chapter fourteen

Managing Customer Waiting Lines and Reservations

Cutting the Wait for Customers in Retail Banking

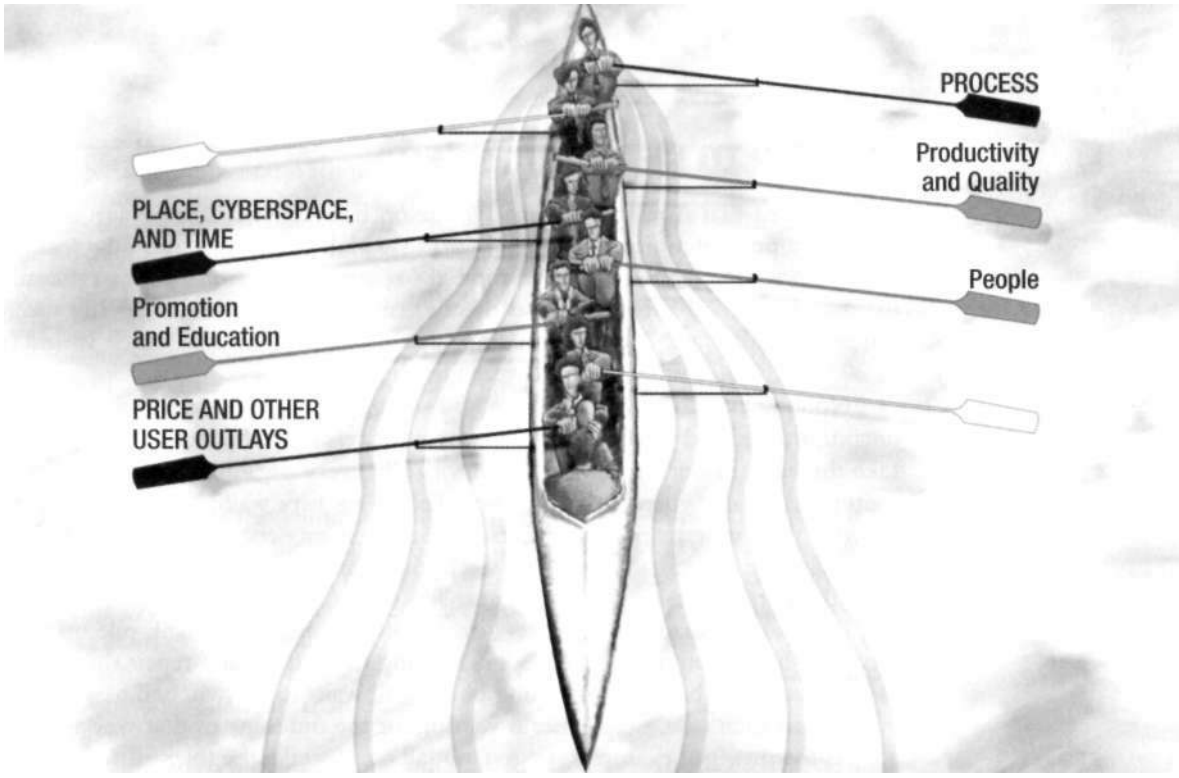
How should a big retail bank respond to increased competition from new financial service providers? A large bank in Chicago decided that enhancing service to its customers would be an important element in its strategy.[1] One opportunity for improvement was to reduce the amount of time that customers spent waiting in line for service in the bank's retail branches—a frequent source of complaints. Recognizing that no single action could resolve the problem satisfactorily, the bank adopted a three-pronged approach.

First, technological improvements were made to the service operation, starting with introduction of an electronic queuing system that not only routed customers to the next available teller station but also provided supervisors with online information to help match staffing to customer demand. Meantime, computer enhancements provided tellers with more information about their customers, enabling them to handle more requests without leaving their stations. And new cash machines for tellers saved them from selecting bills and counting them twice (yielding a time savings of 30 seconds for each cash withdrawal transaction).

Second, changes were made to human resource strategies. The bank adopted a new job description for teller managers that made them responsible for customer queuing times and expediting transactions. It created an officer-of-the-day program, where a designated officer was equipped with a beeper and assigned to help staff with complicated transactions that might otherwise slow them down. A new job category of peak-time teller was introduced, paying premium wages for 12 to 18 hours of work a week. Existing full-time tellers were given cash incentives and recognition to reward improved productivity on predicted high-volume days. Management also reorganized meal arrangements. On busy days, lunch breaks were reduced to half-hour periods and staff received catered meals; in addition, the bank cafeteria was opened earlier to serve peak-time tellers.

A third set of changes centered on customer-oriented improvements to the delivery system. Quick-drop desks were established on busy days to handle deposits and simple requests, while newly created express teller stations were reserved for deposits and check cashing. Lobby hours were expanded from 38 to 56 hours a week, including Sundays. A customer brochure, *How to Lose Wait,* alerted customers to busy periods and suggested ways of avoiding delays.

Subsequently, internal measures and customer surveys showed that the improvements had not only reduced customer wait times but also increased customer perceptions that this bank was "the best" bank in the region for minimal waits in teller lines. Studies also showed that extended lobby hours had transferred some of the "noon rush" customers to before-work and after-work time periods.



O Learning Objectives

After reading this chapter, you should be able to

=^:> recognize the different queue designs

=£> understand the psychology of waiting lines

=£> calculate expected waiting times under defined conditions

=£> know the basics of designing an effective reservation system

=£> discuss the principles of yield management and the use of segmented reservations strategies to improve profitability

WAITING TO GET PROCESSED

It's estimated that Americans spend 37 billion hours a year (an average of almost 150 hours per person) waiting in lines, "during which time they fret, fidget, and scowl," according to *The Washington Post*.[2] Similar (or worse) situations seem to prevail around the world. Richard Larson suggests that, when everything is added up, the average person may spend as much as half-an-hour per day waiting in line, which would translate to 20 months of waiting in an 80-year lifetime![3]

Nobody likes to be kept waiting. It's boring, wastes time, and is sometimes physically uncomfortable. And yet waiting for a service process is an almost universal phenomenon. Like the bank in our opening vignette, virtually every organization faces the problem of waiting lines somewhere in its operation. People are kept waiting on the phone, they line up with their supermarket carts to check out their grocery purchases, and they wait for their bills after a restaurant meal. They sit in their cars waiting for traffic lights to change, to enter drive-in car washes, and to pay at tollbooths.

Physical and inanimate objects wait for processing, too. E-mails pile up in an executive's in-box, equipment sits on racks waiting to be fixed at a repair shop, checks wait to be cleared at a bank, an incoming phone call waits to be switched to a customer service rep. In each case, a customer is waiting for the outcome of that work.

As the previous examples suggest, not all queues take the form of a physical waiting line in a single location. Some queues are geographically dispersed. For example, travelers wait at many different locations for the taxis they have ordered by phone to arrive and pick them up. And some queues are virtual rather than physical. When customers deal with a service supplier at arm's length, as in information-processing services, they interact from home, office, or college using telecommunication channels like voice telephone or the Internet. Calls are typically answered in the order received, often requiring customers to wait their turn in a virtual line. The advent of sophisticated Web sites has created additional opportunities for virtual waits. Although companies often promote the time savings that can be obtained, accessing the Web can sometimes be slow due to the virtual queuing that occurs when too many customers try to log on to a company's site or use the same server to go online simultaneously.

In an ideal world, nobody would ever have to wait to conduct a transaction at a service organization. But since services are performances, they can't typically be stored for later use during periods of excess demand. For example, a bank teller cannot prepackage a check-cashing transaction for the following day—it must be done in real time. This results in delays in service delivery when too many people want the same service at the same time.

As we discussed in Chapter 13, there are a number of ways to balance supply and demand. But what should a manager do when the possibilities for shaping demand and adjusting capacity have been exhausted, and supply and demand are still out of balance? Leaving customers to sort things out themselves is no recipe for service quality or customer satisfaction. Rather than allowing matters to degenerate into a random free-for-all, customer-oriented firms implement strategies for ensuring order, predictability, and fairness in their service delivery processes. In businesses where demand regularly exceeds supply, managers often try to manage demand in one of two ways: (1) by asking customers to wait in line (queuing), usually on a first-come, first-served basis; or (2) by offering them the opportunity to reserve or book space in advance.

queue: a line of people, vehicles, other physical objects, or intangible items waiting their turn to be served or processed.

The Nature of Queues

Waiting lines—known to operations researchers (and also the British) as **"queues"**[4]—occur whenever the number of arrivals at a facility exceeds the capacity of the system to process them. Queues are basically a symptom of unresolved capacity management

TABLE 14.1
Elements of a Queuing System

1. The *customer population*—the population from which demands for service originate (sometimes known to operations researchers as the "calling population")
2. The *arrival process*—the times and volumes of customer requests for service
3. *Balking*—a decision by an arriving customer not to join a queue
4. *Queue configuration*—the design of a system in terms of the number, location, and arrangement of waiting lines
5. *Reneging—a* decision by a customer already in a queue who has not yet been served to leave the line rather than wait any longer
6. *Customer selection policies*—formal or ad hoc policies about whom to serve next (also known as queue discipline)
7. The *service process*—the physical design of the service delivery system, the roles assigned to customers and service personnel, and the flexibility to vary system capacity

problems. The analysis and modeling of waiting lines is a well-established branch of operations management. Queuing theory has been traced back to 1917, when a Danish telephone engineer was given the responsibility of determining how large the switching unit in a telephone system had to be to keep the number of busy signals within reason.

Queuing systems can be divided into seven elements, as shown in Table 14.1. Let's take a look at each, recognizing that strategies for managing waiting lines can exercise more control over some elements than others.

Customer Population When planning queuing systems, operations managers need to know who their customers are and something about their needs and expectations. There is a big difference between a badly injured patient arriving at a hospital emergency unit and a sports fan arriving at a stadium ticket office—obviously, the hospital needs to be more geared for speed than the stadium. Based upon customer research, the population can often be divided into several distinct market segments, each with differing needs and priorities.

Arrival Process The rate at which customers arrive over time relative to the capacity of the serving process, and the extent to which they arrive individually or in clusters, will determine whether or not a queue starts to form. We need to draw a distinction between the *average* arrival rate (e.g., 60 customers per hour = one customer every minute) and the *distribution* of those arrivals during any given minute of that hour. In some instances, arrival times are largely random (for instance, individuals entering a store in a shopping mall). At other times, some degree of clustering can be predicted, such as arrivals of students in a cafeteria within a few minutes of classes ending. Managers who anticipate surges of activity at specific times can plan their staff allocations around such events (for instance, opening an additional checkout line).

Balking If you're like most people, you tend to be put off by a long line at a service facility and may decide to come back later (or go somewhere else) rather than waiting. Sometimes **"balking"is** a mistake, as the line may actually be moving faster than you think. Managers can disguise the length of lines by having them wind around corners, as often happens at theme parks like Disneyland. Alternatively, they may indicate the expected wait time from specific locations in the queuing area by installing information signs.

balking: a decision by a customer not to join a queue because the wait appears too long.

Queue Configuration There are a variety of different types of queues. Here are some common **queue configurations** that you may have experienced yourself in people-processing services (see Figure 14.1 for diagrams of each type).

queue configuration: the way in which a waiting line is organized.

- *Single line, single stage.* Customers wait to conduct a single service transaction. Waiting for a bus is an example of this type of queuing system.
- *Single line, sequential stages.* Customers proceed through several serving operations, as in a cafeteria line. In such systems, bottlenecks will occur at any stage

where the process takes longer to execute than at previous stages. Many cafeterias often have lines at the cash register because the cashier takes longer to calculate how much you owe and to make change than the servers take to put food on your plate (or you take to serve yourself).

>» *Parallel lines to multiple servers (single or sequential stages).* This system offers more than one serving station, allowing customers to select one of several lines in which to wait. Fast-food restaurants usually have several serving lines in operation at busy times of day, with each offering the full menu. A parallel system can have either a single stage or multiple stages. The disadvantage of this design is that lines may not move at equal speed. How many times have you chosen what looked like the shortest line only to watch in frustration as the lines on either side of you move at twice the speed because someone in your line has a complicated transaction?

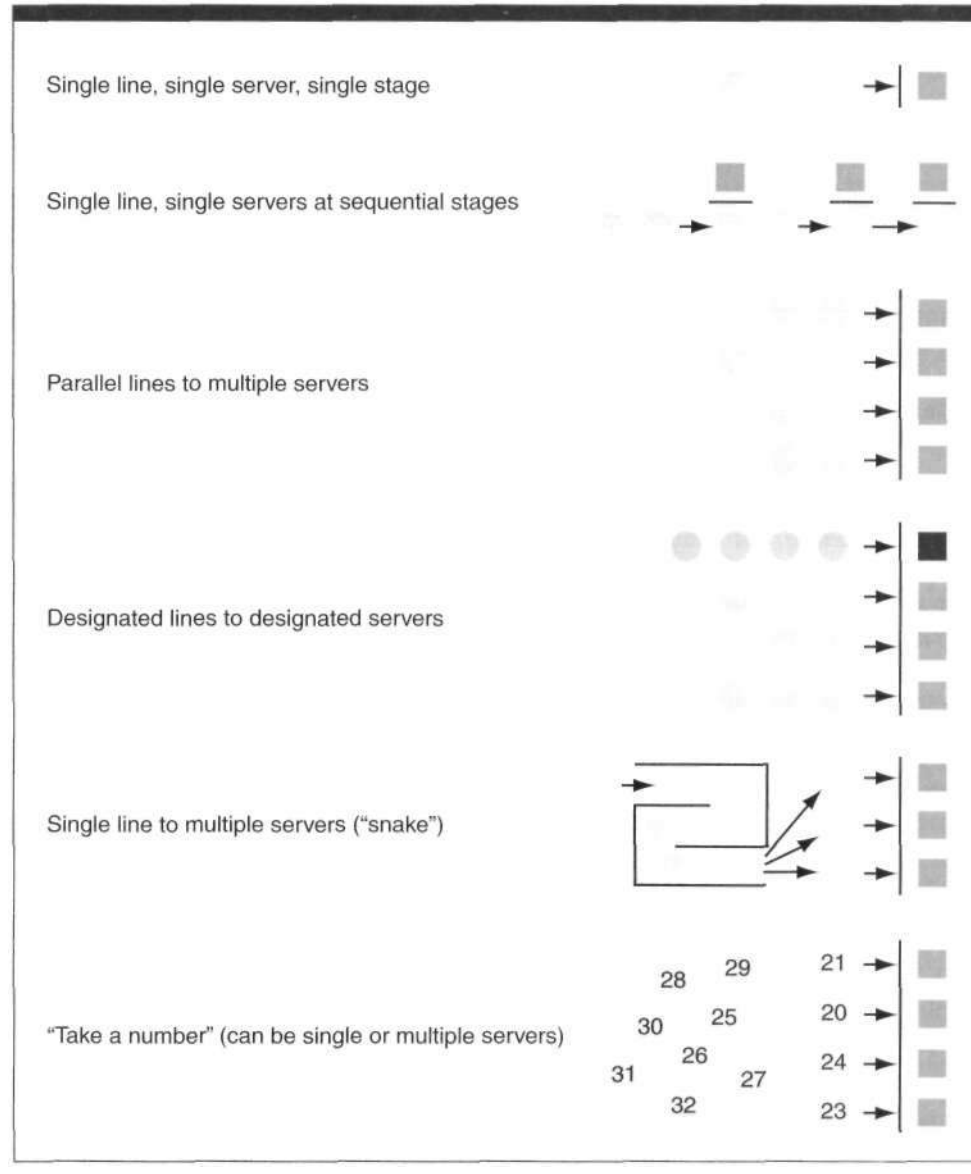


FIGURE 14.1
Alternative Queuing Configurations

5* *Designated lines.* Different lines can be assigned to specific categories of customer. Examples include express lines (six items or less) and regular lines at supermarket checkouts, and different check-in lines for first-class, business-class, and economy-class airline passengers.

>- *Single line to multiple servers ("snake").* Customers wait in a single line, often winding back and forth between rope barriers (hence the name). As each person reaches the head of the queue, he or she is directed to the next available serving position. This approach is encountered frequently in bank lobbies, post offices, and at airport check-ins. Its big advantages are fairness and reduced anxiety. The presence of ropes or other barriers makes it difficult for inconsiderate people to break into line. It may also discourage customers from leaving the line before being served.

>- *Take a number.* In this variation of the single line, arriving customers take a number and are then called in sequence, thus eliminating the need to stand in a queue. This procedure allows them to sit down and relax (if seating is available) or to guess how long the wait will be and do something else in the meantime—but risk losing their place. Users of this approach include ice cream parlors like Baskin-Robbins, large travel agents, or supermarket departments, such as the butcher or baker. Some restaurants use a high-tech version of this queuing strategy. For example, customers who are waiting for tables at the Olive Garden or Outback Steakhouse are given electronic pagers that are numbered by order of arrival. This provides them with more freedom in occupying themselves (e.g., window shopping if the restaurant is located in a mall with other stores) until their pagers vibrate, signaling that their tables are ready.

Hybrid approaches to queue configuration also exist. For instance, a cafeteria with a single serving line might offer two cash register stations at the final stage. Similarly, patients at a small medical clinic might visit a single receptionist for registration, proceed sequentially through multiple channels for testing, diagnosis, and treatment, and conclude by returning to a single line for payment at the receptionist's desk.

Reneging You know the situation—perhaps all too well! The line is not that long, but it's moving at a snail's pace. The person at the front of the queue has been there for at least five minutes and his problem seems nowhere near resolved. There are two other people ahead of you and you have an uneasy feeling that their transactions are not going to be brief either. You look at your watch for the third time and realize that you only have a few minutes left before your next appointment. Frustrated, you leave the line. In the language of queue management, you have reneged. It's important for service providers to determine how long a wait has to be before customers are likely to start **reneging,** because the consequences may include irritated customers who return later as well as business that is permanently lost.

reneging: a decision by a customer to leave a queue before reaching its end because the wait is longer or more burdensome than originally anticipated.

Customer Selection Policies Most waiting lines work on the principle of first come, first served. Customers tend to expect this—it's only fair, after all. In many cultures (but not all), people get very resentful if they see later arrivals being served ahead of them for no obvious reason. But not all queuing systems are organized on a first-come, first-served basis. Market segmentation is sometimes used to design queuing strategies that set different priorities for different types of customers. Allocation to separate queuing areas may be based on the following:

>- *Urgency of the job*—at many hospital emergency units, a triage nurse is assigned to greet incoming patients and decide which ones require priority medical treatment and which can safely be asked to register and then sit down while they

wait their turn. Airline personnel will allow passengers -whose flights are due to leave soon to check in ahead of passengers taking later flights.

>- *Duration of service transaction*—banks, supermarkets, and other retail services often provide "express lanes" for shorter, less-complicated tasks.

>- *Payment of a premium price*—airlines usually offer separate check-in lines for first-class and economy-class passengers, with a higher ratio of personnel to passengers in the first-class line (which results in reduced waits for those who have paid more for their tickets).

>» *Importance of the customer*—special processes may be reserved for members of frequent user clubs. National Car Rental provides express pickup and drop-off procedures for its Emerald Club members and promises these customers "no waiting, no paperwork, no hassles."[6]

Service Process Poorly designed service processes can lead to waits that are longer and more burdensome than necessary. The root cause is sometimes one or more backstage delays, resulting in customer-contact employees that are kept waiting for a necessary action to occur somewhere else in the system. Flowcharts, employee interviews, and analysis of past service failures can help pinpoint where such problems might occur. The physical design of the front-stage service delivery system also plays a key role in effective queue management. Important design issues include:

>- How customers are served (batch processes serve customers in groups, while flow processes serve them individually).

>- Whether personnel, self-service equipment, or a combination of the two will serve customers.

>- How fast service transactions can be executed, thus determining capacity.

*- Whether service comes to customers or whether they must come to the service site and move from one step to another.

>-The quality of the serving and waiting experiences, including personal comfort and design issues such as impression created by the servicescape.

MINIMIZING THE PERCEIVED LENGTH OF THE WAIT

As we've discussed in earlier chapters, customers may view the time and effort spent on consuming services as a burden. People don't like wasting their time on unproductive activities any more than they like -wasting money. They also prefer to avoid unwanted mental or physical effort, including anxiety or discomfort. Research shows that people often think they have waited longer for a service than they actually did. Studies of public transportation use, for instance, have shown that travelers perceive time spent waiting for a bus or train as passing one and a half to seven times more slowly than the time actually spent traveling in the vehicle.

The Psychology of Waiting Time

The noted philosopher William James observed: "Boredom results from being attentive to the passage of time itself." Based on this observation, David Maister formulated eight principles about waiting time.[8] Adding two additional principles gives us a total often, summarized in Table 14.2.

TABLE 14.2
Ten Propositions on the Psychology of Waiting Lines

1. Unoccupied time feels longer than occupied time.
2. **Pre-process** and **post-process** waits feel longer than **in-process** waits.
3. Anxiety makes waits seem longer.
4. Uncertain waits are longer than known, finite waits.
5. Unexplained waits are longer than explained waits.
6. Unfair waits are longer than equitable waits.
7. The more valuable the service, the longer people will wait.
8. Solo waits feel longer than group waits.
9. Physically uncomfortable waits feel longer than comfortable waits.[9]
10. Waits seem longer to new or occasional users than to frequent users.[10]

Unoccupied Time Feels Longer Than Occupied Time When you're sitting around with nothing to do, time seems to crawl. Thus many service organizations give customers something to do to distract them while waiting. Doctors and dentists stock their waiting rooms with piles of magazines for people to read while waiting. Car repair facilities may have a television for customers to watch. One tire dealer goes further, providing customers with free popcorn, soft drinks, coffee, and ice cream while they wait for their cars to be returned. Theme parks supply roving bands of entertainers to amuse customers waiting in line for the most popular attractions.

pre-process wait; a wait before service delivery begins.

in-process wait: a wait that occurs during service delivery.

post-process wait: a wait that occurs after service delivery has been completed.

Pre- and Post-Process Waits Feel Longer Than In-Process Waits There's a perceived difference between waiting to buy a ticket to enter a theme park and waiting to ride on a roller coaster once you're in the park. There's also a difference between waiting for coffee to arrive near the end of a restaurant meal and waiting for the server to bring you the check once you're ready to leave. Customers are typically more patient during the core service delivery process than before it starts or after it's completed.

Anxiety Makes Waits Seem Longer Can you remember waiting for someone to show up to meet you and worrying about whether you had the time and/or the location correct? This makes the perceived waiting time longer, because you are worried about whether you (or the person you're meeting) might have made a mistake.

Customers must wait in line even at fast-food restaurants, but they can pass the time studying the menu.

While waiting in unfamiliar locations, especially out-of-doors and after dark, people are often anxious about their personal safety.

Uncertain Waits Are Longer Than Known, Finite Waits Although any wait may be frustrating, we can usually adjust mentally to a wait of known length. It's the unknown that keeps us on edge. Maybe you've had the experience of waiting for a delayed flight when you haven't been told how long the delay is going to be. This is unsettling, because you don't know whether you have time to get up and walk around the terminal or whether to stay at the gate in case the flight is called any minute. Airlines often try to appease their customers by giving them new take-off times for delayed flights (which are usually extended several times before the aircraft actually leaves the gate).

Unexplained Waits Are Longer Than Explained Waits Have you ever been in a subway or an elevator that has stopped for no apparent reason? Not only is there uncertainty about the length of the wait, there's added worry about what is going to happen. Has there been an accident on the line? Will you have to exit the subway in the tunnel? Is the elevator broken? Will you be stuck for hours in close proximity with strangers?

Unfair Waits Are Longer Than Equitable Waits Expectations about what is fair or unfair sometimes vary from one culture or country to another. In America, Canada, or Britain, for example, people expect everybody to wait their turn in line and are likely to get irritated if they see others jumping ahead or being given priority for no apparent good reason. In some other countries, it is acceptable to push or shove to the front of a line to receive faster service.

The More Valuable the Service, the Longer People Will Wait People will queue overnight under uncomfortable conditions to get good seats at a major concert, movie opening, or sports event that is expected to sell out.

Solo Waits Feel Longer Than Group Waits It's reassuring to wait with one or more people you know. Conversation with friends can help to pass the time, and some people are comfortable conversing with strangers while they wait in line.

Physically Uncomfortable Waits Feel Longer Than Comfortable Waits "My feet are killing me!" is one of the most frequently heard comments when people are forced to stand in line for a long time. And whether sitting or standing, a wait seems more burdensome if the temperature is too hot or too cold, if it's drafty or windy, or if there is no protection from rain or snow.

Unfamiliar Waits Seem Longer Than Familiar Ones Frequent users of a service know what to expect and are less likely to worry while waiting. But new or occasional users of a service are often nervous, wondering about the probable length of the wait and what happens next. They may also be more concerned about such issues as personal safety.

What are the implications of these propositions about the psychology of waiting? When increasing capacity is not feasible, managers should look for ways to make waiting more palatable for customers. An experiment at a large bank in Boston found that installing an electronic news display in the lobby didn't reduce the perceived time spent waiting for teller service but it did lead to greater customer satisfaction.[11] Some large hotels now provide these digital news displays in their elevators to make rides less bor-

ing (in addition to the common practice of putting mirrors near the elevators on each floor to shorten the perceived pre-process wait). And the doorman at a Marriott Hotel in Boston has taken it upon himself to bring a combination barometer/thermometer to work each day, hanging it on a pillar at the hotel entrance where guests waiting can spend a moment or two examining it while they wait for a taxi or for their car to be delivered from the valet parking.[1"]

Heated shelters equipped with seats make it more pleasant to wait for a bus or a train in cold weather. Theme park operators cleverly design their waiting areas to make the wait look shorter than it really is, find ways to give customers in line the impression of constant progress, and make time seem to pass more quickly by keeping customers amused or diverted while they wait. Restaurants solve the waiting problem by inviting dinner guests to have a drink in the bar until their table is ready—an approach that makes money for the house as well as keeping customers occupied. In similar fashion, guests waiting in line for a show at a casino may find themselves queuing in a corridor lined with slot machines.

Giving Customers Information on Waits

Does it help to tell people how long they are likely to have to wait for service? Common sense would suggest that this is useful information for customers, since it allows them to make decisions about whether they should wait now or come back later. It also enables them to plan the use of their time while waiting. An experimental study in Canada looked at how students responded to waits while conducting transactions by computer—a situation similar to waiting on the telephone in that there are typically no visual clues as to the probable wait time. The study examined dissatisfaction with waits of 5, 10, or 15 minutes under three conditions: (1) the student subjects were told nothing, (2) they were told how long the wait was likely to be, or (3) they were told what their place in line was. The results suggested that for 5-minute waits, it was not necessary to provide information to improve satisfaction. For waits of 10 or 15 minutes, offering information appeared to improve customers' evaluations of service. However, for longer waits, the researchers suggest that it may be more positive to let people know how their place in line is changing than to let them know how much time remains before they will be served.

One conclusion we might draw is that people prefer to see (or sense) that the line is moving, rather than to watch the clock. Some companies have adopted this approach to manage the waits that customers encounter when dialing customer service numbers. Recorded messages tell the caller how many people are ahead in the queue—these messages are updated continuously until a customer service representative becomes available.

CALCULATING WAIT TIMES

Queue management involves extensive data gathering. Questions of interest include the rate at which customers (or objects requiring service) arrive per unit of time and how long it takes to serve each one. A typical operations strategy is to plan on the basis of average throughput in order to optimize use of employees and equipment. So long as customers (or objects) continue to arrive at the average rate, there will be no delays. However, fluctuations in arrivals (sometimes random, sometimes predictable) will lead to delays at times as the line backs up following a "clump" of arrivals. Planners need to know how easily customers will just walk away when they spot a lengthy line *(balking)* and how long customers will wait for service before giving up and leaving *(reneging).*

To streamline its check-in service at Boston's Logan International Airport, a major airline turned to MIT Professor Richard Larson, who heads a consulting firm called QED.[14] Technicians from QED installed pressure-sensitive rubber mats on the floor in front of the ticket counters. Pressure from each customer's foot on approaching or leaving the counter recorded the exact time on an electronic device embedded in the mats. From this data, Larson was able to profile the waiting situation at the airline's counters, including average waiting times, how long each transaction took, how many customers waited longer than a given length of time (and at what hours on what days), and even how many bailed out of a long line.This information, which was collected over a long time period, helped the airline plan its staffing levels to more closely match the demand levels projected at different times.

Analyzing Simple Queuing Systems

Complex mathematical models enable planners and consultants to calculate a variety of statistics about queue behavior and thus make informed decisions about changes or improvements to existing queuing systems. For basic queuing situations, the formulas are quite simple and yield interesting insights (see the boxed material on "Using Formulas to Calculate Statistics for Simple Queues"). More complex environments may require powerful simulation models that are beyond the scope of this book. Given certain information about a particular queuing situation, you can use these formulas to calculate such statistics as: (1) average queue length, (2) average wait times for customers, (3) average total time for customers in the service system, (4) the impact of increasing the number of service channels, and (5) the impact of reducing average serving time. The math is easy but requires reference to a one-page statistical table, which we have reproduced as an appendix at the end of the chapter.

Using Formulas to Calculate Statistics for Simple Queues

By using the information provided below and the table in the appendix at the end of this chapter, you will be able to make simple calculations about queue waiting times and how many people are likely to be waiting in a given queue under specified conditions. The formulas are very simple—don't be put off by the use of Greek letters for the notation!

Terminology

Certain terms and notation are used in queue analysis:

M = number of serving channels

A (lambda) = average number of customers actually arriving per unit of time (60 minutes)

ft (mu) = average number of customers per channel that can be served per unit of time (60 minutes)

p (rho) = *k/fi* = flow intensity through serving channel (% utilization)

U = *A/Mn* = capacity utilization of the overall facility

L_q = expected length of line (number of people or objects waiting)

$W_q = L_q/\Lambda$ = expected waiting time before being served

You should note that unless the average number of customers served (p) exceeds the average number of arrivals (A), it would never be possible to serve all the customers desiring service.

Example

Let's take a simple example. Consider the case of a theater ticket office that has one agent (M) who, on average, can serve 25 customers per hour (p). This implies an average serving time of 60/25 = 2.4 minutes per customer. Let's assume that customers arrive at an average rate of 20 per hour (A) in the busy period, which means that p = 20/25 = 0.80. We can now use the table in the appendix to calculate:

>- expected length of the line *(L_q):* Looking down the column for one serving line *(M)* to p = 0.80, we can see that the line length will average 3.2 persons.

Information Needs

Service managers require the following types of information in order to develop effective demand management strategies:

- *Historical data* on the level and composition of demand over time, including responses to changes in price or other marketing variables.
- *Forecasts* of the level of demand for each major segment under specified conditions.
- *Segment-by-segment data* to help management evaluate the impact of periodic cycles and random demand fluctuations.
- *Sound cost data* to enable the organization to distinguish between fixed and variable costs and to determine the relative profitability of incremental unit sales to different segments and at different prices.
- *Identification of meaningful variations in the levels and composition of demand* on a site-by-site basis in multi-site organizations.
- *Customer attitudes* toward queuing under various conditions.
- *Customer opinions* about whether service quality varies with different levels of capacity utilization.

Where might all this information come from? Although some new studies may be required, much of the needed data are probably already being collected within the organization—although not necessarily by marketers. A stream of information comes into most service businesses from distilling the multitude of individual transactions recorded on sales receipts and other routine business documents. Most companies also collect detailed information for operational and accounting purposes. Unfortunately, the marketing value of this data is often overlooked, and it is not always stored in ways that permit easy retrieval and analysis for marketing purposes. But customer transaction data can

- expected waiting time (IV): 3.2 x 60/20 = 9.6 minutes
- expected total time in system (IV + 60//A): 9.6 minutes + 2.4 minutes = 12.0 minutes
- average capacity utilization *(U): \/Mp,* = 20/(1 x 25) = 80%

(In other words, 20 percent of the time, the agent will be idle.)

Let's suppose that customers are complaining about this wait and management wants to speed up service. The choices are to add a second agent with a separate single line of customers so that M= 2, or to purchase new equipment that halves the time required to issue a ticket and receive payment. Here are the comparative results:

(1) Using the table in the appendix, when $M = 2$ (indicating the addition of a second agent) and $p = 0.80$:
   - the expected line length (/.) will be only 0.15 persons
   - the expected wait $(W_q) = L_q/k$ = 0.15 x 60/20 = 0.45 minutes, plus 2.40 minutes for service = 2.85 minutes (down from 12.0 minutes)

(2) However, if we halve the service process time from 2.4 to 1.2 minutes by adding new equipment, we can now serve a maximum of 50 customers per hour per channel and the following results occur:
   - the expected line length, when M= 1 and p = 20/50 = 0.4 is 0.27 persons
   - the expected wait is 0.27 x 60/20 = 0.81 minutes +1.2 = 2.01 minutes total

Both approaches cut the time sharply, but halving the service process time yields slightly better time savings than doubling the number of channels. In this instance, the decision on which approach to adopt would probably depend on the relevant costs involved—the capital cost of adding a second channel plus the wages and benefits paid to a second employee, versus the capital costs of investing in new technology and training (assuming no increase in wages).

often be reformatted to provide marketers with some of the information they require, including how existing segments have responded to past changes in marketing variables.

RESERVATIONS

Ask someone what services come to mind when you talk about reservations and most likely they will list airlines, hotels, restaurants, car rentals, and theater seats. Suggest synonyms like "bookings" or "appointments" and they may add haircuts, visits to professionals like doctors and consultants, vacation rentals, and service calls to fix anything from a broken refrigerator to a neurotic computer.

Reservations are intended to guarantee that service will be available when the customer wants it. Systems vary from a simple appointment book using handwritten entries to a central, computerized data bank for a company's worldwide operations. Reservations systems enable demand to be controlled and smoothed out in a more manageable way. They can also help pre-sell services and provide opportunities to inform and educate customers. A well-organized reservations system allows an organization to deflect demand for service from a first-choice time to earlier or later times, from one class of service to another ("upgrades" and "downgrades"), and even from first-choice locations to alternative ones.

Reservations systems are necessary for possession-processing businesses in fields like repair and maintenance. By requiring reservations for routine maintenance, management can ensure that some time will be kept free for handling emergency jobs that generate much higher margins because they carry a premium price. Households with only one car, for example, or factories with a vital piece of equipment often cannot afford to be without such items for more than a day or two and are likely to be willing to pay more for faster service.

Reservation systems are also used by many people-processing services including restaurants, hotels, airlines, hair salons, doctors, and dentists. Customers who hold reservations should be able to count on avoiding a queue, since they have been guaranteed service at a specific time. However, problems arise when customers fail to show or when service firms over-book. Marketing strategies for dealing with these operational problems include requiring a deposit, canceling nonpaid bookings after a certain time, and providing compensation to victims of over-booking.

The challenge in designing reservation systems is to make them fast and user-friendly for both staff and customers. Whether customers talk with a reservations agent or make their own bookings through a company's Web site, they want quick answers to queries about service availability at a preferred time. They also appreciate it if the system is designed to provide further information about the type of service they are reserving. For instance, can a hotel's reservation system assign a certain type of room for a specific date? (For example, can it guarantee a nonsmoking room with a queen-sized bed and a view of the lake, rather than one with two twin beds and a view of the nearby power station?)

Using Reservations Systems to Manage Yield

Service organizations often use percentage of capacity sold as a measure of operational efficiency. Transport services talk of the "load factor" achieved, hotels of their "occupancy rate," and hospitals of their "census." Professional firms calculate what proportion of a partner's or an employee's time can be classified as billable hours, and repair shops can look at utilization of both equipment and labor. By themselves, however, these percentage figures tell us little of the relative profitability of the business attracted, since high utilization rates may be obtained at the expense of heavy discounting—or even outright giveaways.

yield: the average revenue received per unit of capacity offered for sale.

Many service firms prefer to rely on measurements of their **yield**—that is, the average revenue received per unit of capacity. The goal is to maximize yield in order to improve profitability. As we noted in Chapter 8, pricing strategies designed to achieve this goal are

Getting there is half the fun: Passengers wait to check in at the airport.

widely used in capacity-constrained businesses like passenger airlines, hotels, and car rental agencies. Formalized yield management programs based upon mathematical modeling provide the greatest value to service firms that find it expensive to modify their capacity but incur relatively low costs when they sell another unit of available capacity.[15] Other characteristics encouraging use of such programs include fluctuating demand levels, ability to segment markets by extent of price sensitivity, and sale of services well in advance of usage.

Yield analysis forces managers to recognize the **opportunity cost** of accepting business from one customer or market segment when another might subsequently yield a higher rate. Consider the following problems facing sales managers for different types of capacity-constrained service organizations:

opportunity cost: the potential value of the income or other benefits foregone as a result of choosing one course of action instead of other alternatives.

- Should a hotel accept an advance booking for 200 room nights from a tour group at $80 each when these same room nights might be sold later at short notice to business travelers at the full posted rate of $140?
- Should a railroad with 30 empty freight cars accept an immediate request for a shipment worth $900 per car or hold the cars idle for a few more days in the hope of getting a priority shipment that would be twice as profitable?
- How many seats on a particular flight should an airline sell in advance at special excursion fares or discounted rates?
- Should an industrial repair and maintenance shop reserve a certain proportion of productive capacity each day for emergency repair jobs that offer a high contribution margin and the potential to build long-term customer loyalty? Or should it simply make sure that there are sufficient jobs—involving mostly routine maintenance—to keep its employees fully occupied?
- Should a print shop process all jobs on a first-come, first-served basis, with a guaranteed delivery time for each job? Alternatively, should it charge a premium rate for "rush" work and tell customers with "standard" jobs to expect some variability in completion dates?

Managers who make these types of decisions on the basis of guesswork and "gut feel" are little better than gamblers who bet on rolls of the dice. They need a systematic

way to figure out the chances of getting more profitable business if they wait. The decision to accept or reject business should be based on a realistic estimate of the probabilities of obtaining more profitable business in the future and the need to maintain established (and desirable) customer relationships.

Segmenting Capacity for Reservations Purposes

There has to be a clear plan, based on analysis of past performance and current market data, that indicates how much capacity should be allocated on particular dates to different types of customers at certain prices. Based on this plan, "selective sell" targets can be assigned to advertising and sales personnel, reflecting allocation of available capacity among different market segments on specific future dates.The last thing a firm wants its sales force to do is to encourage price-sensitive market segments to buy capacity on dates when sales projections predict that there will be strong demand from customers willing to pay full price. Unfortunately, in some industries the least-profitable customers often book the furthest ahead.Tour groups, which pay much lower room rates than individual travelers, frequently ask airlines and hotels to reserve space more than a year in advance.

Figure 14.2 illustrates capacity allocation based on systematic yield analysis in a hotel setting, where demand from different types of customers varies not only by day of the week but also by season. These allocation decisions by segment, captured in reservation databases that are accessible worldwide, tell reservations personnel when to stop accepting reservations at certain prices, even though many rooms may still remain unbooked. Charts similar to those presented in Figure 14.2 can be constructed for most capacity-constrained businesses.

Advances in software and computing power have made it possible for managers to use sophisticated mathematical models to address complicated yield management issues. In the case of an airline, for example, these models can integrate massive historical databases on past passenger travel with real-time information on current bookings. The output helps analysts predict how many passengers would want to travel between two cities at a partic-

Pricing Seats on Flight 2015

American Airlines 2015 is a popular flight from Chicago to Phoenix, departing daily from the "windy city" at 5:30 P.M. The 125 seats in coach (economy class) are divided into seven fare categories, referred to by yield management specialists as "buckets," with round-trip ticket prices ranging from $238 for a bargain excursion fare (with various restrictions and a cancellation penalty attached) to an unrestricted fare of $1404. Seats are also available at a higher price in the small first-class section. Scott McCartney tells how ongoing analysis changes the allocation of seats between each of the seven buckets in economy class:

> *In the weeks before each Chicago-Phoenix flight, American's yield management computers constantly adjust the number of seats in each bucket, taking into account tickets sold, historical ridership patterns, and connecting passengers likely to use the route as one leg of a longer trip. If advance bookings are slim, American adds seats to low-fare buckets. If business customers buy unrestricted fares earlier than expected, the yield management computer takes seats out of the discount buckets and preserves them for last-minute bookings that the database predicts will still show up.*
>
> *With 69 of 125 coach seats already sold four weeks before one recent departure of Flight 2015, American's computer began to limit the number of seats in lower-priced buckets. A week later, it totally shut off sales for the bottom three buckets, priced $300 or less. To a Chicago customer looking for a cheap seat, the flight was "sold out."*
>
> *One day before departure, with 130 passengers booked for the 125-seat flight, American still offered five seats at full fare because its computer database indicated 10 passengers were likely not to show up or take other flights. Flight 2015 departed full and no one was bumped.*

Source: Scott McCartney, "Ticket Shock: Business Fares Increase Even as Leisure Travel Keeps Getting Cheaper," *Wall Street Journal,* 3 November, 1997, A1, A10.

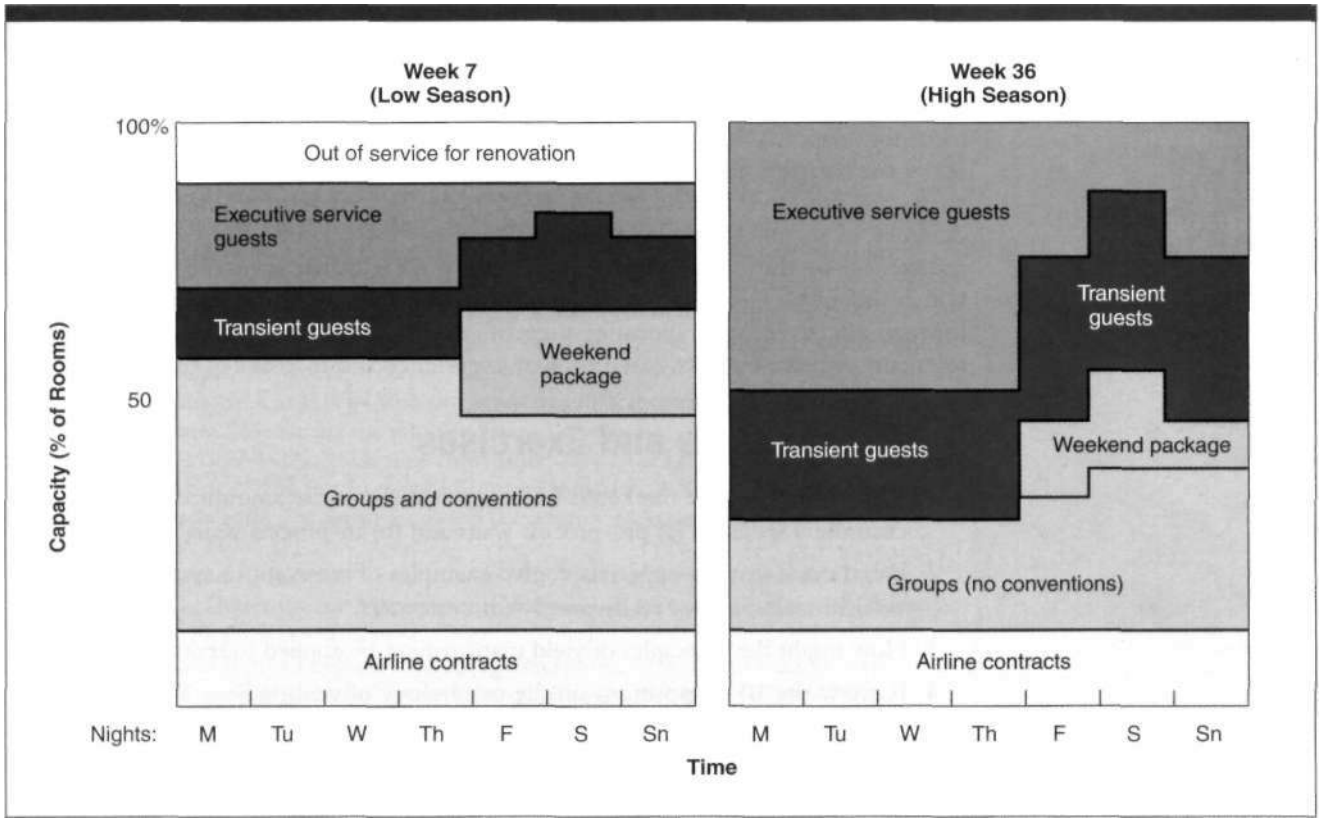


FIGURE 14.2
Setting Capacity Allocation Sales Targets over Time

ular fare on a flight leaving at a specified time and date. The boxed example describes how American Airlines uses yield management analysis to set fares for a specific flight.

There's evidence that yield management programs can improve revenues significantly—many airlines report increases of 5 percent or more after starting such programs. But a word of warning is in order at this point. Yield management shouldn't mean blind pursuit of short-term yield maximization. Over-dependence on the output of computer models can easily lead to pricing strategies that are full of rules and regulations, cancellation penalties, and a cynical strategy of overbooking without thought for disappointed customers who believed they had a firm reservation. To maintain goodwill and build relationships, a company should take a long-term perspective. Managers need to build in pricing strategies for retaining valued customer relationships, even to the extent of not charging the maximum feasible amount on a given transaction. After all, customer perceptions of "price gouging" do not build trust. And as we mentioned in an earlier chapter, firms shouldn't make pricing policies too complex. Jokes abound about travel agents having nervous breakdowns because they get a different quote every time they call the airline for a fare, and because there are so many exclusions, conditions, and special offers. Finally, yield management strategies should include thoughtfully planned contingencies for victims of overbooking, with service recovery efforts designed to restore goodwill when customers have been disappointed.

Conclusion

The time-bound nature of services is a critical management issue today, especially since customers are becoming more conscious of their personal time constraints and availability. When demand exceeds capacity, not all customers can be served immediately.

Waiting lines and reservations are ways of inventorying demand until capacity is available. Advance reservations can shape the timing of arrivals, but sometimes queuing is inevitable. People-processing services are particularly likely to impose the burden of unwanted waiting on their customers, since the latter cannot avoid coming to the "factory" for service. Managers who can adopt strategies to save customers time (or at least make time in the queue pass more pleasantly) may be able to create a competitive advantage for their organizations. Both queuing and reservations systems can be designed to segment customers into different groups, according to the nature of their transaction or the desirability of their business. Yield management strategies, under which different customers pay different prices for effectively the same service, depend for their effectiveness on allocating units of capacity for reservations purposes to specific segments or price buckets, based on past experience and forecasts of future sales.

Study Questions and Exercises

1. Why should service marketers be concerned about the amount of time that customers spend in (a) pre-process waits and (b) in-process waits?
2. Based on your own experience, give examples of reservations systems that worked really well or really poorly for customers.
3. How might the principles of yield management be applied to rental car companies?
4. Review the 10 propositions on the psychology of waiting lines. Which are the most relevant in (a) a supermarket, (b) a city bus stop on a cold, dark evening, (c) check-in for a flight at the airport, (d) a doctor's office, (e) a ticket line for a football game that is expected to be a sell-out?
5. What are the seven elements of a queuing system? Which are under the control of the customer and which does the service provider control?
6. For an organization serving a large number of customers, what do you see as the advantages and disadvantages of the different types of queues shown in Figure 14.1?
7. Using the formulas on page 312 and the table in the appendix, calculate answers to the following problems:
   a. At Frank's office cafeteria, customers select their meals from different food stations and then go to the checkout station to pay. He knows that Maureen, the speedy cashier, can check out a customer every 20 seconds on average. With an arrival rate of 90 customers an hour during the 11 A.M. to 2 P.M. lunch period, what is the average length of the line that Frank can expect at the checkout? How many minutes will he have to wait?
   b. Maureen goes on maternity leave and is replaced by Willy, whom Frank times at one customer every 36 seconds. On average, how much longer will the line now be and how long will Frank have to wait?
   c. In response to complaints about delays at the checkout station, management assigns JoAnn to operate a second cash register during Maureen's absence. Like Willy, JoAnn can process the average customer in 36 seconds. How long, on average, will each line now be and how many minutes can Frank expect to wait (in either line)?
   d. Willy is off sick one day, so JoAnn must work alone. But she manages to improve her performance and to process one customer every 30 seconds. On average, how long is the line now? And how long is the wait?
8. What segmentation principles and variables are illustrated in the yield management example from American Airlines?

Endnotes

1. Based on an example in Leonard L. Berry and Linda R. Cooper, "Competing with Time-Saving Service," *Business* 40, no. 2, (1990): 3—7.
2. Malcolm Galdwell,"The Bottom Line for Lots ofTime Spent in America," *The Washington Post* (syndicated article, February, 1993).
3. Dave Wielenga, "Not So Fine Lines," *Los Angeles Times,* 28 November, 1997, El.
4. This section is based in part on James A. Fitzsimmons and Mona J. Fitzsimmons, *Service Management: Operations, Strategy and Information Technology* 2nd ed. (NewYork: Irwin McGraw-Hill, 1998): 515-537; and David H. Maister,"Note on the Management of Queues" 9-680-053, Harvard Business School Case Services, 1979, rev. 2/84.
5. Richard Saltus, "Lines, Lines, Lines, Lines . . .The Experts Are Trying to Ease the Wait," *The Boston Globe,* 5 October, 1992, 39, 42.
6. From the National Car Rental Web site, www.nationalcar.com, January 2001.
7. Jay R. Chernow, "Measuring the Values ofTravelTime Savings,JowrMd/ *of Consumer Research 1* (March 1981): 360-371. [Note: this entire issue was devoted to the consumption of time.]
8. David H. Maister, "The Psychology ofWaiting Lines," in J. A. Czepiel, M. R. Solomon, and C. F. Surprenant, *The Service Encounter* (Lexington, MA: Lexington Books/DC. Heath, 1986): 113-123.
9. M. M. Davis and J. Heineke, "Understanding the Roles of the Customer and the Operation for Better Queue Management," *International Journal of Operations & Production Management* 14, no. 5 (1994): 21-34.
10. Peter Jones and Emma Peppiatt, "Managing Perceptions ofWaiting Times in Service Queues," *International Journal of Service Industry Management* 7, no. 5 (1996): 47—61.
11. Karen L. Katz, Blaire M. Larson, and Richard C. Larson, "Prescription for the Waiting-in-Line Blues: Entertain, Enlighten, and Engage," *Sloan Management Review* (Winter 1991): 44—53.
12. Bill Fromm and Len Schlesinger, *Tlie Real Heroes of Business and Not a CEO AmongThem* (New York: Currency Doubleday, 1994), 7.
13. Michael K. Hui and David K.Tse,"What toTell Customers in Waits of Different Lengths: An Integrative Model of Service Evaluation," *Journal of Marketing* 80, no. 2 (April 1996): 81—90.
14. Malcolm Galdwell, "The Bottom Line for Lots ofTime Spent in America.
15. Sheryl E. Kimes, "Yield Management: A Tool for Capacity-Constrained Service Firms," *Journal of Operations Management* 8, no. 4 (October 1989): 348—363; Sheryl E. Kimes and Richard B. Chase, "The Strategic Levers of Yield Management," *Journal of Service Research* 1 (November 1998): 156-166.

Appendix: Poisson Distribution Table

Calculating the Expected Number of People Waiting in Line for Various Values of Mandp

Flow Intensity (p)	Number of Service Channels (M)			
0.10	0.0111			
0.15	0.0264	0.0008		
0.20	0.0500	0.0020		
0.25	0.0833	0.0039		
0.30	0.1285	0.0069		
0.35	0.1884	0.0110		
0.40	0.2666	0.0166		
0.45	0.3681	0.0239	0.0019	
0.50	0.5000	0.0333	0.0030	
0.55	0.6722	0.0149	0.0043	
0.60	0.9000	0.0593	0.0061	
0.65	1.2071	0.0767	0.0084	
0.70	1.6333	0.0976	0.0112	
0.75	2.2500	0.1227	0.0147	
0.80	3.2000	0.1523	0.0189	
0.85	4.8166	0.1873	0.0239	0.0031
0.90	8.1000	0.2285	0.0300	0.0041
0.95	18.0500	0.2767	0.0371	0.0053
10		0.3333	0.0454	0.0067