Accounting for the Multi-Period Impact Of Service When Determining Employee Requirements for Labor Scheduling

Gary M. Thompson

David Eccles School of Business

University of Utah

1655 East Campus Center Drive

Salt Lake City, UT 84112

Published in *Journal of Operations Management* (1993), 11, 269-287.

Author Note: The author wishes to express his gratitude for the helpful comments and suggestions of an anonymous Associate Editor and three anonymous referees.

Abstract

Providing good customer service, inexpensively, is a problem commonly faced by managers of service operations. To tackle this problem, managers must do four tasks: forecast customer demand for the service; translate these forecasts into employee requirements; develop a labor schedule that provides appropriate numbers of employees at appropriate times; and control the delivery of the service in real-time.

This paper focuses upon the translation of forecasts of customer demand into employee requirements. Specifically, it presents and evaluates two methods for determining desired staffing levels. One of these methods is a traditional approach to the task, while the other, by using modified customer arrival rates, offers a better means of accounting for the multi-period impact of customer service. To calculate the modified arrival rates, the latter method reduces (increases) the actual customer arrival rate for a period to account for customers who arrived in the period (in earlier periods) but have some of their service performed in subsequent periods (in the period). In an experiment simulating 13824 service delivery environments, the new method demonstrated its superiority by serving 2.74% more customers within the specified waiting time limit while using 7.57% fewer labor hours.

1. Introduction

Service organizations may be categorized based on the degree of control management has over demand for the service. Since one cannot inventory the customer contact activities of service delivery personnel, matching capacity to demand is challenging. Some services managers are fortunate in that they can control demand: appointment systems offer a means of doing this. More common are services that must respond to customer demand. For these organizations, a typical objective is to deliver a managerially- specified threshold level of service as inexpensively as possible.

Delivering a desired level of service at a low cost involves four interrelated tasks. Service delivery task 1 (SDT1) forecasts customer demand for the service. SDT2 translates these demand forecasts into employee requirements. SDT3 develops a labor schedule, using the employee requirements developed in SDT2 as one form of input (other inputs would include such things as employee preferences for shifts and governmental regulations). A common objective in developing this schedule is to provide the required number of employees in each period, while using as few employees as possible. SDT4 modifies the labor schedule in real-time, an action useful for various reasons: employees may be absent or late; customer demand may be higher or lower than anticipated. Thus, real-time control attempts, in reality (not only in a plan), to deliver the threshold level of service as cheaply as possible. Effective real-time control can, to a certain degree, mitigate the effect on customer service of poor decisions made in the earlier tasks, an issue to which we shall return later in the paper.

Our primary focus in this paper is upon SDT2, the translation of forecasts of customer demand into staffing requirements. Obviously, if customer arrival and service rates are constant across the operating day, then the same number of employees would always be desirable. Stochastic arrival or service processes, both common in services, make it desirable to vary, over the course of the operating day, the number of customer-service employees present. Given this, a manager must choose an appropriate duration for the planning periods. Although the labor scheduling literature commonly uses periods of 30 minutes or longer, this is likely a function of the difficulty in developing schedules with shorter duration periods. Short duration periods enable the up-front scheduling of 10 or 15 minutes rest-breaks, rather than taking them ad-hoc, and are preferable with higher variability in the customer arrival rate (Segal, 1974; Green and Kolesar, 1991). For these reasons, this investigation uses 15-minute planning intervals as did Agnihothri and Taylor (1991), Henderson and Berry (1976, 1977), Keith (1979), and Thompson (1990).

Having chosen an appropriate planning interval duration, a service manager must then translate each planning interval's forecast of customer demand into a staffing requirement. The manager will likely derive little guidance in this regard from the retailing, hospitality, or banking literature. For example, Frenk et al. (1991), who examined the relationship between labor volume and sales in retail trade, considered time-independent arrival and service rates (that is they ignored any within-year variability in arrival and service rates) and "... assumed that... the fine tuning between labour capacity and labour requirement is accomplished using part-time labour". One approach to translating customer demand into employee requirements is to simply collect relevant data on customer arrival rates (or transactions) and the service time per customer (or per transaction) and plug the numbers into commercial software (Brewton, 1989). The

obvious disadvantage of this method is the manager may not really know how the software determines the employee requirements.

Another means of translating customer demand into employee requirements is to use a productivity factor, such as 15 seated customers per server (Sill, 1991). Although this may be valid for service facilities operating at capacity and where service of customers overlap, it is likely to be of limited value when service of customers does not overlap, because of the stochastic nature of customer arrivals. Since there is no clear mathematical basis for determining the productivity-based translation factor (to provide a set level of service), the effectiveness of the procedure is susceptible to the quality of observations linking the actual level of customer service to the translation factor in use. Also, without a clear mathematical basis for determining the productivity factor, one must arbitrarily deal with fractional employee requirements. Finally, the translation factor should vary inversely in relation to the total workload, although this greatly increases the burden on those determining the factor.

A manager consulting the labor scheduling literature for guidance about how to determine employee requirements will find that most of this literature has either ignored SDT2, by assuming the employee requirements as prespecified, or has given the issue brief treatment. The manager also may find a statement to the effect that queuing models may be useful in developing the employee requirements: simply use an M/M/s queuing model (Agnihothri and Taylor, 1991; Andrews and Parsons, 1989; Buffa et al., 1976; Holloran and Byrn, 1986; Kolesar et al., 1975; Segal, 1974; Thompson, 1992), or an M/M/s/K queuing model (Quinn et al., 1991), and find the smallest number of staff, s, that provides the threshold level of service, defined as serving a set percentage of customers within a specified waiting time limit. This approach shall be called the traditional method of setting employee requirements, or MT1. Even the queuing

literature supports such an approach. For example, Kwan et al. (1988) state "in a service operation with short scheduling time periods and time-dependent arrival processes, the number of servers for each time period could be derived from the analytical solutions of a queuing model of the system with stationary and steady-state assumptions [i.e., MT1]" (p. 274).

In conducting some initial experiments using MT1 we observed customer service falling below the threshold level under longer customer service times. Specifically, we became aware that customers arriving in a particular period may affect staffing requirements in future periods. Our awareness of this phenomenon and MTl's failure to account for it motivated the development of an alternate method, MT2, that explicitly accounts for service duration and customer waiting time during the process of setting employee requirements. Our initial hypothesis was that over a range of conditions commonly observed in service delivery systems, MT2 should better satisfy the objective of delivering a threshold level of customer service as inexpensively as possible.

To test our hypothesis, we developed a full- factorial simulation experiment, varying ten factors, including the method of setting the employee requirements. In the following sections, we: present in detail two procedures for setting employee requirements; describe the service delivery system simulation experiments; provide the results of these experiments; discuss the significance of our findings; and offer a brief conclusion to the paper.

2. Determining the required number of employees

This section provides a description of two methods for setting desired staffing levels. Each method shares the same objective: find, for each planning period, the smallest number of staff that will deliver the threshold level of customer service, defined as "serving at least c

percent of customers with a wait for service not to exceed z periods". Service levels of this form may be seen in Agnihothri and Taylor (1991), Andrews and Parsons (1989), Brewton (1989), Buffa et al. (1976), Gaballa and Pearce (1979), Holloran and Byrn (1986), Kwan et al. (1988), Segal (1974), and Thompson (1992). Such an approach to determining employee requirements applies when "... the interests of the customer and the server are not mutual... [and so] it may be impossible to assign a monetary value to the cost of waiting" (Taha (1981), p. 47).

We should note at this point that there are other approaches to determining the required number of employees. Quinn et al. (1991) have argued that a total cost perspective to service delivery (minimizing the sum of customer dissatisfaction and service labor costs) offers a superior framework for determining employee requirements. Davis (1991), in contrast, has convincingly argued for service managers taking a longer-term perspective by considering the future competitive advantage offered by good customer service. As addressed in the discussion, our new method of translating forecasts of customer demand into employee requirements can be applied as easily under these service frameworks as it is with the one we study.

The following definitions shall apply throughout the paper:

T: the number of planning periods in the operating day;

z: the managerially specified limit on a customer's waiting time (in periods);

c: the managerially specified proportion of customers to be served within z periods (the threshold service level);

$\mu(\tau)$: the true service rate for customers arriving at time τ (in customers per period per employee);

g_t: the expected customer service rate in period t (in customers per period per employee);

$\lambda(\tau)$: the true customer arrival rate at time τ (in customers per period);

r_t: the expected customer arrival rate in period t (in customers per period);

P_x: the probability of x customers in the service system;

$P\{Y\}$: the probability of event Y;

$\lfloor x \rfloor$: the greatest integer that does not exceed x;

$|x|$: the absolute value of x;

s: the number of servers (employees who serve customers);

$\bar{w}$: an estimate of the average customer waiting time (in periods), given the threshold service level; and

q: a customer's waiting time in the queue (in periods).

Non-stationary customer service rates can originate with customers and with employees. The former arises from changes in the customers' service requirements over the operating day, while the latter arises from differences in the skills and work rates of the specific employees working at particular times during the day. Our investigation considers only the customer-driven cause of nonstationary customer service rates, not because the other is unimportant, but because one is faced with a perplexing dilemma when attempting to account for it. The dilemma is that calculating the effect of the employee-driven non-stationary service rate requires that one know who is working when. However, this information is not available until after the labor schedule is developed. Completing the loop, one must first complete SDT2, which requires known customer service rates, before developing the labor schedule. We return to this issue in the discussion.

2.1. A traditional method of setting employee requirements (MT1)

MT1 represents a traditional approach to setting the desired staffing levels and is the one a manager will see upon examination of the scheduling literature (Andrews and Parsons, 1989; Brewton, 1989; Buffa et al., 1976; Gaballa and Pearce, 1979; Holloran and Byrn, 1986; Kolesar et al., 1975; Kwan et al., 1988; Segal, 1974; Thompson, 1992). The employee requirement for a

period originates from an M/M/s queuing model, and equals the smallest number of employees who can provide the threshold level of customer service (serve the target percentage of customers within the specified waiting time). Assuming exponential service times and a Poisson arrival process, and using as inputs the expected arrival and service rates for period t, the desired staff size for the period equals the smallest value of x such that:

$$P\{q > z\} \leq 1 - c$$

where (Hillier and Lieberman, 1986, pp.421-423)

$$P\{q > z\} = (1 - P\{q = 0\})e^{-sg_t(1-h_t)z}$$

$$P\{q = 0\} = \sum_{n=0}^{s-1} P_n$$

$$P_n = P_0 \frac{(r_t/g_t)^n}{n!}$$

$$P_0 = \left[\left(\sum_{n=0}^{s-1} \frac{(r_t/g_t)^n}{n!}\right) + \frac{(r_t/g_t)^s}{s!\,(1-h_t)}\right]^{-1}$$

and

$$h_t = r_t/(sg_t)$$

Although the true arrival and service rates may be a function of time, MT1 assumes stationary arrival and service rates within each period. While one may assume

$$r_t = \int_{t-1}^{1} \lambda(\tau)d\tau \quad \text{and} \quad g_t = \int_{t-1}^{1} \mu(\tau)d\tau$$

in practice, it is unlikely that either $\lambda(\tau)$ or $\mu(\tau)$ can be precisely determined. As such, either a simple historical average of the number of customer arrivals in period t (the service rate for

customers arriving in period t), or a more sophisticated forecast of period t's anticipated customer arrival (service) rate, most likely serves to define $r_t(g_t)$.

2.2. A new method for setting employee requirements (MT2)

An awareness that customers arriving in a particular period may affect staffing requirements in future periods provided the impetus for MT2's development. Specifically, the proportion of service completed in the period in which a customer arrives declines as the customer's arrival time approaches the end of the arrival period. Figure 1(a) shows a situation where the mean service duration, 1 /g_t, is equal to 70% of the planning interval duration (that is, g_t = 1.43 customers per employee per period). Here, a customer arriving at time *t* — 1 (or at any time up to *t* — 0.7) will be served before the end of the period (ignoring waiting time and any variability in service time). As a customer's arrival time increases beyond *t* — 0.7 and approaches *t*, though, proportionally less of the service will be completed in period *t* and proportionally more of the service will occur in period *t* + 1. Indeed, a customer arriving an instant before the end of period *t* will place almost his entire service demand upon period *t* + 1.

Insert Figure 1 Here

As the service duration grows in relation to the planning interval duration, a customer affects staffing requirements in more periods. Figure 1(b) illustrates the case where g_t = 0.39 customers per employee per period. Here, a customer arriving in the middle of the period has 19.5%, 39.1%, 39.1%, and 2.3% of her required service completed in periods t through *t* + 3,

respectively (again ignoring waiting time). In general, the arrival of customers in period t will continue to place a demand for service in the following *k* periods, where

$k = \lfloor 1 + \bar{w} + 1/g_t \rfloor$

There undoubtedly are many ways to account for the multi-period impact of service when setting desired staffing levels. MT2 modifies the actual expected customer arrival rate to obtain an "effective" expected customer arrival rate for each period. The actual expected customer arrival rate for period *t* is reduced to account for customers who arrive in period *t* but have some of their service performed in subsequent periods, and is increased to account for customers who arrived in earlier periods but who have some service performed in period *t*.

Define $m(t, t + j)$ as the number of customers arriving in period *t* who are "effectively" served in periods $t + j, \ldots, t + k$. Also, define *n* as

$n = \lfloor 1 + \bar{w} \rfloor$

Note that no service occurs in periods $t, \ldots, t + n - 1$ for customers arriving in period *t* (i.e., service begins in period $t + n$ for customers arriving in period *t*). Given (7) and (8), then

$m(t, t + j) = r_t \quad \text{for} \quad j = 1, \ldots, n - 1$

$m(t, t + j) = r_t \left\{ v + g_t \left[-\frac{1}{2}(n^2 + 1 + \bar{w}^2 + v^2) + \bar{w}n - \bar{w} + \bar{w} - nv + n + v \right] \right\} \quad \text{for} \quad j = n$

$m(t, t + j) = r_t v \left[1 + g_t \left(1 + \bar{w} - j - \frac{1}{2} v \right) \right] \quad \text{for} \quad j = n + 1, \ldots, k,$

where

$$v = \begin{cases} 1, & if\ \bar{w} + 1/g_t \geq j \\ 1 - j + \bar{w} + 1/g_t, & otherwise \end{cases}$$

Appendix A supplies the derivations of eqns. (9) - (12).

We estimate the mean customer waiting time ($\bar{w}$), given the threshold level of service, as follows. First, calculate the simple average customer arrival ($\bar{r}$) and service ($\bar{g}$) rates across all periods in the day:

$$\bar{r} = \frac{1}{T}\sum_{t=1}^{T} r_t,$$

$$\bar{g} = \frac{1}{T}\sum_{t=1}^{T} g_t,$$

Next, using eqns. (l) - (6) and the average arrival and service rates, find the smallest number of staff, s, that delivers the threshold service level for an "average" period. Finally (Hillier and Lieberman, 1986, pp. 421-423),

$$\bar{w} = P_0 \frac{(\bar{r}/\bar{g})^s \bar{r}}{s\bar{g}} [s!\,(1 - \bar{r}/(s\bar{g}))^2 \bar{r}]^{-1}.$$

Given (7) - (15) and earlier definitions, an algorithm for determining the "effective" customer arrival rate in each period, e, (measured in customers per period), is:

(1) First set $\pi_t = 0$ for $t = 1, \ldots, T$ and then set $t = 0$

(2) Set $t = t + 1$, $j = k$ and $\pi_{t+k} = \pi_{t+k} + m(t, t+k)\, g_{t+k}/g_t$

(3) Set $j = j - 1$. If $j = 0$ go to step 5; otherwise, continue with step 4.

(4) Set $\pi_{t+j} = \pi_{t+j} + [m(t, t+j) - m(t, t+j+1)]\, g_{t+j}/g_t$. Return to step 3.

(5) Set $\beta_t = m(t, t+1)$. If $t < T$, return to step 2; otherwise, continue with step 6.

(6) Set $e_t = r_t - \beta_t + \pi_t$ for $t = 1, \ldots, T$ and stop.

Step 1 initializes the period index, t, to zero and initializes π_t, the variables recording the net number of customers moved into a period (from customers arriving in earlier periods), to zero for all periods. Steps two through five determine the number of customers moved out of and

into each period. The β_t measure the net number of customers moved out of each period (to place demand upon future periods). Step six calculates the "effective" expected customer arrival rates for each period. The "effective" expected rate for a period equals the net number of customers arriving in the period—the initial expected number of customer arrivals (r_t), less the number of customers moved to future periods (β_t), plus the number of customers transferred in from previous periods (π_t). Appendix B provides a detailed example of the algorithm.

An important observation to make at this point is that the number of customers moved out of a period can be different from the number moved into the receiving period. For example, compare, in step two, the number of customers transferred from period t to period $t + k$ (it is $m(t, t + k)$), to the number of customers transferred to period $t + k$ from period t (it is $m(t, t + k)\, g_{t+k}/g_t$,). These values differ by the factor g_{t+k}/g_t, which adjusts for the difference in service rates between the two periods. See Appendix B for a numerical example of the adjustment.

As does MT1, MT2 uses eqns. (1) - (6) to find the smallest number of staff necessary in a period to provide the threshold customer service level, but uses the effective expected arrival rate in each period, e_t, rather than the unadjusted expected arrival rate for the period, r_t, in eqns. (1), (5) and (6).

If stationary customer arrival and service rates exist, the multi-period impact of service will only be observed at the beginning and end of the operating day, due to the net effect of customer transfers in and out of periods. Since most services, even those with appointment systems, have nonstationary arrival and service rates, our primary hypothesis was that compared to MT1, MT2 will better satisfy the objective of delivering a specified threshold level of customer service as cheaply as possible.

We also hypothesized that MT2's advantage over MT1 in terms of inexpensively delivering the specified threshold service level would increase with (a) relatively longer service durations (achieved with longer service durations or shorter planning intervals), (b) lower threshold service levels, and greater variability in customer (c) arrival or (d) service rates across periods. Consider these expectations. First, longer service durations, shorter planning periods, and lower threshold service levels serve to increase the number of periods affected by customer arrivals in a period (by increasing the parameter k). Second, more variability in customer arrival or service rates results in greater imbalances between the workload arising from customers arriving in earlier periods but who continue to place a demand for service in the current period and the workload lost to future periods from customers who continue to place a demand for service upon future periods. A measure of the true within-day variation in arrival and service rates is:

$$\frac{(T-1)^{-1}\sum_{t=1}^{T-1}\left|\int_{t}^{t-1}\varphi\,(\tau)d\tau-\int_{t-1}^{t}\varphi(\tau)d\tau\right|}{T^{-1}\int_{0}^{T}\varphi(\tau)d\tau}$$

When $\varphi(t)=\lambda(t)(\varphi(t)=\mu(t))$, eqn. (16) measures the average change in arrival (service) rates between adjacent periods, divided by the mean customer arrival (service) rate.

3. The simulation experiment

Our goals in conducting the simulation experiment were to determine if MT2 is superior to MT1, and if so, to identify the conditions under which the superiority occurs. To meet these objectives, the experiment had a full-factorial design, as outlined in Table 1. In aggregate, the diversity and levels of the factors make the simulation experiment representative of many real service environments. Moreover, the diversity of the simulated environments helps to ensure that

the better method of setting employee requirements is widely effective, and not just best in an atypical environment.

In the following subsections we identify assumptions of the simulation experiment, the experimental factors associated with customer arrivals, customer service, the accuracy of customer arrival and service rate forecasts, and labor scheduling flexibility. Then we introduce the performance measures used in the simulation and, finally, present details on the simulation process.

Insert Table 1 Here

3.1. Assumptions of the simulation experiment

The assumptions of the simulation experiment include: an 18-hour operating day; 15-minute planning intervals; a mean arrival rate of one customer per minute; exponential service times (Andrews and Parsons, 1989; Harris et al., 1987; Holloran and Byrn, 1986; Kwan et ah, 1988; Paul and Stevens, 1971; Quinn et ah, 1991; Segal, 1974); Poisson customer arrivals (Andrews and Parsons, 1989; Gaballa and Pearce, 1979; Harris et al., 1987; Holloran and Byrn, 1986; Kwan et al,, 1988; Paul and Stevens, 1971; Quinn et al., 1991; Segal, 1974); stable underlying customer arrival-rate and service-rate functions; the employees work as scheduled (no employees are tardy or absent); once service starts for a customer, it is never interrupted.

Insert Figure 2 Here

Four assumptions need further explanation. First (and second), the assumptions of stable underlying customer arrival-rate and service-rate functions refer to the underlying arrival and service *processes*. That is, service times follow an exponential distribution and arrivals follow a Poisson distribution, but the underlying processes are stable over the simulation period. For a customer scheduled to arrive at time τ (where $t - 1 \leq \tau \leq t$), the time until the arrival of the next customer arrival is given by time until next arrival $= \lambda(\tau)^{-1}\ln(rnd_1)$, while the service time for the customer is given by service duration for a customer arriving at time

$\tau = \mu(\tau)^{-1}\ln(rnd_2)$,

where rnd_i is a number selected randomly from the uniform [0,1) distribution. The implication of eqns. (17) and (18) is that if a simulated environment starts with an underlying arrival pattern with a single, mid-day peak, for example, it does not switch in mid-simulation to one with three peaks (the function $\lambda(\tau)$ remains unchanged over the simulated environment).

Third, by assuming the employees work as scheduled, we avoid any confounding effects introduced by real-time control activities (SDT4). More important, however, is the absence of any service- based literature that presents, investigates, or offers guidance regarding effective real-time labor schedule control mechanisms.

Fourth, to mimic the operation of organizations that deliver good customer service, we assume that once started, a customer's service proceeds uninterrupted. We do, however, allow service to be transferred between employees at or after the end of a shift. This assumption sometimes has the effect of extending shifts, an issue we consider again in the section on performance measures.

Insert Table 2 Here

3.2. Customer arrivals

To make our study as broadly based as possible, we varied two experimental factors related to customer arrivals. First, we selected four customer arrival-rate patterns representing those existing in service organizations: unimodal (seen, for example, in organizations with one mid-day peak, such as a grocery store on a Saturday), bimodal (occurring in services with "commute" peaks, like highway toll facilities), trimodal (a typical pattern for many restaurants), and random (with many peaks and valleys in daily demand).

Second, all arrival-rate patterns were generated with two levels of variation: coefficients of variation of 0.20 and 0.45. Combining the levels of the two customer arrival-related factors yielded a total of eight different underlying customer arrival-rate curves, as illustrated in Fig. 2. Equation (16), a measure of the within-day variation in the customer arrival rate, varies across levels of the factors representing (a) the arrival pattern and (b) the variability of the arrival pattern, as shown in Table 2. In order of declining variation, the arrival-rate patterns are: random, trimodal, bimodal, and unimodal.

3.3. Customer service

The experiment contained four factors related to customer service. One factor represented the true service-rate pattern within a day. Sinusoidal curves, selected for convenience, provided the service-rate pattern, with the four levels of this factor having a periodicity of 60, 120, 180, and 540 minutes. A second factor, the variability of underlying customer service rate, had two levels: coefficients of variation in the true service-rate pattern of 0.15 and 0.30. Table 2 shows the values of eqn. (16), a measure of the within-day variation in the customer service rate, across

levels of the factors representing (a) the service-rate pattern and (b) the variability of the service-rate pattern. Noteworthy is that longer periodicity in the servicerate pattern equated with lower values of eqn. (16).

The third factor, the mean service rate, had four levels: mean service rates (durations) of 6.00 (2.5), 3.00 (5.0), 1.50 (10.0), and 0.75 (20) customers per employee per period (minutes). There were three threshold service levels, the fourth service- related factor. These levels required that 75% of customers be served within waiting times of 0.0333, 0.1333, and 0.5333 periods (0.5, 2.0, and 8.0 minutes, respectively). Although it may seem more desirable to have a threshold service level that requires serving a higher percentage of customers within some specified waiting time limit, choosing a high value for the target percentage makes it more difficult to observe differences in customer service provided by competing SDT2 methods. Moreover, serving a high percentage of customers within a specified waiting time is equivalent to serving a lower percentage of customers within a shorter waiting time.

The range of service durations, combined with the customer arrival-rate curves and threshold service levels, yielded service environments having a wide range of mean employee requirements. With the fastest (slowest) customer service rate and a mean arrival rate of one customer (two customers) per minute and the eight (one-half) minute waiting time limit, MT1 specified a mean requirement of approximately three (46) employees, per period.

3.4. Accuracy of customer arrival and service rate forecasts

Since, in practice, managers must translate *forecasts* of customer demand into desired staffing levels, forecast accuracy is a concern. In this paper, the amount of "historical" information available to the procedure for setting the employee requirement influences the

degree of forecast accuracy, another environmental experimental factor. Using eqns. (17) and (18) with separate random number streams enabled us to simulate arrivals and service durations for customers that arrived to the facility in the "past". The three levels of forecast accuracy used observations from 18, 6, and 2 simulated "historical" days to develop simple averages of the customer arrival and service rates for each planning interval.

Clearly, using more information in developing each period's average arrival and service rate improves the accuracy of these estimates. For example, given the variability inherent in the exponential distribution, one expects that the observed average arrival rate (and average service duration) for a period may vary substantially from the period's true mean arrival rate (true mean service duration) when the observed averages are based only on a few simulated days.

3.5. Scheduling flexibility

Labor scheduling, SDT3, is a key component of the simulation experiment. Since the employee requirements developed in SDT2 are an input to the scheduling function, it is important to examine the downstream effect of different required staffing levels. For example, although the staffing levels specified by competing SDT2 methods may be significantly different, there may be very little difference in labor schedules developed using the specified staffing levels.

The simulation experiment includes an environmental factor for the degree of scheduling flexibility available when developing the labor schedule. This factor has two levels, representing extremes of labor scheduling flexibility. The low level of scheduling flexibility allows shifts of eight working hours, with an hour-long break taken between the fourth and fifth hours of work, for a total of only 37 different shifts. With low scheduling flexibility, the labor schedule is solved

optimally using the branch-and-bound integer programming procedure of SAS-OR (SAS Institute, 1988).

At the high level of scheduling flexibility, the simulation uses the employee requirements generated by MT1 and MT2 directly. Doing this has the same effect as using many different shifts when developing the labor schedule, since with arbitrarily high scheduling flexibility it is possible to satisfy each period's employee requirement exactly. No labor schedules are developed at this level of scheduling flexibility, however, because it is difficult to determine, a priori, the degree of flexibility necessary to satisfy the employee requirements exactly, and second, because of the difficulty in optimally solving labor scheduling problems having extensive flexibility.

3.6. Performance measures

This paper uses two measures to evaluate service delivery system performance, as shown in Table 1. One measure is the average quarter-hours of paid labor. As addressed in the section on simulation assumptions, employees' shifts occasionally extend beyond the planned duration to ensure that a customer's service proceeds uninterrupted once started. Thus, the actual labor usage sometimes exceeds the amount of scheduled labor, and so is a more accurate measure of the true labor cost one might expect when implementing a particular schedule.

The second performance measure is the actual percentage of customers whose wait for service does not exceed the specified limit. As the goal of the simulation was to serve 75% of customers with a wait not exceeding the duration specified by the threshold service level factor, actual service levels above (below) 75% would represent acceptable (unacceptable) customer service.

3.7. Simulation details

The simulation was written in FORTRAN and took approximately 110 hours to run on a personal computer with a 80486DX2-66 processor. Conducting the simulation experiment required that we:

(1) Select a customer arrival curve (one of the eight illustrated in Fig. 2.

(2) Generate "historical" customer arrival information.

(3) Generate 30 days of "future" information: store in a data file an arrival time and a random number to be used in determining service duration for each customer arriving over this 30-day period.

(4) Select, in order, some combination of customer service rate, service-rate pattern, service-rate pattern variability, threshold service level, forecast accuracy and method of setting the employee requirements. Develop each period's employee requirement using the appropriate "historical" mean customer arrival and service rates for the period.

(5) Select a value of scheduling flexibility. If the scheduling flexibility is low, develop an optimal labor schedule using the employee requirements determined in step four as an input. Otherwise, we assume the employee requirements and the labor schedule's staffing levels are identical.

(6) Simulate the operation of the service facility for 30 days (using the stored data on customer arrival times and service duration determining random numbers) with the staffing levels from the labor schedule identified in step five, and measure system performance.

(7) Repeat steps five and six for both levels of scheduling flexibility.

(8) Repeat steps four through seven for all combinations of customer service rate, service-rate pattern, service-rate pattern variability, threshold service level, forecast accuracy and method of setting the employee requirements (i.e., 4 x 4 x 2 x 3 x 3 x 2 = 576 times).

(9) Repeat steps one through eight, replicating each customer arrival curve three times (selecting a different random number seed each time), thus executing a total of 27648 (2 x 576 x 8 x 3) simulations.

One should not construe from our simulation of 30 days with the same schedule in effect that we advocate keeping the same schedule for more than a day. The 30-day period served only to give an accurate measure of the true performance that might be expected when implementing a particular schedule.

4. Results

Table 3 provides a summary of results. Overall, MT1 and MT2 used 1375.86 and 1271.72 labor quarter-hours (343.96 and 317.93 labor hours) and served 85.83 and 88.19 percent of customers within the specified waiting time limit, respectively.

To evaluate the statistical significance of the results we developed two ANOVA models with our performance measures as dependent variables. In each ANOVA model, the main effect of the method of setting the employee requirements and all first-order interaction effects incorporating this factor were significant at the α = 0.0001 or the α= 0.0000 level. The only exception was the interaction of the method of setting the employee requirements with the threshold service level factor in the labor-usage ANOVA model, which was not significant at the a = 0.10 level.

Figure 3 graphically illustrates the first-order MT-based interactions. Each horizontal axis in Fig. 3 is the service level provided by MT2 less the service level provided by MT1, while each vertical axis is the actual labor savings offered by MT2, measured as a percentage of MTl's labor actual cost.

Insert Table 3 Here

Insert Figure 3 Here

---

5. Discussion

The results offer excellent support for our general hypothesis of MT2's superiority. Compared to MT1 over all 13824 simulated environments, MT2 used 7.57% fewer labor hours while serving 2.74% more customers within the specified waiting time limit, differences significant at the α= 0.0000 level. In the single case where MT1 served more customers within the specified waiting time limit than MT2 (low scheduling flexibility), (a) MT2 still provided much better service than required by the threshold service level and (b) the labor savings offered by MT2 was, as a percentage, very much greater than MTl's higher service level. Thus, across all levels of all experimental factors, MT2 better satisfied the objective of providing the threshold service level as inexpensively as possible. Given the difference in performance over the extensive range of simulated conditions, we recommend that service managers use MT2 rather than MT1. Compared to MT1, MT2 provided better customer service and used less labor to do it,

simply because MT2 offers a superior means of ensuring that labor is present *at the times needed*.

The overall results raise a subtle point; specifically, that one cannot judge a scheduling method only by the level of customer service that it provides. To illustrate the importance of this point, consider a statement by Holloran and Byrn (1986). In reporting on a comprehensive system for managing a telephone reservation system, these authors state "Monitoring the telephone service factor has shown that service factors are within one percent of the desired level" (p. 41). As noted earlier, these authors used MT1. Had they instead used MT2, they would very likely have maintained or increased their customer service level while using less labor.

While our general hypothesis was well supported, the results incompletely supported our specific hypotheses regarding MT2's relative superiority. As expected, MT2's performance advantage over MT1 (relatively better service and relatively lower labor costs) increased with (a) longer service durations (Fig. 3(c)), (b) lower threshold service levels (Fig. 3(f)), and (c) greater variation in the service-rate pattern (Fig. 3(e)). There was only partial support for MT2's performance advantage increasing under (a) arrival-rate patterns with more across-period variability, (b) a higher coefficient of variation in the arrival-rate pattern, and (c) service-rate patterns having shorter periodicity. This incomplete support may be a result of the interaction between the setting of employee requirements (SDT2) and the process of labor scheduling (SDT3), but should be put in perspective relative to the well-supported primary hypothesis of MT2's broad superiority.

The relative effectiveness of MT1 and MT2 was markedly different across the levels of scheduling flexibility, as Fig. 3(g) shows. When scheduling flexibility was high, MT2 served 7.04% more customers within the specified waiting time limit and used 1.02% fewer labor hours

than did MT1. In contrast, MT2 used 10.71% fewer labor hours than MT1 but served 0.51% fewer customers within the desired time at the low level of scheduling flexibility. That MT2 served fewer customers in the desired time when scheduling flexibility was low should not be cause for alarm since, on average, MT2 served 22.23% *more* customers within the waiting time limit than required by the threshold service level. Much more problematic is MTl's failure to deliver the threshold service level under the high level of scheduling flexibility (MT1 served 1.06% *fewer* customers than required by the threshold service level). Although most services fall between the extremes of scheduling flexibility examined in this paper, the results offer useful information to managers attempting to increase the flexibility at their disposal. As expected, labor costs fall and the actual level of service moves much closer to the threshold level as scheduling flexibility increases. Although the labor savings offered by MT2 is greater with lower scheduling flexibility, MT2 proves very beneficial in lowering labor costs and delivering the threshold level of customer service even with extreme scheduling flexibility.

In the introduction, we stated that real-time control activities may serve to mitigate the effect of poor decisions made in the three temporally earlier service delivery tasks. Does this statement, assuming it is legitimate, invalidate our results? No— we hypothesize that real-time control will be more easily, and hence more cheaply, accomplished with MT2's more appropriate numbers of employees on-hand. We assert this because although MTl's schedules have more labor available than do MT2's, MTl's surplus labor is not present at the times needed (otherwise, the level of service provided by MT1 would significantly exceed that provided by MT2, and this did not happen). Real-time control could serve to reduce the surplus, unproductive labor that MT1 schedules, bringing MTl's true labor cost closer to that of MT2. However, for

this to happen, the real-time control necessary for MT1 will be substantially harder to implement than the real-time control needed for MT2.

The concept of using an effective arrival rate, which takes into account the multi-period impact of service, should have wide applicability. In short, using effective arrival rates proved superior because doing so better ensures the right numbers of staff are present when needed. Both the concept of effective arrival rates, and the algorithm presented earlier for calculating the effective arrival rates, operate independently of the service framework for determining the employee requirements. Because of this, the concept and the algorithm can likely be applied with equal effectiveness in environments having other service frameworks, such as those espoused by Davis (1991) and Quinn et al. (1991), or those having other service and arrival rate distributions or queue structures. It is also important to remember that one can use MT2 (and effective arrival rates) even for services having constant arrival and service rates. If such a service operates around the clock, the number of customers transferred out of and in to each period are identical, and so MT1 and MT2 specify the same number of employees in all periods. If such a service does not operate around the clock, MT1 and MT2 still specify the same employee requirements for periods $k + , \ldots, T$, but they differ in the employee requirements for periods, $1, \ldots, k$. Here, MT1 specifies the same employee requirements for periods $1, \ldots, k$ as it did for all other periods, while MT2 lowers the actual arrival rate for periods $1, \ldots, k$ since these periods have more customers transferred out than in.

Hall (1991), in considering non-stationary customer arrival rates, provides some interesting suggestions regarding when to add service capacity to minimize queue size. Unfortunately for service managers, the best times to change capacity are usually mid-period. The implication of this is that one should use shorter-duration periods than labor scheduling

typically uses. A trend toward shorter-duration periods does exist in the labor scheduling literature, driven in part by the increases in computing ability enabling the development of schedules based on shorter planning intervals. As planning intervals become shorter, any service duration grows in relation to the duration of the planning interval, thus further increasing the relative advantage of MT2 (as Fig. 3(d) shows) and the use of effective arrival rates.

Various extensions are possible to the research presented in this paper. Some of the possible extensions are (1) investigating the value of effective arrival rates in environments using other frameworks for determining employee requirements (such as those suggested by Davis (1991) and Quinn et al. (1991)) and modifying the calculation of effective arrival rates to account for (2) nonstationary customer service rates originating with employees, (3) within-period non-stationary arrival and service rates, and (4) inherent variability in service duration and waiting time across customers.

6. Conclusions

We have presented a new method, MT2, for determining the employee requirements used as inputs to the labor scheduling process. MT2, by accounting for the multi-period impact of service, offered significant improvements over the traditional approach to the task. Over 13824 simulated service environments, MT2 used 7.57% fewer labor hours while serving 2.74% more customers in the desired time (differences significant at the $\alpha = 0.0000$ level). Given MT2's simplicity of use and effective performance, it is likely to have wide applicability in service organizations.

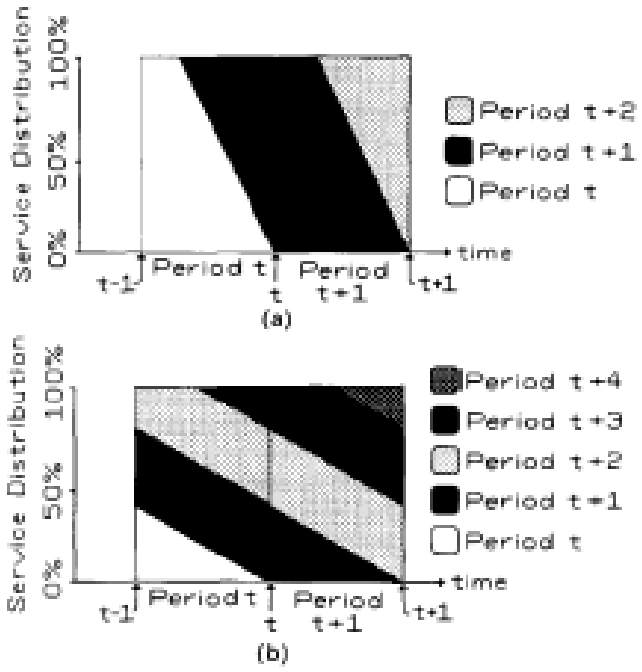

Figure. 1. Examples of the multi-period impact on service of customer arrivals: The proportion of service performed in the arrival and subsequent periods as a function of a customer's arrival time, when the service duration is (a) 70% and (b) 256% of the planning interval duration and where service commences upon the customer's arrival.

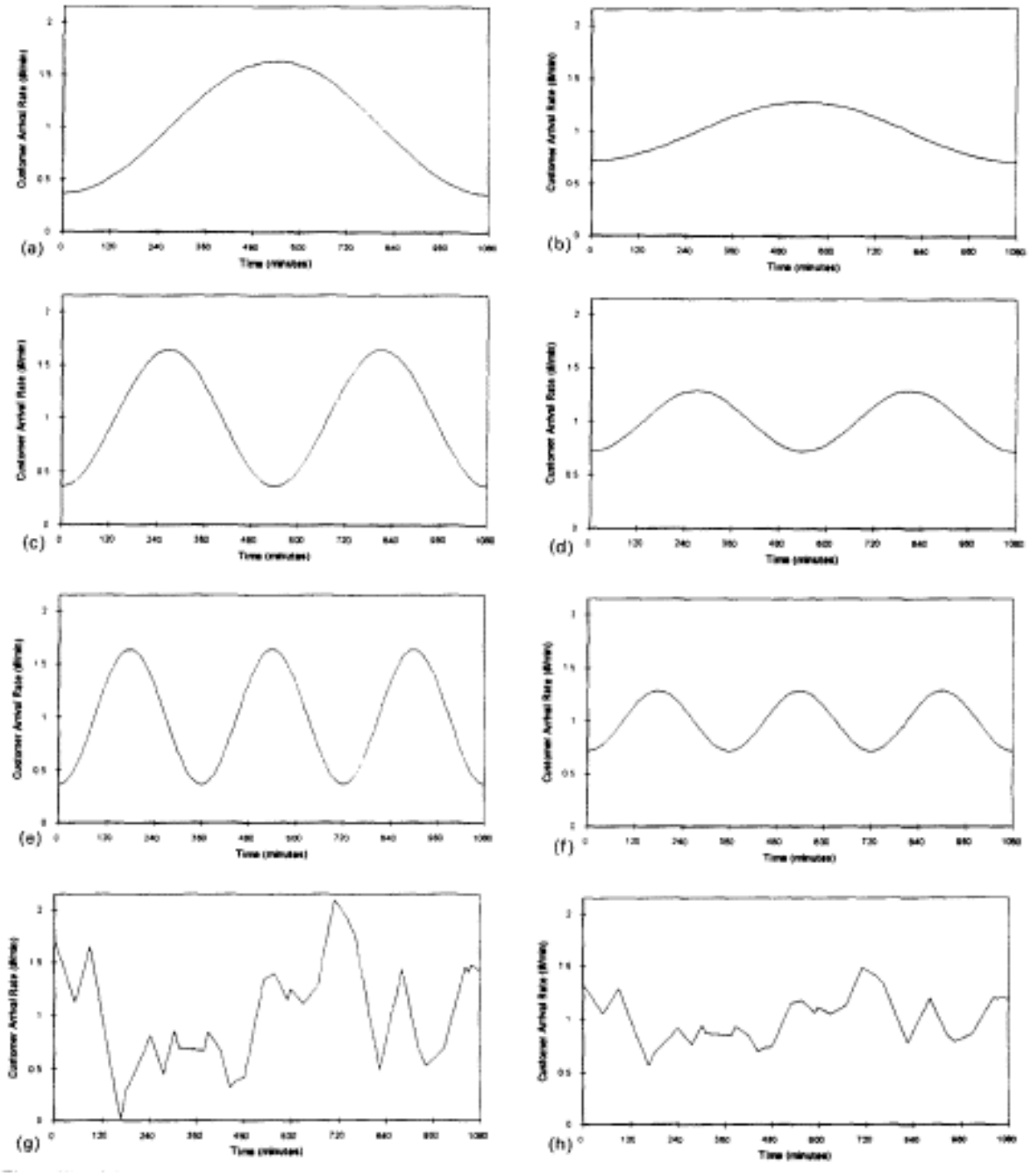

Figure. 2. The eight true customer arrival-rate curves: (a) unimodal, high-variation pattern; (b) unimodal, low-variation pattern; (c) bimodal, high-variation pattern; (d) bimodal, low-variation pattern; (e) trimodal, high-variation pattern; (0 trimodal, low- variation pattern; (g) random, high-variation pattern; and (h) random, low-variation pattern.

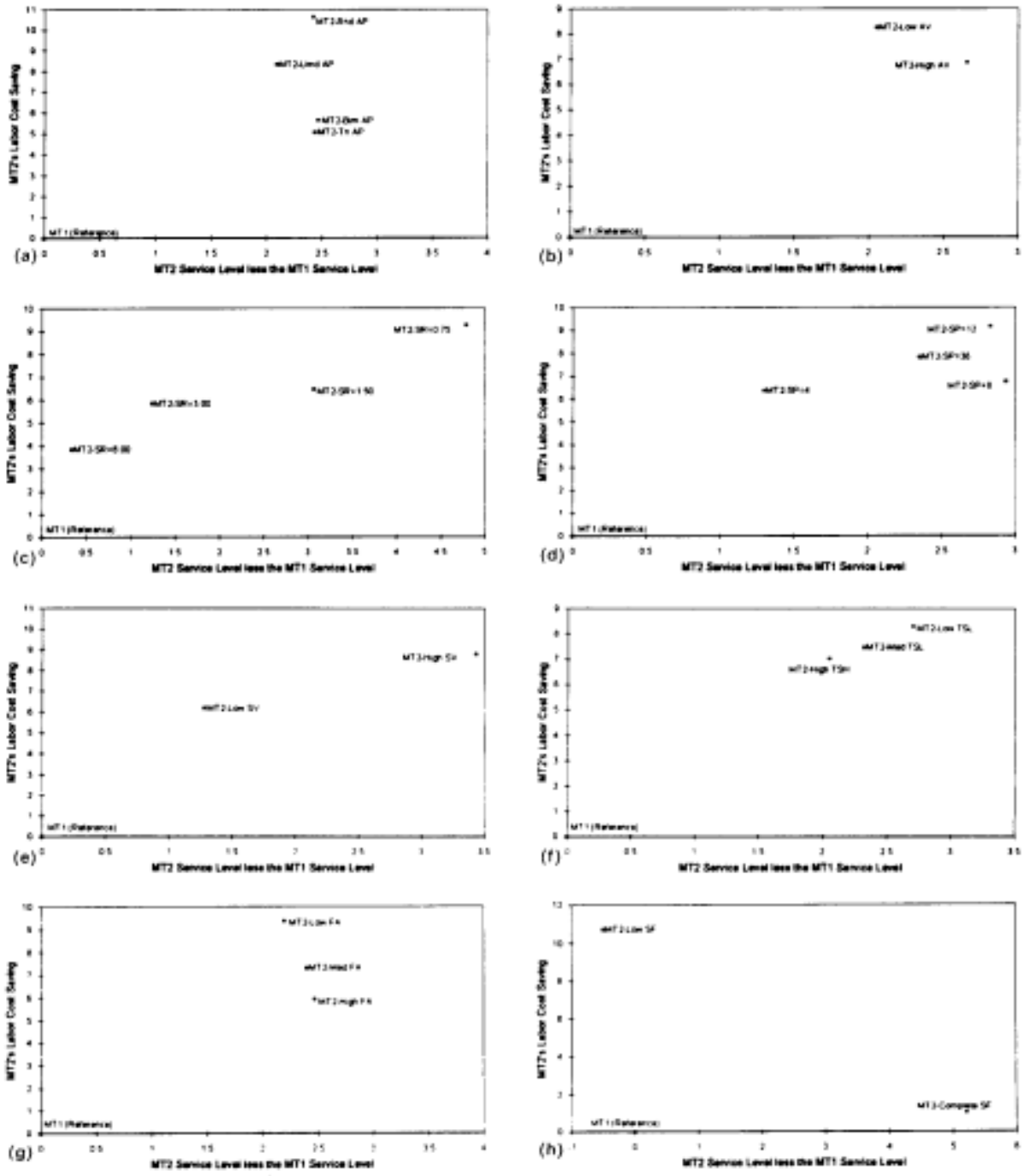

Figure. 3. Interactions between the method of setting the employee requirements (MT) and (a) arrival-ratebpattern (AP); (b) arrival- rate pattern variation (AV); (c) service rate (SR); (d) service-rate pattern (SP); (e) variation in service rate (SV); (f) threshold service level (TSL); (g) forecast accuracy (FA); and (h) scheduling flexibility (SF).

Table 1. Experimental design overview.

Factor (Number of levels)	Measured as	Levels
Underlying (true) customer arrival-rate pattern (4)	A pattern	Unimodal, bimodal, trimodal, and random (see Fig. 2)
Coefficient of variation of the true customer arrival-rate pattern (2)	A coefficient of variation (standard deviation/mean)	0.20 (low) and 0.45 (high)
Mean service rate (4)	A daily average of the number of customers served per period per employee	6.00, 3.00, 1.50, and 0.75 customers per period per employee (2.5, 5, 10 and 20 minutes per customer)
Underlying (true) service-rate pattern (4)	The periodicity of the true service-rate pattern, in periods	4, 8, 12, and 36 periods (60, 120, 180, and 540 minutes)
Variation in the service-rate pattern (2)	A coefficient of variation	0.15 (low variation), and 0.30 (high variation)
Threshold customer service level (3)	The maximum waiting time, in periods, within which to serve at least 75% of customers	0.0333, 0.1333, and 0.5333 periods (0.5, 2.0, and 8.0 minutes)
Accuracy of the customer arrival- and service-rate forecasts (3)	Number of "historical" periods' observations used in determining the average rates	18 (high accuracy), 6 (medium accuracy), and 2 (low accuracy)
Scheduling flexibility (2)	The degree of flexibility available in SDT3	Low flexibility and high flexibility (see text for details)
Method of setting the employee requirements (2)	Not applicable	MT1, a traditional method, and MT2, a new method
Dependent variable	**Measured as**	
Actual labor cost	Average daily paid quarter-hours of work (see text for details)	
Actual customer service level	Average percentage of customers served within the specified waiting time	

Table 2. Across-period variability in the arrival and service rates*

Arrival-rate pattern	Arrival-rate pattern variation		Service-rate pattern	Service-rate pattern variability	
	High	Low		High	Low
Random	0.1437	0.0637	60 min	0.2737	0.1369
Trimodal	0.1062	0.0472	120 min	0.1935	0.0968
Bimodal	0.0713	0.0317	180 min	0.1368	0.0684
Unimodal	0.0358	0.0159	540 min	0.0475	0.0238

* As measured using eqn. (16).

Table 3. Summary results.

Factor	Level	Actual performance measures			
		Service level		Paid labor quarter-hours	
		MT1	MT2	MT1	MT2
Arrival-rate pattern	Unimodal	85.70	87.79	1248.80	1144.73
	Bimodal	86.02	88.48	1368.94	1291.88
	Trimodal	85.24	87.67	1282.42	1217.60
	Random	86.38	88.81	1603.27	1432.67
Arrival-rate pattern variation	Low	85.88	87.93	1396.16	1281.05
	High	85.79	88.45	1355.55	1262.39
Service rate	6.00 cpppe[a]	91.71	92.04	445.12	428.04
	3.00 cpppe	88.22	89.45	796.09	749.67
	1.50 cpppe	83.90	86.96	1473.09	1377.83
	0.75 cpppe	79.52	84.30	2789.12	2531.34
Service-rate pattern	4 periods	87.22	88.52	1303.26	1649.12
	8 periods	84.83	87.77	1336.63	1246.37
	12 periods	85.23	88.06	1444.60	1312.31
	36 periods	86.06	88.41	1418.93	1307.92
Service-rate pattern variation	Low	87.56	88.84	1285.80	1206.30
	High	84.11	87.53	1465.92	1337.15
Threshold service level	$z = 0.0333$ periods	84.06	86.11	1450.74	1348.95
	$z = 0.1333$ periods	84.87	87.18	1388.11	1284.54
	$z = 0.5333$ periods	88.57	91.27	1288.72	1181.68
Forecast Accuracy	Low	84.19	86.39	1425.34	1291.89
	Medium	86.29	88.69	1365.72	1265.99
	High	87.02	89.49	1336.51	1257.28
Scheduling flexibility	Low	73.94	79.15	892.50	883.43
	High	97.73	97.23	1859.21	1660.01
Overall results		85.83	1375.86	88.19	1271.72

[a] cpppe = customers per period per employee.

Appendix A

This appendix derives the formulas for the number of customers transferred out of each planning period (to place demands for service upon future periods). To simplify the derivation that follows, we ignore any across-customer variation in service duration.

To begin, consider a customer arriving at time τ in period t (i.e., $t-1 \leq \tau \leq t$). As does MT1, we ignore any inter-period variability in arrival or service rates, so service for this customer will take $1/g_t$ periods and service will finish at time $\tau + \bar{w} + 1/g_t$ (i.e., at the arrival time, plus the estimated waiting time, plus the service duration). Consider the proportion of this customer's service performed in periods $t + j, t + j + 1, \ldots, t + k$ where eqn. (7) defines k. One finds this proportion by first taking the difference between the time of service completion and time at which period $t + j$ starts, and then taking the ratio of the resultant time to the total service duration, or:

$$\frac{(\tau + \bar{w} + 1/g_t) - (t + j - 1)}{1/g_t}$$

Equation (A.l) only applies if the numerator is positive, or

$$\tau \geq t + j - 1 - \bar{w} - 1/g_t$$

Note that the right side of eqn. (A.2) is the earliest time at which a customer can arrive and still place some demand for service upon period $t + j$. To find the number of customers who arrive in period t but effectively place their demand for service upon periods $t + j, t + j + 1, \ldots, t + k$, we must integrate

$$\int_{t_1}^{\text{end of period } t} (\text{customer arrival rate}) \times (\text{proportion of service performed in periods } t+j, t+j+1, \ldots, t+k)\, d\tau, \qquad \text{(A.3)}$$

where the integration starts at time t_1 = max {(start of period t), (earliest time at which a customer can arrive and still place a demand for service upon period $t + j$)}.

Substituting appropriate notation, eqn. (A.3) becomes:

$$m(t, t+j) = \int_{t-1}^{t} r_t \, d\tau \quad \text{for } j = 1, \ldots, n-1, \qquad \text{(A.4)}$$

$$m(t, t+j) = \int_{t-v}^{t+n-1-w} r_t \times \left[\frac{(\tau + \bar{w} + 1/g_t) - (t+j-1)}{1/g_t}\right] d\tau + \int_{t+n-1-w}^{t} r_t \, d\tau \quad \text{for } j = n, \qquad \text{(A.5)}$$

$$m(t, t+j) = \int_{t-v}^{t} r_t \left[\frac{(\tau + \bar{w} + 1/g_t) - (t+j-1)}{1/g_t}\right] d\tau \quad \text{for } j = n+1, \ldots, k, \qquad \text{(A.6)}$$

where

$$v = \begin{cases} 1 & \text{if } \bar{w} + 1/g_t \geqslant j, \\ 1 - j + \bar{w} + 1/g_t & \text{otherwise.} \end{cases} \qquad \text{(A.7)}$$

Integrating (A.4) through (A.6) and simplifying yields eqns. (9) - (11), respectively.

Appendix B

This appendix presents a simple numerical example of the algorithm used in MT2. To illustrate the use of this algorithm, consider a four- period example with arrival and service rates as given in Table B.1, where each period is 15 minutes long and management's threshold service level requires serving 75% of customers within 7.5 minutes (i.e., c = 0.75 and z = 0.5). The average arrival rate, $\bar{r}$, is 65 customers per period, while the average service rate, $\bar{g}$, is 0.8 customers per employee per period. Applying eqns. (1) (6) using the average arrival and service rates, one finds that 84 employees will deliver the threshold service level, and that the expected customer waiting time, $\bar{w}$, is 0.308 periods (found using eqn. (15)).

Table B.1
Example parameters

Period	g_t [a]	r_t [b]
1	0.72	50
2	0.88	100
3	0.96	80
4	0.64	30

[a] The service rates are measured in customers per period per employee.
[b] The arrival rates are measured in customers per period.

Table B.2 shows the sequential steps in the algorithm and the values determined at each step. Of the 50 customers expected to arrive in period one, 8.73 are "moved" to period three, and 32.64 (= 41.37 — 8.73) are "moved" to period 2, leaving an effective arrival rate of 8.63 customers in period one. In "moving" customers between periods, it is important to consider the differences in service rates of customers arriving in the two periods. Thus the 32.64 customers

"moved" from period one to period two are equivalent to 39.89 (= 32.64 x 0.88/0.72) of the period-two customers.

Recall that MT1 and MT2 both use eqns. (l) - (6) to find the smallest number of staff necessary in a period to serve at least the threshold percentage of customers within the specified waiting time. Also recall that while MT1 uses the unadjusted expected arrival rate for each period (the r_t) in eqns. (1), (5) and (6), MT2 uses the effective expected arrival rate in each period (the e_t). A comparison of the actual and effective arrival rates shows the relevance of this point: the actual arrival rate exceeds the effective arrival rate for periods one and two but the reverse is true in periods three and four. Applying eqns. (1) - (6) with the specified threshold service level gives MT2's requirements of 14, 72, 114, and 80 employees for periods one through four, respectively.

Finally, note that the algorithm attempts to calculate π_5 and π_6. The importance of these parameters is that they show the number of staff needed in the periods after the facility closes, and no longer accepting new arrivals, to process customers currently in the system. In our example, there were no customers "moved" into period one, and we ignored the customers "moved" into periods 5 and 6. In contrast, consider an organization operating around the clock. Here, MT2 accounts for the customers who arrive just before midnight when determining the number of employees needed just after midnight (early morning the next day). Thus, all periods are recipients of "moved" employees and the algorithm eventually accounts for all "moved" employees (but not necessarily in the current planning cycle).

Table B.2
Steps in the algorithm

Step	t	k	n	j	v	$m(t,t+j)$	π_t	β_t	e_t
1	0						$\pi_1 = 0, \pi_2 = 0, \pi_3 = 0, \pi_4 = 0$		
2	1	2	1	2	0.697	$m(1,1+2) = 8.73$	$\pi_3 = 0 + 8.73 \times 0.96/0.72 = 11.65$		
3				1					
4					1	$m(1,1+1) = 41.37$	$\pi_2 = 0 + (41.37 - 8.73) \times 0.88/0.72 = 39.89$		
3				0					
5								$\beta_1 = 41.37$	
2	2	2	1	2	0.444	$m(2,2+2) = 8.68$	$\pi_4 = 0 + 8.68 \times 0.64/0.88 = 6.31$		
3				1					
4					1	$m(2,2+1) = 78.91$	$\pi_3 = 11.65 + (78.91 - 6.31) \times 0.96/0.88 = 88.27$		
3				0					
5								$\beta_2 = 78.91$	
2	3	2	1	2	0.349	$m(3,3+2) = 4.69$	$\pi_5 = \text{NA}$		
3				1					
4					1	$m(3,3+1) = 61.60$	$\pi_4 = 6.31 + (61.60 - 4.69) \times 0.64/0.96 = 44.25$		
3				0					
5								$\beta_3 = 61.60$	
2	4	2	1	2	0.870	$m(4,4+2) = 7.27$	$\pi_6 = \text{NA}$		
3				1					
4					1	$m(4,4+1) = 25.40$	$\pi_5 = \text{NA}$		
3				0					
5	0							$\beta_4 = 25.40$	
6	1								$e_1 = 50 - 41.37 + 0 = 8.63$
6	2								$e_2 = 100 - 78.91 + 39.89 = 60.98$
6	3								$e_3 = 80 - 61.60 + 88.27 = 106.67$
6	4								$e_4 = 30 - 25.40 + 44.25 = 48.85$

References

Agnihothri, S.A. and Taylor, P.F., 1991. "Staffing a centralized appointment scheduling department in Lourdes Hospital". *Interfaces*, vol. 21, no. 5, 1-11.

Andrews, B.H. and Parsons, H.L., 1989. "L.L. Bean chooses a telephone agent scheduling system". *Interfaces*, vol. 19, no. 6, 1-9.

Brewton, J.P., 1989. "Teller staffing models: Instruments to achieve superior customer service". *Financial Managers' Statement,* vol. 11, no. 4, 22-24.

Buffa, E.S., Cosgrove, M.J. and Luce, B.J., 1976. "An integrated work shift scheduling system". *Decision Sciences*, vol. 7, no. 4, 620-630.

Davis, M., 1991. "How long should a customer wait for service?". *Decision Sciences*, vol. 22, no. 2, 421-434.

Frenk, J.B.G., Thurik, A.R. and Boot, B.A., 1991. "Labor costs and queuing theory in retailing". *European Journal of Operational Research*, vol. 55, no. 2, 260-267.

Gaballa, A. and Pearce, W., 1979. "Telephone sales manpower planning at Qantas". *Interfaces*, vol. 9, no. 3, 1-9.

Green, L. and Kolesar, P., 1991. "The pointwise stationary approximation for queues with nonstationary arrivals". *Management Science*, vol. 37, no. 1, 84-97.

Hall, R.W., 1991. *Queuing Methods for Services and Manufacturing*. Prentice-Hall, Englewood Cliffs, NJ.

Harris, C M., Hoffman, K.L. and Saunders, P.B., 1987. "Modeling the IRS taxpayer information system". *Operations Research*, vol. 35, no. 4, 504-523.

Henderson, W.B. and Berry, W.L., 1976. "Heuristic methods for telephone operator shift scheduling: an experimental analysis". *Management Science*, vol. 22, no. 12, 1372-1380.

Henderson, W.B. and Berry, W.L., 1977. “Determining optimal shift schedules for telephone traffic exchange operators”. *Decision Sciences*, vol. 8, no. 2, 239-255.

Hillier, F.S. and Lieberman, G.J., 1986. *Introduction to Operations Research*. 4th edn., Holden Day, Oakland, CA.

Holloran, T.J. and Byrn, J.E., 1986. “United Airlines station manpower planning system”. *Interfaces*, vol. 16, no. 1, 39-50.

Keith, E.G., 1979. “Operator scheduling”. *AIIE Transactions*, vol. 11, no. 1, 37-41.

Kolesar, P.J., Rider, K.L., Crabill, T.B. and Walker, W.E., 1975. “A queuing-linear programming approach to scheduling police patrol cars”. *Operations Research*, vol. 23, no. 6, 1045-1062.

Kwan, S.K., Davis, M.M. and Greenwood, A.G., 1988. “A simulation model for determining variable worker requirements in a service operation with time-dependent customer demand”. *Queuing Systems*, vol. 3, 265-275.

Paul, R.J. and Stevens, R.E., 1971. “Staffing service activities with waiting line models”. *Decision Sciences*, vol. 2, no. 2, 206-218.

Quinn, P., Andrews, B. and Parsons, H., 1991. “Allocating telecommunications resources at L.L. Bean”. *Interfaces*, vol. 21, no. 1, 75-91.

SAS Institute, 1988. SVfS, Version 6.03. SAS Institute, Inc., Cary, NC.

Segal, M., 1974. “The operator-scheduling problem: A network-flow approach”. *Operations Research*, vol. 22, no. 4, 808-823.

Sill, B.T., 1991. “Capacity management: Making your service delivery more productive”. *Cornell Hotel and Restaurant Administrative Quarterly*, vol. 31, no. 4, 76-87.

Taha, H.A., 1981. “Queuing theory in practice". *Interfaces*, vol. 11, no. 1, 43-49.

Thompson, G.M., 1990. “Shift scheduling when employees have limited availability: An LP approach”. *Journal of Operations Management*, vol. 9, no. 3, 352-370.

Thompson, G.M., 1992. “Improving the utilization of frontline service delivery system personnel”. *Decision Sciences*, vol. 23, no. 5, 1072-1098.