower-bound term of the SLA is not met ("implicit" penalties include the potential loss of long-term business from the client). Note, however, that no penalties are assessed if the vendor's failure is caused by circumstances beyond its control, such as an empty universal queue. During our period of data collection, however, the universal queue was never empty.

The lower productivity bound, L, ensures some minimum productivity from the vendor and therefore helps to ensure that the client can serve its customers promptly. The purpose of the upper bound of the SLA, U, is not so clear. We can envision at least three reasons for the client to set such an upper bound. First, the limit may be a mechanism for the client to prevent competing vendors from monopolizing the universal queue and extracting as many e-mails as possible. This ensures that each vendor serves a substantial fraction of the demand, so that the client maintains relationships with multiple vendors, avoids potential hold-up problems when one vendor dominates, and mitigates the risk of relying on one vendor. (Note that another method for ensuring multiple-vendor participation is to form dedicated queues for each vendor. This strategy, however, may be costly because of the loss of economies of scale.) Second, recall that the client operates its own internal contact center. Upper thresholds for vendors may also help to ensure that the client's internal contact center has a sufficient workload. Finally, the upper bound may reduce the vendor's incentive to answer e-mails too quickly with low quality. These three reasons for the upper bound in the SLA, however, are speculative, and modeling the client's allocation and contracting problem is an interesting area for additional research.

Given the contract, the vendor makes the staffing, hiring, and training decisions. Once hired and trained, the vendor's agents are dedicated to answering this particular client's e-mail. Based on the daily target, managers set a staffing level. The desired staffing levels are reported to a central operations facility, which creates staff schedules for the entire organization (the vendor provides customer support service for a number of different clients). The central facility takes transportation needs into account, as well as other constraints such as vacations.

As mentioned above, line managers at the vendor earn bonuses by keeping agent utilization high as well as by meeting SLA targets. Individual agents employed by the vendor are given bonuses based on their work quality and productivity. Quality is measured by the rate of errors identified in e-mail replies; productivity is measured by the rate of e-mails answered. The weight of the productivity measure in calculating bonuses, however, is small relative to the weight given to the quality measure. We speculate that the firm may have formulated the bonus scheme in this way to emphasize quality over quantity. The scheme may also be designed to reduce the conflict between the productivity incentive, which encourages agents to work faster, and the need to occasionally slow agents down so they do not violate the upper bound.

To avoid such violations, the individual agents do not have to know details about the targets or the facilitywide productivity. The managers see the complete picture and can alter productivity in a variety of ways. On days when the facility is overstaffed, general managers can "walk the floor," explaining to individual agents that they should look forward to a light day and that they can relax. Throughout the day managers also distribute information about production volumes and targets to team leaders, agents responsible for small groups of coworkers. The team leaders pass

Downloaded from informs.org by [155.246.103.35] on 25 March 2017, at 23:16 . For personal use only, all rights reserved.

the information along to their agents using an instant messaging system.

When the authors began work with the firm, the managers knew that staffing levels were at times too low and at other times perhaps too high. These staffing errors could be caused by a variety of reasons. Two possibilities are poor estimates of worker capacity and poor staffing decisions, given a capacity estimate. Our impression is that both factors were at work at this firm. To create a staffing plan for each week, management combined the targets submitted by the client with "ballpark" estimates of the individual capacities of the agents. Specific staffing levels were generated using a simple spreadsheet model and adjusted by the manager's "gut feeling." Staffing decisions, therefore, were haphazard and open to systematic biases similar to those discussed in Schweitzer and Cachon (2000). The firm needed a software tool that would generate reliable staffing recommendations. An important component of this tool is a method to rigorously estimate capacity.

3. Related Literature and Contribution

Goodale and Thompson (2004) categorize labor scheduling problems into four steps:

(i) Forecast demand for service

(ii) Convert the forecast into staffing requirements

(iii) Satisfy staffing requirements by creating the least costly schedule

(iv) Control real-time delivery of service.

The problem described in this paper falls into the second step. We estimate the capacity of agents using historical productivity data to determine the staffing level for a given target. This paper is also related to the literature on estimating the distribution of agent service times. Gans et al. (2003, §6.3.2) describe research that attempts to characterize the distributional form of service durations in call centers. Other related work includes that of Schruben and Kulkarni (1982), who examine the consequences of using data-driven parameter estimates for demand and service rates when formulating an $M/M/1$ model to predict system performance. Ross and Shanthikumar (2005) pose the problem of an internet service provider (ISP) that hires a second ISP to handle some of its local traffic. The first ISP cannot observe the second ISP's capacity or the competing traffic, but for planning purposes the first ISP must estimate these quantities. Ross and Shanthikumar propose fast estimation methods to solve this problem. All these papers make the assumption that observed service times are samples from the actual service-time distribution, an assumption that is violated in our application.

As we mentioned above, Brown et al. (2005) describe another situation in which service times are distorted by incentives; in their case, the incentives encouraged agents to reduce service time by hanging up on customers. They show that after the call center corrected the incentives problem, the lognormal distribution provides a good fit for the service time data. They also mention that distorted data may be corrected "by using a mixture model or, in a somewhat less sophisticated manner, by deleting from the service time analysis all calls with service times <10 seconds." In our case the service-time data cannot be analyzed correctly without taking aggregate daily targets and daily productivity into account. In the following procedure we work only with the aggregate data and do not examine the actual histogram of service times (in fact, the vendor only provided aggregate data). We then derive estimates of the first two moments of the service time, rather than a full description of the service-time distribution. We will see that estimates of the mean and variance are sufficient to generate effective staffing recommendations.

Diwas and Terwiesch (2007) study the impact of workload on the productivity of employees of a hospital. They show that as workload increases, hospital employees responsible for transporting patients temporarily speed up their service rate. They also show that the length of stay and quality of care for cardiac patients in the hospital decrease as congestion increases. We will discuss models of speed-up behavior in §5.4.

Also related to our work is the literature that links inventory levels and processing times in production systems. Using a set of laboratory experiments, Schultz et al. (1998) show that the processing times of workers in a serial system vary with the size of the adjacent inventory buffers. In particular, workers reduce processing times when buffer sizes are small and they risk idling their neighbors. Schultz et al.

RIGHTSLINK

Downloaded from informs.org by [155.246.103.35] on 25 March 2017, at 23:16 . For personal use only, all rights reserved.

(1999) use the same laboratory environment to examine the effects of inventory levels and incentives on task feedback, group cohesiveness, task norms, and peer pressure. Their findings provide an explanation, based on behavioral theory, for the results in Schultz et al. (1998).

Our model is motivated by the observation that external and internal incentives affect agent behavior. Empirical research in operations management that examines the impact of incentives on agent behavior includes Lee and Zenios (2007), who show how outcome data from dialysis therapy may be used to construct an effective reimbursement system for dialysis providers. The authors also develop methods to estimate the quantities necessary to implement their scheme, given data collected from individual patients. Olivares et al. (2008) focus on the perceived costs of over- and under-use of hospital operating room capacity and develop a structural model to estimate those costs. In general, papers in this research stream develop relatively detailed models of agents' optimizing behavior in the presence of particular incentives. In our case, it was not possible to build such a detailed model, because data about individual agent and management incentives were not available. In addition, a detailed model of agent behavior was not desired by the firm and, we believe, not necessary: the firm was not planning to change its incentive system but was focused on improving its staffing system, given the available data. One of the strengths of the model developed here is that it is driven by routinely collected aggregate productivity data rather than by records of individual service times. In many service environments data on service times for individual jobs may be expensive to collect, so in many environments aggregate models are more useful than detailed microlevel models.

Most of the modeling literature on state-dependent work rates is drawn from the queueing setting (e.g., Harris 1966). In that literature, customers are actively waiting in a queue and their time is a direct cost. Therefore, as queue lengths rise, workers work faster to reduce the queue. This is similar to the speed-up model in §5.4, although on a shorter time scale.

Although beyond the scope of this study, the choices the client firm has made in structuring the work allocation system are intriguing and worthy of additional study. There exists a related literature on server competition and work allocation policies with strategic server behavior (e.g., Cachon and Zhang 2007).

4. Staffing Models

The vendor must determine the number of agents to staff on day j given target T_j, a lower limit on productivity LT_j, and an upper limit UT_j. The vendor's staffing problem for day j can be described as

$$\operatorname*{Max}_{N_j \in Z^+} \mathrm{E}[r \min(C(N_j), UT_j) - G((LT_j - C(N_j))^+) - SN_j], \qquad (1)$$

where N_j is the number of agents, r is the revenue earned per e-mail, and $C(N_j)$ is the total capacity, where total capacity is defined as the upper bound on the productivity. That is, capacity is defined as the maximum number of e-mails that can be handled in a day given N_j agents while maintaining an acceptable level of service quality. Productivity, in contrast, is the actual number of e-mails handled on a given day. The function $G(x)$ is the total penalty incurred by the vendor if the vendor's productivity is less than the lower bound specified in the SLA by x e-mails, and S is the staffing cost per agent. Note that the function $G(x)$ includes both short-term financial consequences (actual payments to the client) as well as significant, but uncertain, long-term penalties related to the potential damage to the relationship with the client.

The precise value of $G(x)$ is difficult to specify, and the vendor has chosen to use a different staffing model based on a service level for the lower bound on the SLA. The vendor's formulation is

$$\begin{aligned} &\operatorname*{Min}_{N_j \in Z^+} \; N_j \\ &\text{s.t.} \;\; \Pr(C(N_j) \geq LT_j) \geq \alpha, \end{aligned} \qquad (2)$$

where $0 < \alpha < 1$ is the desired probability of satisfying the lower bound on the SLA (values of α are typically at least 0.95). The value of α chosen by the vendor implicitly captures the trade-offs among the revenue earned per e-mail (r), the financial and goodwill costs of underage (G), and the cost of capacity (S). We discuss the implications of solving (2) instead of (1) in more detail in §7.

To calculate the optimal staffing level N_j^* we must determine the distribution of total daily capacity,

RIGHTSLINK

Downloaded from informs.org by [155.246.103.35] on 25 March 2017, at 23:16 . For personal use only, all rights reserved.

given N agents (for the remainder of this section we suppress the subscript j). Define C_i as the capacity of agent i per day, and we assume that the C_is for all agents are described by independent and identically distributed nonnegative random variables with finite mean and variance. In practice, the independence assumption may be violated, as when many agents slow down together because of problems with the vendor's information system or when many agents speed up together because the center receives a flurry of e-mails that are particularly easy to resolve. Unfortunately, the aggregate data collected by the vendor did not contain sufficient information to reliably estimate correlation among agents. We will discuss this problem again in §§5 and 8.

Given our assumptions, the total daily capacity given N agents is

$$C(N) = \sum_{i=1}^{N} C_i.$$

For this facility, the number of agents required to meet the service level agreement is usually in the range of 50 to 400, so it is reasonable to invoke the central limit theorem. Let $f(C(N))$ be the probability density function (PDF) of daily capacity, given N agents. Therefore,

$$\begin{aligned} f(C(N)) &\approx N(\mu N, \sigma\sqrt{N}), \\ \mu &= \mathrm{E}(C_i), \qquad (3) \\ \sigma^2 &= \mathrm{Var}(C_i). \end{aligned}$$

Given the distribution described in (3), our model (2) reduces to

$$\begin{aligned} &\operatorname*{Min}_{N\in Z^+} \; N \\ &\text{s.t.} \;\; 1-\Phi\left(\frac{LT-N\mu}{\sqrt{N}\sigma}\right) \geq \alpha, \end{aligned} \qquad (4)$$

where Φ is the standard normal cumulative distribution function (CDF). The challenge now is to determine μ and σ, the mean and standard deviation of one agent's daily capacity.

5. Estimation

In this section we present methods to estimate μ and σ from aggregate agent productivity data. In §5.1 we describe the available data. In §5.2 we describe three estimation techniques that we considered for implementation, and we demonstrate the methods by producing three estimates of the pair (μ, σ) from one month of historical data (Month 1). In §5.3 we discuss how we chose from among the three models. One criterion was the performance of each model in a prototype staffing algorithm. In particular, we combine estimates from each model with data from another month (Month 2) to generate recommended staffing levels, and we compare those levels to levels that were implemented by the vendor. Given the observed historical performance, this retrospective analysis allows us to assess which estimates lead to reasonable recommendations. Based on these results and other criteria described in §5.3, we chose one of the estimation methods, and the vendor implemented a staffing algorithm using that method. We report on the vendor's implementation results in §6. In §5.4 we describe models that take additional factors that may influence capacity into account, such as speed-up by agents when demand outstrips the agents' baseline capacity.

5.1. Data for Fitting the Models

The following historical data from Month 1 were collected by the vendor and used to make our estimates: (i) N_j, the number of agents staffed on day j, (ii) Y_j, the number of e-mails processed in one day by the N_j agents (the productivity on day j), (iii) M, the number of days in the sample period, and (iv) T_j, the daily target. Figures 2 to 4 show these data in three different formats. Figure 2 displays the data over the 31 days of Month 1 (data labeled "resolved" represents the number of e-mail resolved each day, the actual productivity). Figure 3 shows a histogram of the ratio Y_j/T_j. During this month the contact center never exceeded the upper bound $1.1T_j$ but frequently produced less than the lower bound $0.9T_j$.

Figure 4 highlights the relationship between workload and performance. Each point on the plot represents one day's data. Both axes represent a service rate, specifically, a number of e-mails per day per agent. On the horizontal axis we plot T_j/N_j, the daily rate needed from each agent to precisely meet the target (we will call this the *target rate*) given the available staff.[1] On the vertical axis we plot Y_j/N_j,

[1] Note the client determines a target volume T_j for day j, not the target *rate*. If the firm has staffed N_j workers that day, then T_j/N_j, the target rate, is the rate at which the workers would need to work to meet the target.

RIGHTSLINK

Downloaded from informs.org by [155.246.103.35] on 25 March 2017, at 23:16 . For personal use only, all rights reserved.

Figure 2 Target Volumes, Bounds, and Actual Performance in Month 1

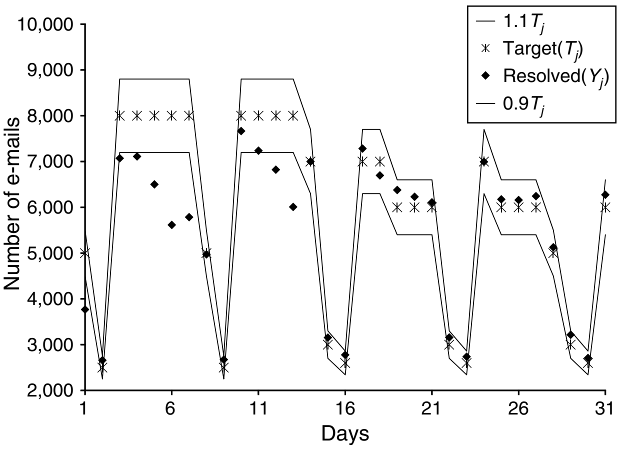
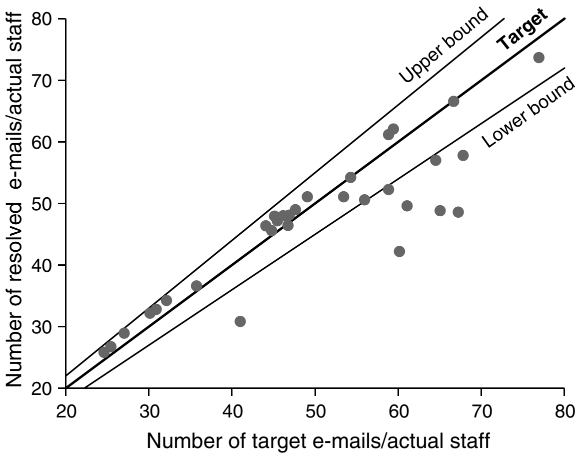


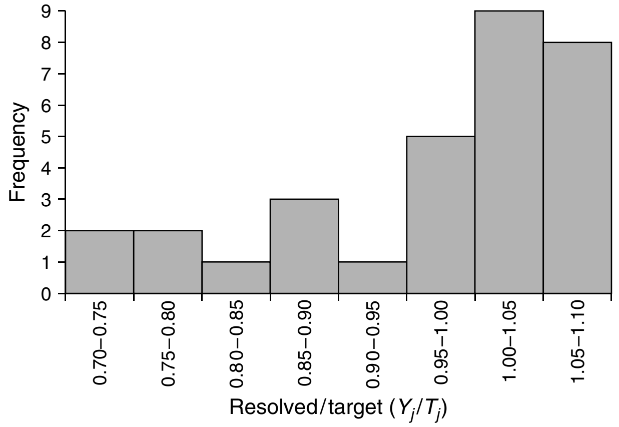
Figure 4 Target vs. Actual Daily Productivity



the *realized rate* of productivity. Therefore, points on the 45° diagonal represent days on which exactly T_j e-mails are processed, and points on the higher (lower) diagonal represent days on which the upper (lower) bounds are reached.

Figure 4 shows an extremely wide range of realized rates, from 26 e-mails per day to 74 e-mails per day. The figure strongly suggests that the most significant contributor to this variation was slow-down behavior by the agents. Of the 17 days when the target rate was less than 50 e-mails per day, only 2 saw a realized rate that fell below the target rate. Of the 14 days that the target rate was above 50 e-mails per day, 12 of the realized rates fell below the target rate. Therefore, from the plot we can make a rough estimate that the true agent capacity is a bit above 50 e-mails per day; any realization below 50 per day probably represents slow-down behavior.

Figure 3 Histogram of Productivity/Target Ratio for Month 1



Of course, such qualitative judgments are difficult to automate, and we want to formulate a model that uses all the data to make a more precise estimate of the mean capacity. We also want to estimate the variance of capacity, a quantity that our staffing model will need. Therefore, we must also examine the variation in realized capacity on days without slow-down behavior.

There are many reasons why the productivity of a fully loaded facility would work at a variety of speeds. There are environmental effects: variations due to factors such as the types of e-mails arriving that day and glitches in the information system. These effects may be present every day, are usually independent of the load, and therefore should be incorporated into our mean and variance estimates.

There may also be days on which agents put in extra effort to handle particularly large workloads; an example of such speed-up behavior may be the point on the upper right of Figure 4, which represents a target rate of 77 and a realized rate of 74. In general, the realized capacity continues to increase, on average, as target rates rise, even above the 50 e-mails per day threshold. One could in theory estimate a volume-dependent work rate, a function linking workload and "maximum" capacity, as in Diwas and Terwiesch (2007). See §5.4 for a discussion of such a model.

Downloaded from informs.org by [155.246.103.35] on 25 March 2017, at 23:16 . For personal use only, all rights reserved.

Table 1 Parameter Estimates for the Three Models of Productivity

	Parameter	Estimate (e-mails/day)	Std. err.	*p*-value	R^2
Naïve	μ	47.1	2.1	<0.0001	
	σ	119.5	10.7	<0.0001	0.434
Censored	μ	51.3	1.8	<0.0001	
(with $\varepsilon = 0.05$)	σ	95.3	13.6	<0.0001	0.827
Truncated	μ	57.0	2.6	<0.0001	
	σ	83.7	12.1	<0.0001	0.822

5.2. Three Models of Productivity

Figures 2 to 4 suggest that the center's agents did not exceed the upper limit UT_j , even when there was sufficient capacity to do so. When constructing a model of realized productivity, this information might be ignored, as in the naïve model below, or we might assume that the agents exhibited some type of slow-down or stopping behavior, as in the censored and truncated models below. All three models are standard in the econometrics literature and satisfy useful regularity conditions (see Cohen 1959 and Hayashi 2000, §§8.2 and 8.3). Given the data from Month 1, we use the SAS statistical package to fit μ and σ to the conditional distributions described below. Table 1 presents the results. In this section we will discuss the parameter estimates. In §5.3 we will discuss the fit statistics in the last column of the table.

1. *Naïve Model*: First we assume that the historical productivity is equivalent to the capacity; i.e., we assume that the total productivity of N_j agents, Y_j, is distributed according to a normal distribution with mean and standard deviation equal to μN_j and $\sigma\sqrt{N_j}$, respectively. The parameters μ and σ may be estimated by maximizing the log-likelihood function (H) over these parameters:

$$H = -M\mathrm{Ln}(\sigma) - \sum_{j=1}^{M} \frac{(Y_j - \mu N_j)^2}{2\sigma^2 N_j}, \tag{5}$$

where M is the number of days for which the historical data is collected. It is well known that (5) is maximized by setting μ equal to the sample mean and setting σ^2 equal to the sample variance (the latter is a biased estimate that can be corrected by dividing by $M-1$ instead of M). This is the logical approach if the agents on any given day work at their full capacity and there is no interaction between agent productivity and the total productivity of the vendor on a given day.

2. *Censored Model*: Now suppose that the productivity data are distributed according to a censored normal distribution. Therefore, all realizations of capacity near or above UT_j are collected on a mass point near UT_j. In this case, the log-likelihood function is

$$H = \sum_{j=1}^{M} H_j$$

$$\begin{aligned} H_j = {} & w_j \mathrm{Ln}\left(\phi\left(\frac{Y_j - \mu N_j}{\sigma\sqrt{N_j}}\right)\right) \\ & + (1 - w_j)\mathrm{Ln}\left(1 - \Phi\left(\frac{UT_j - \mu N_j}{\sigma\sqrt{N_j}}\right)\right) \end{aligned} \tag{6}$$

$$w_j = \begin{cases} 1 & \text{if } Y_j < (U - \varepsilon)T_j \\ 0 & \text{if } Y_j \geq (U - \varepsilon)T_j, \end{cases}$$

where ϕ is the standard normal PDF. The estimate is appropriate if the agents on any given day work at their full capacity but are made to stop immediately when the total daily production level is close to UT_j, the upper threshold in the SLA. Because an over-capacitated system will lead to productivity that is close to the upper bound but may not exactly reach the upper bound, we treat any observations in the range $(U - \varepsilon)T_j$ to UT_j as if productivity reached the upper bound. The choice of a value for ε depends on our assessment of how accurately the agents can hit the upper target. Note, however, that as ε approaches 0 the estimate from the censored model approaches the estimate from the naïve model because few observations are classified as days with overcapacity, the first term in H_j dominates, and (6) becomes equal to (5). As ε grows larger, a higher proportion of low-productivity observations are ignored (essentially, replaced by UT_j) and the estimate is based on less information. We chose $\varepsilon = 0.05$ because it captured a majority of the points close to UT_j but was not so large that it captured points far from the upper bound that we believed were inconsistent with the behavioral assumption of the censoredp model. The challenge of identifying an appropriate value of ε complicates the use of this estimation

Downloaded from informs.org by [155.246.103.35] on 25 March 2017, at 23:16 . For personal use only, all rights reserved.

approach and is one reason to consider the following truncated model. Another reason to consider the following model is if one believes that a substantial number of low-capacity observations are actually realizations from high-capacity days.

3. *Truncated Model*: Now suppose that the productivity data are distributed according to a truncated normal distribution; i.e., the distribution of productivity is the distribution of the capacity with the density above UT_j set to zero and the remaining density rescaled. In this case the log-likelihood function is

$$H = \sum_{j=1}^{M} H_j$$
$$H_j = \mathrm{Ln}\left(\phi\left(\frac{Y_j - \mu N_j}{\sigma\sqrt{N_j}}\right)\right) - \mathrm{Ln}\left(\Phi\left(\frac{UT_j - \mu N_j}{\sigma\sqrt{N_j}}\right)\right). \qquad (7)$$

Use of this truncated distribution implies that the agents adjust their service rates so that at the end of the day the total production level for the vendor is not above the upper threshold. The estimation procedure also relies on the assumption that the adjusted capacity above UT_j "lands" at a point below UT_j according to the likelihood specified by the original normal distribution. Note that our use of the term "truncation" is nonstandard. The traditional description of an experiment that produces a right-truncated distribution states that when a sample falls above a threshold it cannot be observed at all. Here, we do observe productivity on days when capacity exceeds UT_j, but we assume that on these days the productivity has been adjusted downward so that the original capacity is, for our purposes, unobservable. The resulting likelihood function is identical to the likelihood function of the standard right-truncated normal distribution.

We can further refine this model by assuming that any capacity above UT_j is seen as productivity adjusted downward to a point between y and UT_j, where $LT_j \le y < UT_j$. In such an approach we assume that the realized productivity in the region $y \le Y_j < UT_j$ follows the original distribution of capacity, conditional on being above y and below UT_j. This method is reasonable under the assumption that managers would not slow down so much that the lower bound LT_j might be violated. The likelihood function for this estimate is then maximized over three parameters: μ, σ, and y. We found that these refined estimates were close to those found for the log-likelihood function (7), and the resulting staffing results were close to those presented below.

Table 1 shows the results of the three estimation methods, fitted with the data from Month 1. Note that the standard deviations are quite high, relative to the mean capacity. This is due, in part, to real variations in the skills of agents and the content of e-mails. Some part of the high estimated variance may also serve as a proxy for positive correlation among agents, for significant correlation increases the variability of daily productivity. As we mentioned above, the data are not sufficiently fine grained to provide reliable correlation estimates, but we show below that a model based on the independent service-time assumption generates staffing levels that work well in practice.

In general, the sizes of the differences among the estimates depend on the staffing levels that generate the data. If all days are understaffed, then all three methods generate similar (correct) results. If there is overstaffing, then the naïve model, which ignores the upper bound, produces a low—and incorrect—estimate of the average capacity. It may, however, be difficult to choose between the other two models.

5.3. Selecting a Model

There were three types of information to be considered when choosing between the censored and truncated models: statistical measures of fit, observations of the work environment, and the performance of each model in a prototype staffing algorithm. In this section we describe all three sources of information. In the end, however, our analysis did not point definitively toward one model or another. We chose the truncated model, although good arguments can certainly be made for the censored model as well.

First, to measure statistical fit we calculate R^2, the fraction of variance explained by each model. Let $\overline{Y}$ be the overall average productivity across all days, a statistic that ignores all the other information at hand, including the staffing levels N_j. Let $\widehat{Y}(N_j \mid \hat{\mu}, \hat{\sigma})$ be a model's expected (mean) productivity for N_j agents, given parameter estimates $\hat{\mu}$ and $\hat{\sigma}$. Then, for each model we calculate

$$R^2 = 1 - \frac{\sum_{j=1}^{M} (Y_j - \widehat{Y}(N_j \mid \hat{\mu}, \hat{\sigma}))^2}{\sum_{j=1}^{M} (Y_j - \overline{Y})^2}.$$

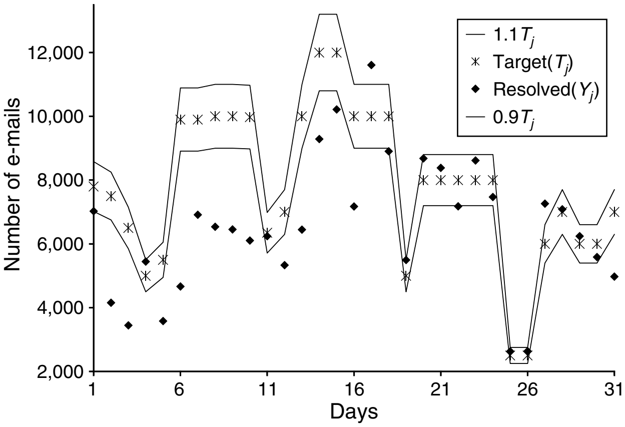
Downloaded from informs.org by [155.246.103.35] on 25 March 2017, at 23:16 . For personal use only, all rights reserved.

Table 1 shows that the R^2 for the two models is virtually identical. When we refitted the models using data from three other months, we again found that the two models produced virtually identical values of R^2, with values of R^2 always within 0.006 of each other.

Now we discuss how our field observations compare with the behavioral assumptions that motivate the censored and truncated models. On days when the center was overstaffed or productivity was unusually high, the managers could avoid the upper limit by either (i) asking all agents to log off the system once 110% of the target was achieved or (ii) gradually slowing down the agents when they observed that the agents were working fast and were likely to reach 110% of the target. If (i) happened frequently, then we would see a large mass point near 1.1 in the histogram, Figure 3, and a pure censored model would be appropriate. On days with event (ii), agents may be able to reach the upper bound, but there may be other days when agents could have reached the upper bound but fell short because of the variability in service times while they were working at a slower pace. Productivity on such days would be distributed randomly below the upper bound, behavior that could be approximated either by the adjusted censored distribution (using ε) or by the truncated distribution.

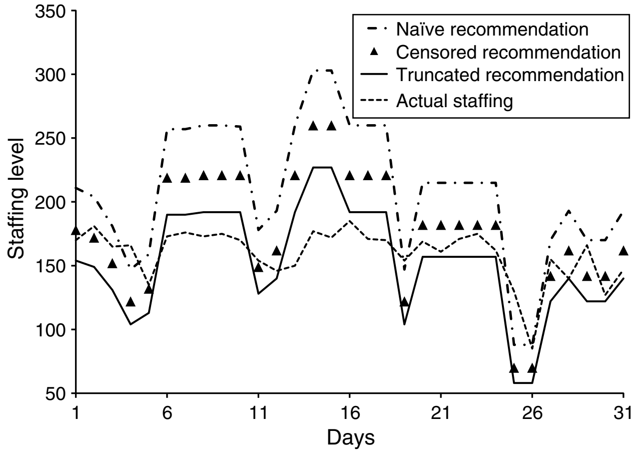
As we described in §2, our own observations are consistent with event (ii). During interviews, general managers further confirmed that they do slow down agents when they foresee that the productivity will be more than 110% of the target at the current speed. This strategy reduces the idle time of agents and thereby increases the utilization artificially, behavior consistent with the manager's incentives. The interviews also revealed another reason that the vendor prefers to slow the agents down rather than letting them finish their daily targets earlier. The vendor provides transportation service to and from the homes of its agents and uses a sophisticated algorithm to plan the transportation schedule one week in advance. Therefore, the vendor prefers that its agents leave the workplace at their scheduled times, a goal accomplished on low-demand/high-capacity days by slowing the agents down.

Figure 5 Target E-Mail Volume and Actual Performance in Month 2



This information, however, does not point definitively toward the censored or truncated model. To gather further evidence we developed a prototype staffing algorithm and retrospectively applied the algorithm using each of the estimates. Specifically, using each of the estimates and historical data from Month 2, we generated staffing levels and compared those recommendations with the vendor's staffing decisions. Figure 5 displays the actual performance of the vendor during Month 2, and Figure 6 shows the recommended staffing levels, given each estimate. Staffing models based on both the truncated and censored methods seemed to produce reasonable recommendations. On days of the month when the vendor fell short of the target (e.g., Days 5–10 in Figure 5), staffing models using estimates from the censored and truncated methods suggested slightly higher staffing levels, with the truncated estimate

Figure 6 Actual Staffing and Recommendations for Month 2, Given Each Estimate



Downloaded from informs.org by [155.246.103.35] on 25 March 2017, at 23:16 . For personal use only, all rights reserved.

leading to more conservative staffing than the censored estimate. On days when the vendor was able to meet the target (e.g., Days 20–25), both estimates led to staffing levels close to those chosen by the vendor, again with the line based on the truncated method being more conservative. This provides evidence in favor of the truncated model, because on days when the vendor met the target it should not be necessary to raise staffing levels, and it may make sense to lower them.

In the end we remain uncertain as to whether the censored or truncated model is a better description of reality, whether the productivity of slowing agents consistently approaches the upper bound (the censored model) or is more widely distributed (the truncated model). The truth is probably somewhere in between. We chose to implement the truncated model because of the evidence provided by the prototype. In addition, the truncated model is relatively simple to implement, but the censored model requires the identification of an appropriate ε, a judgment call that is difficult to automate.

5.4. Models with Additional Factors

After creating and implementing our estimation model and staffing algorithm (see the next section), we examined more sophisticated methods for analyzing the data. These models incorporate factors such as speed-up behavior and variation in model parameters by day of week and month.

As mentioned above, Figure 4 suggests that agents may speed up on days when it is likely that the overall capacity may fall short of the lower bound, LT_j. We now modify the truncated model to incorporate such behavior, examine whether the data provide strong evidence of such behavior, and discuss how such a model might be used in a staffing algorithm. In this section we describe one model of speed-up behavior, and in Appendix A we present two additional models. The insights generated by all three speed-up models are similar.

Let $\bar{\mu}_j$ be the e-mail processing rate necessary to meet the lower bound, given T_j and N_j. Therefore, $\bar{\mu}_j = LT_j/N_j$. For this speed-up model, we assume that the productivity data follow a truncated normal distribution with mean μ and standard deviation σ when $\bar{\mu}_j \le \mu$, i.e., on days when the vendor is sufficiently staffed so that its agents can meet the lower bound when working at the standard service rate μ. On days when $\bar{\mu}_j > \mu$, the vendor is understaffed, so the agents may need to speed up to meet the lower bound of the SLA. We assume that the greater the capacity deficit, the greater the speed-up behavior, up to some limit γ. Specifically, we assume that the speedup is linear in the fractional *capacity shortfall*, $(\bar{\mu}_j - \mu)/\bar{\mu}_j$, and has an upper bound γ. Therefore, for days when $\bar{\mu}_j > \mu$, productivity follows a truncated normal distribution with mean $\mu + \gamma(\bar{\mu}_j - \mu)/\bar{\mu}_j$ and standard deviation σ. We estimate (μ, σ, γ) by maximizing the log-likelihood function,

$$H = \sum_{j=1}^{M} H_j$$

$$H_j = \left(\mathrm{Ln}\left(\phi\left(\frac{Y_j - \hat{\mu}_j N_j}{\sigma\sqrt{N_j}}\right)\right) - \mathrm{Ln}\left(\Phi\left(\frac{UT_j - \hat{\mu}_j N_j}{\sigma\sqrt{N_j}}\right)\right).\right)$$

$$\hat{\mu}_j = \mu + \gamma \frac{(\bar{\mu}_j - \mu)^+}{\bar{\mu}_j}.$$

Table 2 shows the maximum likelihood estimates from the model, given the data from Month 1. First, note that the value of R^2 has increased from 0.82 for the pure truncated model to 0.86. Given the extra degree of freedom provided by γ, the speed-up model generates expected production levels that provide a better fit for the observed data. In Table 2 the result $\mu = 55$ suggests that the base capacity under the speed-up model is close to the capacity suggested by the simple truncated model: 55 versus 57 e-mails per day. The estimate for the maximum speed-up rate, $\gamma = 56$, has an extremely large amount of uncertainty associated with it; a confidence interval using 2 standard errors includes maximum speedup estimates of both 0 and 112 e-mails per day. If we take the point estimate of 56 per day at face value, however, it implies that the capacity of any server can roughly double, from 55 to a maximum of approximately $(55 + 56)$ as the staffing level falls.

Table 2 **Parameter Estimates for the Speed-Up Model**

Parameter	Estimate (e-mails/day)	Std. err.	*p*-value	R^2
μ	55.0	3.3	<0.0001	
σ	75.6	13.0	<0.0001	0.86
γ	56.4	29.5	0.065	

Downloaded from informs.org by [155.246.103.35] on 25 March 2017, at 23:16 . For personal use only, all rights reserved.

A doubling of capacity may seem unrealistic, but such large shortfalls in capacity are not included in the data set, so this model is not appropriate (and we would advise against it being applied) for such an extreme situation. For example, during Month 1, the day with the lowest predicted capacity, relative to the target, had a capacity shortfall $(\bar{\mu}_j - \mu)/\bar{\mu}_j$ of 20%, so that the model's predicted capacity on that day would be $55 + 56 * (0.2) = 66$. This seems to be a reasonable increase above the base speed of 55 e-mails per day.

In general, these results suggest that there may be a speed-up effect, but there is great uncertainty about the size of that effect. We found similar results using two alternate models (see Appendix A) as well as from a model that included speed-up adjustments to σ as well as to μ. We also found similar results when the model presented above was expanded to include data from multiple months (see below). We believe that significantly more data would be needed to generate precise estimates of the speed-up effect.

Should more evidence of speed-up behavior be found, the implications for staffing are unclear. If managers believe that it is stressful for agents to speed up and that frequent speedups may lead to higher turnover, or if frequent speedups may harm quality, then setting staffing levels according to the base rate (e.g., $\mu = 55$) is quite reasonable. If, however, managers believe that taking advantage of speed-up behavior is worth the potential cost, they may choose to raise the capacity estimate by some amount based on γ and intentionally understaff.

In addition to these relatively simple speed-up models, we also formulated more complex models that attempt to control for speed-up, day-of-week, and month-to-month effects on our estimates of capacity. We find that there is a significant variation in the baseline capacity across months. This confirms the facility managers' intuition; they suspected that service rates changed over time because of changes in e-mail content, agent learning curves, and turnover. In fact, such changes in capacity increase the value of an automated estimation procedure that reestimates parameters as time passes.

These more complex models, however, did not provide strong evidence for speed-up and day-of-week effects. A model with day-of-week parameters improved the fit of the model to the data (as measured by R^2), but the additional parameters were not statistically significant, either individually or as a group. This is not entirely surprising, for it is difficult to disentangle the effects of regular changes in demand across weekdays and speed-up behavior. In particular, we find that certain weekdays tend to have both high demand and high productivity, so it is impossible with these data to determine whether the higher productivity is caused by increased capacity that is unique to those days or to speed-up behavior caused by increased demand.

There are a variety of other models that might be considered. In the spirit of the speed-up model presented above, one could also imagine a more complex truncated model that takes a range of slow-down behaviors into account. For example, managers may be unsure of the likelihood of exceeding the threshold on days when they are close to the upper bound, so they may slow down unnecessarily. Such a model would apply a "slow-down factor" on days where the total capacity would push productivity near the upper threshold.

Finally, all of the estimation procedures described in this paper assume that all workers have the same capacity. In some environments there may be heterogeneity in worker capacity, and the type of worker scheduled could vary across days with different workloads. This might be true because managers would tend to schedule the faster workers on most days and only use slower workers when they were needed on days with high loads. In our environment, however, the managers are responsible for setting staff levels and worker scheduling is handled by a central facility (see §2). The centralized worker staffing algorithm does not incorporate any criteria related to individual performance.

6. Implementation and Initial Results

The vendor implemented the algorithm in a spreadsheet model that contained two components: a parameter estimation module and a staffing requirements module. Because e-mail content and agent skills can change over time, the parameter estimation module updates the parameter estimates twice each month by maximizing the likelihood function of the truncated model, conditioned with the latest historical data. The staffing requirements module then uses the updated

Downloaded from informs.org by [155.246.103.35] on 25 March 2017, at 23:16 . For personal use only, all rights reserved.

Table 3 Performance During Pre- and Postimplementation Sample Periods

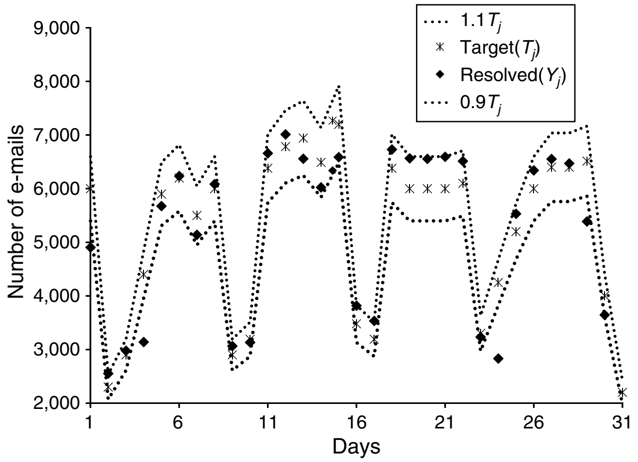
	Pre	Post	Postnaïve (estimated performance using naïve estimates)
Number of days	59	61	61
Average target (T_j)	5,115	4,603	4,603
Coefficient of variation of targets	0.35	0.32	0.32
No. e-mails resolved	294,164	281,851	290,566
% of days above lower bound	83	92	99.8
E-mail shortfall below lower bound (as a % of target)	2.2	1.1	0.13
Estimated average load factor	0.78	0.81	0.72

parameter estimates as model inputs when calculating the recommend staffing levels.

In Table 3 we compare the performance of the contact center before and after implementation of the model. The postimplementation data represent performance during the first two months after implementation, and the preimplementation data reflect performance during two months before implementation with similar levels of target demand, volume, and variability as in the two postimplementation months. For illustration, in Figure 7 we show one month of postimplementation data. The last column of Table 3 shows the estimated performance of the center if we had based our staffing algorithm on a naïve estimate of capacity rather than on the estimate generated by the truncated model (below we provide more details of how the numbers in this column were derived).

Table 3 shows that the percentage of days above the lower bound, $Y_j > LT_j$, rose from 83% to 92%. A *t*-test shows that this change is significant at a 0.075 level of significance (see D'Agostino et al. 1988, for a discussion of the suitability of the *t*-test here). If Month 2 is included in the preimplementation data, then the level of significance of the change is 0.001. In the postimplementation data, the 92% rate is below the desired 95% rate used in the model; there may have been a variety of reasons for this shortfall, including (i) poor performance of the shift-scheduling algorithm (our algorithms only generated the staffing requirements and did not schedule shifts), (ii) an unusual number of agent no-shows that reduced staffing below the plan, (iii) sudden changes in model parameters that were not accurately captured by the estimation updates, and (iv) other random variations, such as unexpectedly large service times. The table also shows the size of the e-mail shortfall as a percentage of the sum of the targets, $\sum(LT_j - Y_j)^+/\sum T_j$. This quantity fell by 50% after model implementation.

Figure 7 Target E-Mail Volume and Actual Performance During Month 3 (Postimplementation)



It is possible that any improvement in satisfying the lower bound is the result of overstaffing: with extra agents the center can easily meet the lower bounds and then conform to the upper bound by slowing down the agents. Therefore, we wish to confirm that our algorithm is recommending cost-effective staffing levels. The comparison of staffing levels is complicated by possible changes over time in the underlying capacity. To control for these changes, we first estimated the agents' preimplementation capacity by applying the truncated model to the two months of preimplementation data. We then reestimated capacity using the postimplementation data. Given these two capacity estimates, μ_{pre} and μ_{post}, the daily targets T_j, and the daily staffing levels N_j, we calculated the estimated daily load factors, $T_j/(N_j\mu_i)$, where $i = pre$ or $post$. The last line of Table 3 shows these load factors, averaged over all days before and after implementation. We see that the postimplementation load factor is slightly

Downloaded from informs.org by [155.246.103.35] on 25 March 2017, at 23:16 . For personal use only, all rights reserved.

higher, indicating that the performance gains were not caused by relative increases in staffing levels.

Finally, to determine whether the improvement in performance is a result of better capacity estimates or improved staffing recommendations, we reapplied our staffing module to the postimplementation period, but used a naïve capacity estimate rather than the estimate from the truncated model. The results are shown in the last column of Table 3. We have seen that naïve capacity estimates tend to be low (Table 1); therefore, it is not surprising that the recommended staffing levels rose. Specifically, when using the naïve estimates, the recommended number of staff hours was 13% higher than the actual staffing, lowering the load factor from 0.81 to 0.72. Then, assuming that the capacity estimate from the truncated model is correct and that agents follow the behavioral assumptions that underlie the truncated model, we calculated the performance levels for the naïve staffing recommendation. We found that the higher staffing levels would have produced a service level of 99.8% rather than the target 95%. Therefore, the naïve estimate would have led to significant overstaffing, and the observed performance improvement can be ascribed to improvements in both the new capacity estimation procedure as well as the staffing recommendations.

7. Analysis of the Model Formulation

Here we compare the vendor's expected performance when using the profit optimization formulation (1) and the service-level constraint formulation (2) for determining staffing. For simplicity, in (1) we assume that the vendor faces a linear penalty function: $G(x) = px$, where p is the penalty per e-mail paid by the vendor for a shortfall below LT_j. Note that when $G(x)$ is linear, the objective function in (1) can be written analytically in terms of the PDF and CDF of the normal distribution (see Appendix B):

$$\begin{aligned}\pi_v(N) &\equiv \mathrm{E}[r\min(C(N), UT) - p((LT - C(N))^+) - SN]\\ &= r\mu N - r\sigma\sqrt{N}H(z_U) - pLT\\ &\quad + p\mu N - p\sigma\sqrt{N}H(z_L) - SN, \qquad (8)\end{aligned}$$

where

$$Z_U = (UT - \mu N)/(\sigma\sqrt{N}), \quad z_L = (LT - \mu N)/(\sigma\sqrt{N})$$
$$\text{and} \quad H(z) = [\phi(z) - z(1 - \Phi(z))].$$

When π_v is evaluated over the positive real numbers, π_v is concave, given our data (see Appendix B). Therefore, π_v exhibits decreasing differences in integer values of N, so the optimal staffing level can be found by simply increasing N until the marginal change is less than zero.

The particular parameters we use in the following analysis correspond to our vendor's cost and revenue structure. The vendor's facility is in India, where it incurs a staffing cost of roughly INR 10,000 per month per agent, which is approximately $11 per agent per day (using an exchange rate of 45 INR per $1 and 20 working days in a month). Based on our interviews with the managers, we estimate that the client pays the vendor approximately 30 cents per e-mail, although the per-e-mail penalty for a shortfall may be larger than this because of both explicit and implicit penalties.

First we find the implied values of p that would generate the same staffing levels from (1) that are generated by the vendor using (2) with $\alpha = 0.95$. Put another way, *if* the vendor is rational when it uses (2), what is the implied shortage penalty in (1)? This is a descriptive rather than a normative analysis, for we are attempting to infer p from the data. Figure 8 presents bounds on the implied values of p for each of the target e-mail volumes during Month 3 (Figure 7). Here, any per-e-mail shortage penalty p within the range $p_{\min}$ to $p_{\max}$ will lead to the same staffing levels from (1) as from (2), where (2) is solved with $\alpha = 0.95$. We find that as the target increases, the implied per-e-mail penalty decreases. This effect is driven by economies of scale: in smaller systems it is more expensive on a per-customer basis to satisfy high service levels. Therefore, the implied penalty motivating the service level of $\alpha = 0.95$ must be higher in small systems. Because it is unlikely that the per-e-mail shortage penalty varies significantly as the daily target varies, it seems clear that the vendor is making suboptimal staffing decisions.

Now we examine the magnitude of the loss incurred by the vendor when it uses constraint formulation (2) to solve its staffing problem rather than finding the optimal staffing from formulation (1). Motivated by Figure 8, we examine this loss for values of p in the range $[\$0.20, \$2.50]$. In Figure 9 we show the percentage loss when using $\alpha = 0.95$ in (2).

Downloaded from informs.org by [155.246.103.35] on 25 March 2017, at 23:16 . For personal use only, all rights reserved.

Figure 8 Ranges of Penalties p in (1) That Lead to Same Staffing Level as (2) with $\alpha = 0.95$

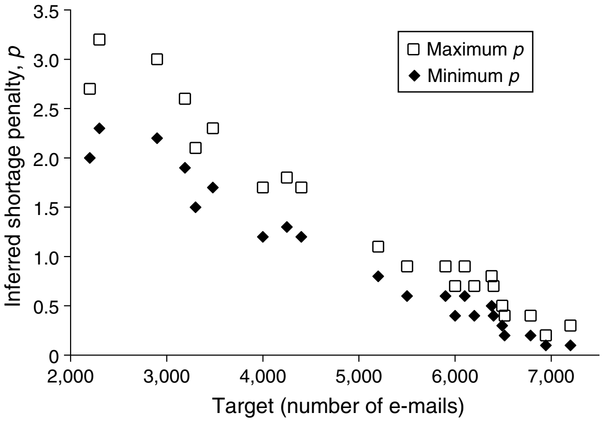
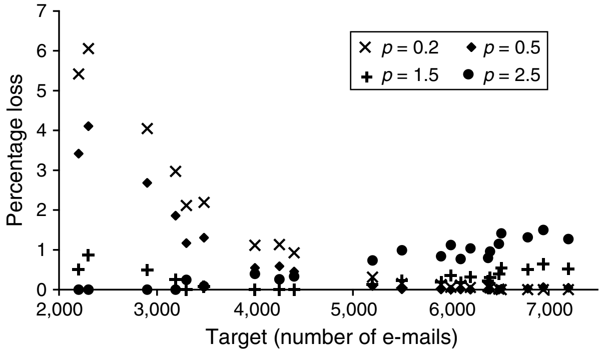


Figure 9 Percentage Loss When Solving (2) and Using $\alpha = 0.95$



Again, these losses were evaluated for each of the target e-mail volumes shown in Figure 7. Generally, the loss is low (less than 2%), although the cost can climb rapidly for low volumes and low values of p. For the lowest values of p, the constraint $\alpha = 0.95$ is too tight, and this error is particularly expensive in small systems. Based on this plot, we advise that if the vendor's volumes frequently vary both above and below 3,500 e-mails per day, then the vendor should conduct additional analyses to specify p and staff according to the optimization problem (1) rather than the simpler constraint problem (2). If e-mail volume is consistently above 3,500 per day, however, the service constraint model is robust.

8. Conclusion

In this paper we describe methods to estimate capacity and determine staffing levels for a large e-mail contact center. Given productivity data distorted by limited demand and internal incentives, we find that the productivity data can be effectively described by a truncation of the underlying capacity distribution, and we design our estimation procedure accordingly. Postimplementation results show that the procedure performs well. We then compare profit-maximization and service-level constraint formulations of the model and show that a service-level constraint formulation can be quite robust in this application as long as the volume is sufficiently high and the implicit cost parameters are not extremely low. Using more complex statistical models, we also examine whether agents exhibit speed-up behavior when capacity is low relative to demand and whether capacity varies by day of week. The results generated by these models are ambiguous, and we would recommend additional data collection and analysis to clarify these issues.

One limitation of our model is the assumption that service times are independent. With agent-level data one could examine estimation procedures that take correlation among agents into account. Unfortunately, such detailed data were not available to us.

An alternate method for estimating service times is a controlled experiment. For example, one might randomly select a large group of servers, provide them with a sufficient amount of work, and observe their service times. Repetition of such an experiment over a period of time can produce estimates of the mean and variance of agent capacity (see Neter et al. 1996). This can be expensive, however, for the experiments must be repeated as workforce attributes and e-mail content changes. In addition, if the servers know that they are being observed, their behavior may change, so results of the experiment may not accurately reflect the long-term capacity of actual agents under real-world conditions.

Another method for estimating agent capacity is to use the naïve model with the productivity data that do not include days during which the agents slowed down. The approach would require managers to identify the days when agents slowed down and therefore cannot be automated, although the truncated model described in this paper has been encoded in software. Alternatively, because it is likely that the first half of

Downloaded from informs.org by [155.246.103.35] on 25 March 2017, at 23:16 . For personal use only, all rights reserved.

most days will not exhibit significant slowing behavior, the vendor can collect hourly data and estimate capacity from the productivity of agents during the first half of the day. Converting first-half capacity into daily capacity, however, may require addition analysis of the hourly productivity data to capture the effects of agent fatigue.

Acknowledgments

The authors thank Sanjog Misra and Vithal Sanapala. In addition, the authors received helpful comments from Gerard Cachon, three anonymous reviewers, and an anonymous associate editor.

Appendix A

This appendix describes two additional models that incorporate speed-up behavior. Table A.1 shows the maximum likelihood estimates from the following speed-up models, given the data from Month 1.

Speed-Up Model 1: Double Truncation

In this model we assume that managers would never allow the agents' collective productivity to fall below some level. The simplest method for modeling this effect is to truncate the distribution at some point below LT, as well as at UT. We conducted sensitivity analysis on the left-truncation point. Clearly, a very low left-truncation point near zero has no impact on the parameter estimates. The highest reasonable truncation point, and the one with the greatest impact on the estimates, is at the lowest observed ratio Y/T (any higher left-truncation point would not be able to explain that observation). In the data from month 1, the lowest Y/T ratio was 0.7. Table A.1 shows that the parameter estimates, given left truncation at 0.7, are extremely close to the original truncated parameter estimates shown in Table 1.

Speed-Up Model 2: All or Nothing

This model offers a more explicit description of speed-up behavior. For each day we compute the capacity, given staffing N_j and standard rate μ, and determine whether the capacity is sufficient to meet the lower bound $0.9T_j$. If it is (if $\mu N_j \geq 0.9T_j$), then we assume that the realized productivity is distributed as a truncated normal distribution with mean μ and standard deviation σ. If the facility is understaffed and the standard capacity is not sufficient to meet the lower bound $\mu N_j < 0.9T_j$), we assume that the realized productivity is distributed as a truncated normal distribution with a mean $\theta\mu$ and a standard deviation σ. The factor θ for such days adds an extra degree of freedom in our estimation model and is intended to capture any speed-up effect. We call this the "all-or-nothing" model because either agents speed up according to the multiplier θ or they do not.

We estimate (μ, σ, θ) by maximizing the log-likelihood H,

$$H = \sum_{j=1}^{M} H_j$$

$$H_j = \left(\mathrm{Ln}\left(\phi\left(\frac{Y_j - \mu N_j}{\sigma\sqrt{N_j}}\right)\right) - \mathrm{Ln}\left(\Phi\left(\frac{UT_j - \mu N_j}{\sigma\sqrt{N_j}}\right)\right).\right)\delta_j + \left(\mathrm{Ln}\left(\phi\left(\frac{Y_j - \theta\mu N_j}{\sigma\sqrt{N_j}}\right)\right) - \mathrm{Ln}\left(\Phi\left(\frac{UT_j - \theta\mu N_j}{\sigma\sqrt{N_j}}\right)\right).\right)(1-\delta_j)$$

$$\delta_j = \begin{cases} 1 & \text{if } LT_j/N_j \leq \mu \\ 0 & \text{if } LT_j/N_j \leq \mu. \end{cases}$$

In Table A.1 we see that the parameter values that maximize H are $\mu = 55.0$, $\sigma = 74.2$, and $\theta = 1.066$. That is, on days when the facility is understaffed, agents speed up by 6.6%. We see that the value $\mu = 57$ estimated from the truncated model in §5.2 is enveloped by the estimate of capacity $\mu = 55$ and the estimate of speedup capacity $\theta\mu = 58.6$. This is consistent with the intuition that the estimate of capacity from the model that does not account for speedup should be a value higher than the true capacity but lower than the speeded-up capacity.

Although the results are statistically significant, the practical significance is less clear. First, both the base capacity μ and the increased capacity $\theta\mu$ are quite close to the capacity estimate from the original truncated model, and therefore staffing recommendations based on either capacity will be close to staffing recommendations based on the truncated model. In addition, it is not clear whether it is advisable to consistently staff under the assumption that agent capacity is μ or $\theta\mu$. Finally, the "all-or-nothing" aspect of this model is unsatisfying; one would not expect the same speed-up behavior whether the facility is a bit or extremely understaffed. Therefore, a model in which the speed-up factor is scaled by the capacity shortfall, such as the one described in §5.4, may be more reasonable.

Table A.1 **Parameter Estimates for the Speed-Up Models**

	Parameter	Estimate (e-mails/day)	Std. err.	*p*-value
Double truncation	μ	55.7	2.7	<0.0001
(at 0.7 and 1.1)	σ	86.0	13.7	<0.0001
All or nothing	μ	55.0	2.9	<0.0001
	σ	74.2	10.7	<0.0001
	θ	1.066	0.076	<0.0001

Appendix B

That the objective function in (1) can be written as (8) follows from the well-known identity, $\mathrm{E}[(D-y)^+] = \hat{\sigma}[\phi(z) - z(1-\Phi(z))]$, where y is a constant, D is distributed according

Downloaded from informs.org by [155.246.103.35] on 25 March 2017, at 23:16 . For personal use only, all rights reserved.

to the normal distribution with mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$, and $z=(y-\hat{\mu})/\hat{\sigma}$ (see Porteus 2002, p. 12).

We now show that π_v is concave under extremely mild conditions, such as conditions satisfied by our data. Taking the first derivative of $\pi_v(N)$ w.r.t. to N, we find

$$\begin{aligned}\pi_v'(N) &= r\left[\mu-\frac{\sigma}{2\sqrt{N}}H(z_U)-\sigma\sqrt{N}\frac{\partial H(z_U)}{\partial N}\right]\\ &\quad +p\left[\mu-\frac{\sigma}{2\sqrt{N}}H(z_L)-\sigma\sqrt{N}\frac{\partial H(z_L)}{\partial N}\right]-S,\\ \frac{\partial H(z_U)}{\partial N} &= \frac{\partial H(z_U)}{\partial z_U}\frac{\partial z_U}{\partial N},\quad \frac{\partial H(z_U)}{\partial z_U}=-(1-\Phi(z_U)),\\ \frac{\partial z_U}{\partial N} &= -\frac{1}{2N}\left[z_U+\frac{2\mu\sqrt{N}}{\sigma}\right],\\ \frac{\partial H(z_L)}{\partial N} &= \frac{\partial H(z_L)}{\partial z_L}\frac{\partial z_L}{\partial N},\\ \frac{\partial H(z_L)}{\partial z_L} &= -(1-\Phi(z_L)),\quad \text{and}\\ \frac{\partial z_L}{\partial N} &= -\frac{1}{2N}\left[z_L+\frac{2\mu\sqrt{N}}{\sigma}\right].\end{aligned}$$

Thus, we have

$$\begin{aligned}\frac{\partial H(z_U)}{\partial N} &= \frac{\partial H(z_U)}{\partial z_U}\frac{\partial z_U}{\partial N}\\ &= (1-\Phi(z_U))\frac{1}{2N}\left[z_U+\frac{2\mu\sqrt{N}}{\sigma}\right]\quad\text{and}\\ \frac{\partial H(z_L)}{\partial N} &= \frac{\partial H(z_L)}{\partial z_L}\frac{\partial z_L}{\partial N}\\ &= (1-\Phi(z_L))\frac{1}{2N}\left[z_L+\frac{2\mu\sqrt{N}}{\sigma}\right].\end{aligned}$$

Therefore,

$$\pi_v'(N)=rJ(z_U)+pJ(z_L)-S,\quad\text{where}$$

$$J(z)=\left[\mu\Phi(z)-\frac{\sigma}{2\sqrt{N}}\phi(z)\right].$$

The second derivative of $\pi_v(N)$ is

$$\begin{aligned}\pi_v''(N) &= r\frac{\partial J(z_U)}{\partial z_U}\frac{\partial z_U}{\partial N}+r\frac{\partial J(z_U)}{\partial N}\\ &\quad +p\frac{\partial J(z_L)}{\partial z_L}\frac{\partial z_L}{\partial N}+p\frac{\partial J(z_L)}{\partial N},\end{aligned}$$

where

$$\begin{aligned}\frac{\partial J(z_U)}{\partial z_U} &= \mu\phi(z_U)+\frac{\sigma}{2\sqrt{N}}z_U\phi(z_U),\\ \frac{\partial J(z_L)}{\partial z_L} &= \mu\phi(z_L)+\frac{\sigma}{2\sqrt{N}}z_L\phi(z_L),\\ \frac{\partial J(Z_U)}{\partial N} &= \frac{\sigma}{4N\sqrt{N}}\phi(z_U)\quad\text{and}\\ \frac{\partial J(Z_L)}{\partial N} &= \frac{\sigma}{4N\sqrt{N}}\phi(Z_L).\end{aligned}$$

Therefore,

$$\begin{aligned}\pi_v''(N) &= \frac{1}{4\sigma N^2\sqrt{N}}[r\phi(z_U)(N\sigma^2-(\mu N+UT)^2)\\ &\quad +p\phi(z_L)(N\sigma^2-(\mu N+LT)^2)].\end{aligned}$$

Because $N\sigma^2-(\mu N+LT)^2 > N\sigma^2-(\mu N+UT)^2$, $r>0$, $p>0$, $\phi(z_U)>0$, and $\phi(z_L)>0$, $r\phi(z_U)(N\sigma^2-(\mu N+UT)^2)+p\phi(z_L)(N\sigma^2-(\mu N+LT)^2)<(r\phi(z_U)+p\phi(z_L))(N\sigma^2-(\mu N+LT)^2)$.

Therefore, a sufficient condition for the concavity of $\pi_v(N)$ is

$$N\sigma^2-(\mu N+LT)^2<0. \tag{B1}$$

By solving the equation $N\sigma^2-(\mu N+LT)^2=0$ using the quadratic formula, we find that the equation has nonreal (complex) roots if and only if $\sigma^2<4\mu LT$. Therefore, $\sigma^2<4\mu LT$ is a sufficient condition for (B1).

To determine whether this condition is met by our data, consider the approximation $LT\approx N\mu$, so that (B1) is roughly equivalent to the following condition on $C_v=\sigma/\mu$, the coefficient of variation of each agent's daily capacity:

$$C_v<2\sqrt{N}. \tag{B2}$$

In our data, the optimal value of N never falls below 40 servers, so (B2) only requires that coefficient of variation to be less than 12. In the data, the coefficient of variation is consistently less than 2, so the condition is easily satisfied.

References

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, S. Zeltyn, L. Zhao, S. Haipeng. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* **100** 36–50.

Cachon, G., F. Zhang. 2007. Obtaining fast service in a queueing system via performance-based allocation of demand. *Management Sci.* **53**(3) 408–420.

Cohen, A. C. 1959. Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics* **1**(3) 217–237.

D'Agostino, R. B., W. Chase, A. Belanger. 1988. The appropriateness of some common procedures for testing the equality of two independent binomial populations. *Amer. Statistician* **42**(3) 198–202.

Diwas, K. C., C. Terwiesch. 2007. The impact of work load on productivity: An econometric analysis of hospital operations. Working paper, The Wharton School, Philadelphia.

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

Goodale, J. C., G. M. Thompson. 2004. A comparison of heuristics for assigning individual employees to labor tour schedules. *Ann. Oper. Res.* **128**(1–4) 47–63.

Harris, C. 1966. Queues with state-dependent stochastic service rates. *Oper. Res.* **15**(1) 117–130.

Hayashi, F. 2000. *Econometrics*. Princeton University Press, Princeton, NJ.

Downloaded from informs.org by [155.246.103.35] on 25 March 2017, at 23:16 . For personal use only, all rights reserved.

Keblis, M. F., M. Chen. 2006. Improving customer service operations at Amazon.com. *Interfaces* **36**(5) 433–445.

Lee, D. K. K., S. A. Zenios. 2007. Evidence-based incentive systems with an application in health care delivery. Working paper, Stanford University, CA.

Neter, J., M. H. Kutner, C. J. Nachtsheim, W. Wasserman. 1996. *Applied Linear Statistical Models*. McGraw-Hill, New York.

Olivares, M. O., C. Terwiesch, L. Cassorla. 2008. Structural estimation of the newsvendor model: An application to reserving operating room time. *Management Sci.* **54**(1) 41–55.

Parkinson, D. F. 1955. Parkinson's law. *Economist* **19**(November) 635–637.

Porteus, E. L. 2002. *Foundations of Stochastic Inventory Theory*. Stanford Business Books, Stanford, CA.

Ross, A. M., J. G. Shanthikumar. 2005. Estimating effective capacity in Erlang loss systems under competition. *Queueing Systems* **49** 23–47.

Schruben, L., R. Kulkarni. 1982. Some consequences of estimating parameters for the M/M/1 queue. *Oper. Res. Lett.* **1**(2) 75–78.

Schultz, K. L., D. C. Juran, J. W. Boudreau. 1999. The effects of low inventory on the development of productivity norms. *Management Sci.* **45**(12) 1664–1678.

Schultz, K. L., D. C. Juran, J. W. Boudreau, J. O. McClain, L. J. Thomas. 1998. Modeling and worker motivation in JIT production systems. *Management Sci.* **44**(12) 1595–1607.

Schweitzer, M., G. Cachon. 2000. Decision bias in the newsvendor problem: Experimental evidence. *Management Sci.* **46**(3) 404–420.