Labor Scheduling, Part 2: Knowing How Many On-duty Employees to Schedule

Gary Thompson
Cornell University School of Hotel Administration, gmt1@cornell.edu

Follow this and additional works at: https://scholarship.sha.cornell.edu/articles

Part of the Hospitality Administration and Management Commons

Recommended Citation

This Article or Chapter is brought to you for free and open access by the School of Hotel Administration Collection at The Scholarly Commons. It has been accepted for inclusion in Articles and Chapters by an authorized administrator of The Scholarly Commons. For more information, please contact hotellibrary@cornell.edu.

If you have a disability and are having trouble accessing information on this website or need materials in an alternate format, contact web-accessibility@cornell.edu for assistance.
Labor Scheduling, Part 2: Knowing How Many On-duty Employees to Schedule

Abstract
[Excerpt] You’ve figured out how many customers to expect, and now you need to decide how many employees to schedule. Alas, to do that well is more complicated than you might think.

Labor scheduling constitutes a large portion of the costs under the control of front-line managers in service organizations. Scheduling too few employees can result in poor service, overworked employees, and lost sales. Scheduling too many employee hours can reduce operating margins. Moreover, one needs to schedule employees who have the necessary skills and to try to honor employees’ requests for specific work hours. The manager has to take into account customer demands, employee skills, and employee work-hour requests—not to mention government regulations, company policies, and contractual obligations—while keeping an eye on profits.

Keywords
schedules, service systems, queuing models, staffing levels

Disciplines
Hospitality Administration and Management

Comments
Required Publisher Statement
© Cornell University. Reprinted with permission. All rights reserved.
Part 2

Knowing How Many On-duty Employees to Schedule

by Gary M. Thompson

Labor scheduling constitutes a large portion of the costs under the control of front-line managers in service organizations. Scheduling too few employees can result in poor service, overworked employees, and lost sales. Scheduling too many employee hours can reduce operating margins. Moreover, one needs to schedule employees who have the necessary skills and to try to honor employees’ requests for specific work hours. The manager has to take into account customer demands, employee skills, and employee work-hour requests—not to mention government regulations, company policies, and contractual obligations—while keeping an eye on profits.

This article is the second in a four-part series that focuses on the steps of workforce scheduling:

1. forecasting customer demand,
2. calculating employee requirements,
3. scheduling employees,
4. fine-tuning the schedule in real time.

In this paper I show how one might tackle the second step.


Gary M. Thompson, Ph.D., an associate professor of operations management at the Cornell University School of Hotel Administration, has written a four-part series on labor management, of which this is Part 2; Part 1 appeared in the October 1998 Cornell Quarterly.

© 1998, Cornell University
setting the number of employee hours needed to serve customers adequately during a given time period. The second task of workforce scheduling uses as its input the forecasts of demand discussed in Part 1 of this series of articles. The output of the second task, the number of on-duty employees needed, will become the input to the third task—the development of the schedule—which I will discuss in Part 3.

The output of step two depends on whether the work is controllable. Uncontrollable work is work over which managers and employees have no temporal control. The serving of customers by the wait staff and the preparation of made-to-order meals in a full-service restaurant are classic examples of uncontrollable work. Controllable work is work over which there is some temporal control—for example, the preparation of rooms by a hotel’s housekeeping staff.

For uncontrollable work, the process of translating demand forecasts into employee requirements should result in the ideal number of employees working in each planning period. That process can also identify the effect of deviating from the ideal staff size. For controllable work the translation process specifies the total workload in labor-hours (or a similar measure) and the window during which the work can occur; for example, for turning over guest rooms, from just after a guest departs to just before the next guest arrives. The window is based on forecasts of the events that determine those two points.

The Three Approaches
There are three basic approaches to translating demand forecasts into employee requirements: using productivity standards, using service standards, and using economic standards.

Productivity standards. The aim of using productivity standards is to define and then rely on consistent (and reasonable) productivity from employees. Productivity standards are the easiest means of translating demand forecasts into employee requirements. An example of a productivity standard in a restaurant might be that a server can handle 14 customers an hour; a productivity standard in a hotel might be that a housekeeper can process 15 rooms a day.

Productivity standards are easily applicable for controllable work. Since an employee performing controllable work does not have to wait for customers to arrive, the employee can work uninterrupted. For example, if housekeepers are scheduled correctly, they can work uninterrupted, since they will not have to wait for guests to check out.

If, however, one cannot exactly predict when customers will arrive (that is, if the work is uncontrollable), one needs to allow for idle time. For example, if servers are scheduled correctly, they can work uninterrupted, since they will not have to wait for guests to check out.

For example, a server may actually be able to handle 16 customers an hour, yet the productivity standard might be set at one server for every 14 customers. The difference between the maximum service level and the productivity standard is the planned idle time.

The challenge of using productivity standards for uncontrollable work is that planned idle time, and therefore the productivity standard itself, should not be a constant. During times when there are more customers, one can more fully use employees while providing the same level of service. This is possible because having more employees on duty increases the flexibility inherent in the work-management system.

Service standards. The aim of using service standards is to deliver a consistent level of customer service, regardless of the time of day. Service can be measured in many ways. Some possibilities are the average length of time customers wait for service, the average number of customers waiting for service, and the percentage of customers who have to wait more than a specified amount of time for service. Implicit in service standards is the recognition that there will be less idle time per employee when the workload is high than when it is low.

The most difficult aspect of developing service standards is determining an appropriate level of service. Customer surveys, focus groups, direct observation, and experimentation can help set the standard.

Economic standards. The aim of using an economic standard is to deliver the service most economically. Typically that means delivering better service at high-demand times than at low-demand times. Better service is economically warranted at high-demand times because it is experienced by more customers. Conversely, the cost of delivering such high service to the few customers arriving in low-demand periods may outweigh the benefits.

Economic standards vary in complexity. A straightforward approach is to estimate the cost of having customers wait for the service. For example, customer-waiting time would be valued much lower in a QSR restaurant where the average sale is relatively small compared to the value of time for those customers kept waiting on the telephone

---


Using a Queuing Model

There are two key tools for determining employee requirements with service standards and economic standards: queuing models and simulation.

Queuing models are formulas that describe the performance of a service system. Many queuing models have been developed using various sets of assumptions. One of the most famous queuing models describes a multiple-server, single-queue system. The model has three inputs: the customer-arrival rate ($\lambda$); the rate at which a server can serve customers ($\mu$); and the number of servers ($m$). These formulas describe the performance of the system:

- Server utilization: $\rho = \frac{\lambda}{m \mu}$
- Probability of zero customers in the system: $P_0 = \left[ \sum_{n=0}^{m-1} \frac{\left( \frac{\lambda}{\mu} \right)^n}{n!} \cdot \left( 1 - \frac{\lambda}{\mu} \right)^m \right]^{-1}$
- Probability of $n$ customers in the system: $P_n = \left\{ \begin{array}{ll} \frac{\left( \frac{\lambda}{\mu} \right)^n}{n!} & \text{if } 0 \leq n \leq m \\ \frac{m!}{m-m} & \text{if } n \geq m \end{array} \right.$
- Average number of customers in the queue: $L_q = \frac{P_0 \left( \frac{\lambda}{\mu} \right)^m \rho}{m! (1 - \rho)^2}$
- Average time a customer spends in the queue: $W_q = \frac{L_q}{\lambda}$
- Average number of customers in the system: $L_s = L_q + \frac{1}{\mu}$
- Average time a customer spends in the system: $W_s = W_q + \frac{1}{\mu}$
- Probability that a customer’s wait exceeds $t$: $P(W_q > t) = \left( 1 - \sum_{n=0}^{m-1} P_n \right) e^{-\frac{\mu t}{m} (1 - \rho)}$

Take, for example, a case in which customers arrive at the rate of 58.7 an hour, each employee can serve 16 customers an hour, and there are five servers:

- Server utilization = 0.734 (73.4%)
- Probability of zero customers in the system = 0.021 (2.1%)
- Average number of customers in the queue = 1.19
- Average time a customer spends in the queue = 0.020 hrs (1.22 min)
- Average number of customers in the system = 4.86
- Average time a customer spends in the system = 0.083 hrs (4.97 min)
- $P(W_q > 3)$ (wait time exceeds three minutes [0.05 hours]) = 0.149

The second key tool for determining employee requirements with service standards and economic standards is simulation. Although it is more robust, it is harder to implement.—G.M.T.

A Comparison of the Standards

To illustrate the differences in the three labor standards, I created two scenarios.

For reservations agents for an upscale resort hotel. The economic standard would then be used to determine the staffing levels that result in the lowest cost of delivering the service, taking into account labor cost and customer-waiting cost. A more-complex approach is discussed later in this paper.

Economic standards have been applied in quick-service restaurants and telemarketing with purported success. That success, I believe, arises for two reasons. The more important one is that overall service has been improved. The other is that the service has been tailored to the volume. The major factor—improved service—suggests that service standards are commonly set too low in many businesses that use service standards.

The ease of implementing economic standards varies. For example, applying economic standards to a hotel’s telephone-reservation center would be easier than applying them to its concierge position. One can determine the number of lost calls and estimate the average call value for telephone reservations. It would be difficult, however, to determine the cost of having customers wait for a concierge. At their most complex level, using economic standards requires defining the link between customer-waiting time and long-run (future) business volume. Doing so is the most difficult aspect of using economic standards and may be the reason they are not widely applied.

**Scenario one.** In the first scenario the prime labor driver is the number of customers. The hourly wage, including benefits, is $10. I considered volumes of business from zero to 700 customer arrivals an hour. Working at 100-percent capacity, employees can serve 16 customers an hour.

I set the productivity standard at 14 customers an hour for each employee; that is, an additional employee is scheduled for every additional 14 customers. In determining the number of employees to schedule, fractions are rounded up to the next whole number. For example, if there are 29 customers an hour, three employees are scheduled. Using that standard, I determined how many employees would be scheduled for each of the volumes of business.

The staffing levels set by the productivity standard gave an average waiting time of 0.8 minutes. To make the service standard comparable to the productivity standard, I then used a service standard such that customers should be served with an average wait not to exceed 0.8 minutes. (See the box about using a queuing model on page 28.)

To ensure a fair, unbiased comparison among the three standards (i.e., productivity, service, and economic), I performed an involved analysis to determine a reasonable customer-waiting cost that could be used with all three standards. I found that the particular productivity standard defined above was consistent with a customer-waiting cost of $13.46 per hour. Thus, I used that value to compare the three standards, even though the $13.46 may not be an accurate economic standard. Later in this paper I discuss how to determine the appropriate economic standards.

Compared with the staffing levels specified by the economic standard, the productivity-standard approach tends to understaff for many customer-arrival rates (see Exhibit 1). The service standard approach tends to overstaff at low business volumes and increasingly understaff as business volumes increase (based on the specific standard used here, namely, an average wait of less than 0.8 minutes). The staffing levels specified by either the productivity or the service standards do not match those of the economic standard because the former standards remain constant across the range of business volumes.

Since the staffing levels determined by using the economic standard provide the best economic performance, any deviations from those levels will show up as inferior economic performance (see Exhibit 2, on the next page). When fewer than the “ideal” number of workers are scheduled, labor costs are kept low, but those savings are more than
Exhibit 2

**Scenario one:** How much does the cost of the productivity- and service-standard approaches differ from that of the economic-standard approach?

offset by the higher cost of customers' waiting. Conversely, when more than the "ideal" number of workers are scheduled, labor costs are higher than they need to be and those additional labor costs are more than the savings earned through customers' reduced waiting costs.

The economic performance of the productivity and service standards match that of the economic standard only in narrow ranges of customer-arrival rates. The service standard's understaffing at high customer-arrival rates and its overstaffing at low customer volumes both decrease the economic performance. The productivity standard's higher costs across a broad range of arrival rates are a result of the costs of poor service associated with its frequent understaffing.

**Scenario two.** I then considered a second scenario, where the productivity and service standards were inconsistent with the economic value of customers. I set the customer-waiting cost at twice its previous value, or $26.92 per hour; that is, twice as high as the valuation of customers implied by the productivity and service standards.

Both the productivity standard and the service standard consistently understaffed compared to the economic standard's ideal staffing levels as determined by optimum income for the period (see Exhibit 3). The service standard was particularly troublesome; it understaffed by as many as five employees. Setting the service standard higher—that is, reducing the amount of time customers wait in line—would reduce the amount by which the service standard understaffs at high customer volumes, but it would increase the amount that it overstaffs at low customer volumes.

Given that the results show that the economic standard is the best one to use, one may say "That's obvious, it should be." Yet in practice the economic standard is not used as widely as it should be, perhaps because of the relative ease with which one can use a productivity standard. That is, staffing levels based on productivity standards can be calculated in one's head or on a single sheet of paper. In contrast, economic and service standards require the use of queuing models or simulations.

Both the productivity standard and the service standard are consistently more costly than the economic standard (see Exhibit 4). Both the productivity standard and the service standard have a narrower range of customer-arrival rates where they match the economic standard's costs than they did in scenario one. Further, the excess costs are even higher than in scenario one. The increased costs result from the greater staffing discrepancies under scenario two.
In contrast to the second scenario, if the economic standard was lower than in scenario one (i.e., the customers were deemed less valuable), the economic standard would still be best. The other approaches would result in overstaffing.

**The implications.** There are three important implications of the two scenarios. First, though a productivity standard and a service standard can sometimes match the performance of an economic standard, they do so only in narrow ranges of business volume. That suggests that if those standards are to be applied correctly, either they must be applied only in narrow ranges of business volumes or they must change across business volumes. Identifying the relevant business volume ranges and values of the standards would be difficult, which is perhaps why their implementation does not result in the anticipated favorable outcome.

Second, productivity standards and service standards are largely inaccurate, even when they are consistent with a particular customer-waiting cost, as was shown in scenario 1. This inaccuracy leads to staffing levels that are higher or lower than ideal and therefore service-delivery costs that are higher than those resulting from use of the economic standard.

Third, regardless of whether one implements a productivity standard, a service standard, or an economic standard, one is faced with the difficult task of setting an appropriate benchmark. The problem with productivity standards and service standards is that the labor standards they implement are only surrogates for an economic standard. Since only the economic-standard approach directly addresses the economic effect of good and poor service, it yields better staffing decisions. I therefore contend that hospitality businesses should use economic
standards, despite the difficulty of determining the economic value of good and poor service.

**Developing an Economic Standard**

I've shown that the economic standard is a better tool than either productivity standards or service standards. To apply this knowledge, however, we must be able to develop the economic standard. Here are some ways to do it. I will illustrate the alternatives using a scenario where employees can serve 16 customers per hour and where 112 customers are expected to arrive in a given hour (necessitating a minimum of eight employees during that hour). Also, I'll assume that the total hourly labor cost (including benefits) is $10.00 per employee, and that the contribution value of a transaction is $5.00 (unless otherwise indicated).

**Method 1—Applying the economic standard.** As noted earlier, the simplest (but not necessarily best) way to implement an economic standard is to estimate a customer-waiting cost. For example, since the transaction contribution is lower than our hourly labor cost per employee, one might initially assume that the per-hour customer waiting cost is approximately equal to an employee's hourly labor cost ($10.00). The chart in Exhibit 5 applies that assumption to the queuing formulas shown in the box on page 28.

Using the data shown in Exhibit 5, it is possible to select a staffing level that minimizes the total service-delivery costs. In this case, the ideal number of employees to have on duty for that particular hour is nine.

One of the problems with using the customer-waiting-cost technique is that it assumes a linear relationship between waiting time and cost. For example, it assumes that 100 customers each waiting one minute has the same cost to a firm as does one customer waiting 100 minutes. The next method offers a way to overcome this shortcoming.

**Method 2—Using a revenue focus.** In switching from a cost to a revenue focus, one attempts to answer the question, how will current and future sales be affected by a wait of x minutes? To apply the simplest form of the revenue-based approach, one need only identify the waiting time at which customers are lost (i.e., after how many minutes’ waiting does a customer walk out without making a purchase?), which can be done through observation, experimentation, and experience. For example, let's say we've observed that making customers wait for less than 10 minutes has no effect on sales, yet we lose the current sale for any customer forced to wait more than 10 minutes (loss of future sales is addressed later on). Exhibit 6 shows that, in this case, nine servers is the ideal workforce to have on duty for the hour, since the net benefit is maximized at nine servers.

Now, consider the effect of a higher transaction contribution, such as $100. The table in Exhibit 7 shows that the ideal number of servers increases to 10. The reason that the tenth server is worthwhile in this example is that the extra

---

**Exhibit 5**

**Applying an economic standard**

<table>
<thead>
<tr>
<th>Number of Servers</th>
<th>Wq&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Total waiting time&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Total waiting cost&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Labor cost&lt;sup&gt;d&lt;/sup&gt;</th>
<th>Total cost&lt;sup&gt;e&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>2.382</td>
<td>4.447</td>
<td>$44.47</td>
<td>$80.00</td>
<td>$124.47</td>
</tr>
<tr>
<td>9</td>
<td>0.722</td>
<td>1.347</td>
<td>$13.47</td>
<td>$90.00</td>
<td>$103.47</td>
</tr>
<tr>
<td>10</td>
<td>0.277</td>
<td>0.517</td>
<td>$5.17</td>
<td>$100.00</td>
<td>$105.17</td>
</tr>
</tbody>
</table>

<sup>a</sup> Average time a customer spends in the queue (from the queuing-model results), in minutes.

<sup>b</sup> Total hours in the queue across all customers, equal to number of customers (112) times the average waiting time, in hours.

<sup>c</sup> Equal to the total waiting time, in hours, times the estimated hourly cost of customer waiting ($10.00 in this example).

<sup>d</sup> Equal to the number of servers times the hourly labor cost per employee.

<sup>e</sup> Equal to the total waiting cost plus the labor cost.
server's ability to reduce the number of transactions lost more than offsets the employee's hourly labor cost. In general, the higher the contribution per transaction, the higher the number of servers that are economically warranted.

That simple revenue-based approach can be further refined. Rather than trying to identify a single cut-off point, beyond which all sales are lost, it is possible to track the proportion of transactions that are lost based on waiting times within certain ranges. For example, let's say that we've observed that for waits up to three minutes no customers are lost, but for waits of three to five minutes we lose 20 percent of the customers, and we lose 60 percent of the customers for waits between five and ten minutes. As before, all sales are lost if the wait exceeds ten minutes. We can then perform an analysis of different staffing levels, as shown in Exhibit 8 (on the next page). The ideal staffing level is that level which returns the greatest benefits—in this case, 10 employees.

The calculation of "transactions lost" in Exhibit 8 merits further elaboration. The transactions-lost figures are found by multiplying the number of expected customers times the likelihood that a customer will leave without making a purchase (e.g., the wait was too long). The value of the likelihood that a customer will leave is found by summing across the different wait categories the probabilities of customers having to wait that long times the observed likelihood that the customer will be lost given that wait. So, in this example, the possible waiting periods are defined as 0 to 3 minutes, 3 to 5 minutes, 5 to 10 minutes, and more than 10 minutes. If there are eight employees on duty, the probabilities for

---

Exhibit 6

<table>
<thead>
<tr>
<th>Number of servers</th>
<th>P(W ≤ 10)</th>
<th>P(W &gt; 10)</th>
<th>Transactions made</th>
<th>Transactions lost</th>
<th>Total transaction value</th>
<th>Labor cost</th>
<th>Net benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0.9559</td>
<td>0.0441</td>
<td>107.056</td>
<td>4.944</td>
<td>$535.28</td>
<td>$80.00</td>
<td>$455.28</td>
</tr>
<tr>
<td>9</td>
<td>0.9981</td>
<td>0.0019</td>
<td>111.792</td>
<td>0.208</td>
<td>$525.96</td>
<td>$80.00</td>
<td>$445.96</td>
</tr>
<tr>
<td>10</td>
<td>0.9999</td>
<td>0.0001</td>
<td>111.992</td>
<td>0.008</td>
<td>$525.96</td>
<td>$100.00</td>
<td>$425.96</td>
</tr>
</tbody>
</table>

aProportion of customers waiting 10 minutes or less (from the queuing-model results).
bProportion of customers waiting more than 10 minutes (from the queuing-model results).
cEqual to the number of customers expected that hour (112) times the proportion whose wait does not exceed 10 minutes.
dEqual to the number of customers expected that hour (112) times the proportion whose wait exceeds 10 minutes.
eEqual to the number of transactions made times the transaction contribution (assumed to be $5.00 in this example).
fEqual to the number of servers times the hourly labor cost per employee ($10).
gEqual to the total transaction value minus the labor cost.

Exhibit 7

<table>
<thead>
<tr>
<th>Number of servers</th>
<th>P(W ≤ 10)</th>
<th>P(W &gt; 10)</th>
<th>Transactions made</th>
<th>Transactions lost</th>
<th>Total transaction value</th>
<th>Labor cost</th>
<th>Net benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0.0660</td>
<td>0.9340</td>
<td>107.056</td>
<td>4.944</td>
<td>$10,705.60</td>
<td>$80.00</td>
<td>$10,625.60</td>
</tr>
<tr>
<td>9</td>
<td>0.0019</td>
<td>0.9981</td>
<td>111.792</td>
<td>0.208</td>
<td>$11,179.18</td>
<td>$90.00</td>
<td>$11,089.18</td>
</tr>
<tr>
<td>10</td>
<td>0.0001</td>
<td>0.9999</td>
<td>111.992</td>
<td>0.008</td>
<td>$11,199.17</td>
<td>$100.00</td>
<td>$11,099.17</td>
</tr>
<tr>
<td>11</td>
<td>0.0000</td>
<td>1.0000</td>
<td>112.000</td>
<td>0.000</td>
<td>$11,200.00</td>
<td>$110.00</td>
<td>$11,090.00</td>
</tr>
</tbody>
</table>

aProportion of customers waiting 10 minutes or less (from the queuing-model results).
bProportion of customers waiting more than 10 minutes (from the queuing-model results).
cEqual to the number of customers expected that hour (112) times the proportion whose wait does not exceed 10 minutes.
dEqual to the number of customers expected that hour (112) times the proportion whose wait exceeds 10 minutes.
eEqual to the number of transactions made times the transaction contribution (assumed to be $100.00 in this example).
fEqual to the number of servers times the hourly labor cost per employee ($10).
gEqual to the total transaction value minus the labor cost.

---

Quinn, Andrews, and Parsons, pp. 75–91.
Exhibit 8
Accounting for different customer-waiting times

<table>
<thead>
<tr>
<th>Number of servers</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) $P(W \leq 3)^A$</td>
<td>0.7145</td>
<td>0.9222</td>
<td>0.9798</td>
<td>0.9951</td>
</tr>
<tr>
<td>(2) $P(3 &lt; W \leq 5)^B$</td>
<td>0.1180</td>
<td>0.0510</td>
<td>0.0161</td>
<td>0.0043</td>
</tr>
<tr>
<td>(3) $P(5 &lt; W \leq 10)^C$</td>
<td>0.1233</td>
<td>0.0249</td>
<td>0.0040</td>
<td>0.0006</td>
</tr>
<tr>
<td>(4) $P(W &gt; 10)^D$</td>
<td>0.0441</td>
<td>0.0019</td>
<td>0.0001</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Transactions lost$^E$ | 15.875 | 3.022  | 0.636  | 0.137  |
Transactions made$^F$ | 96.125 | 108.978| 111.364| 111.863|
Total transaction value$^G$ | $480.63$ | $544.89$ | $556.62$ | $559.32$ |
Labor cost$^H$ | $80.00$ | $90.00$ | $100.00$ | $110.00$ |
Net benefit$^I$ | $400.63$ | $454.89$ | $456.82$ | $449.32$ |

$^A$Proportion of customers waiting three minutes or less (from the queuing-model results).
$^B$Proportion of customers waiting between three and five minutes (from the queuing-model results).
$^C$Proportion of customers waiting between five and ten minutes (from the queuing-model results).
$^D$Proportion of customers waiting more than 10 minutes (from the queuing-model results).
$^E$Equal to the number of customers expected (112) times the likelihood of a transaction (customer) being lost; that is, $(row 1 \times 0) + (row 2 \times 0.20) + (row 3 \times 0.60) + (row 4 \times 1.00)$ = all lost transactions.
$^F$Equal to the number of customers expected (112) minus the number of transactions lost.
$^G$Equal to the number of transactions made times the transaction contribution (assumed to be $5.00 in this example).
$^H$Equal to the number of servers times the hourly labor cost per employee ($10).
$^I$Equal to the total transaction value minus the labor cost.

Future sales. Davis takes this analysis one step further by attempting to identify the long-term effects of good and poor service. The argument he uses is that poor service affects not only the current transaction, but it reduces future business. Similarly, very good service customers falling into the different waiting periods are 0.715, 0.118, 0.123, and 0.044, respectively, and the observed likelihood of losing those waiting customers is 0, 0.2, 0.6, and 1.0, respectively, for the waiting periods. This translates into an overall likelihood of 0.142 of losing the transaction, as follows: $(0.715 \times 0.0) + (0.118 \times 0.2) + (0.123 \times 0.6) + (0.044 \times 1.0) = 0.142$. Multiply that value by the 112 customers expected for that hour and the transactions-lost figure equals 15.875 (Exhibit 8).

The first thing to note from the table above is that serving customers under 0.15 minutes (about 10 seconds of wait time) will increase future business by 0.5 transactions, on average, for every customer who experiences that exceptional service. The second thing to note is that very poor service (a wait over 10 minutes) results not only in the loss of the current sale, but also the loss of another transaction. In other words, poor service results in the loss of two transactions for every customer who experiences that poor service (for example, through the customer's failing to return or through negative word-of-mouth that influences other potential patrons). Once those effects have been estimated, setting the staffing level is straightforward, as shown in Exhibit 9. The ideal staff level, then, is that number of employees that can provide the maximum benefit. In this case, 12 employees would be ideal for the hour (Exhibit 9). In general, the longer the time horizon one considers as being affected by good or poor service, the higher the staffing levels that are appropriate. Since estimating future effects can be problematic, managers are likely to underestimate their importance. Again, underestimating the long-run effects can result in a manager setting staffing levels that are lower than they should be to maximize profits.

$^7$Davis, pp. 421–434.
Other Issues

To apply economic standards, it's necessary to take into account some confounding factors.

**Forecast error.** When translating demand forecasts into employee requirements, one should take forecast error into consideration. The greater the forecast inaccuracy, the greater the staffing requirements (see Exhibit 10, on the next page). When there is a high level of forecast inaccuracy, required staffing levels can be as much as 50 percent higher, and service-delivery costs as much as 39 percent higher, than when there are perfectly accurate forecasts. The effects of inaccurate forecasts illustrates the connection between the four tasks of workforce scheduling. Because each task has only a limited ability to correct problems or deficiencies that occurred in an earlier task, one must perform each task well.

**Work spillover.** When translating the demand forecasts into labor requirements one might ask, Is it safe to do this translation separately for each planning period? In general the answer is no.

The reason is simple: service spills over from one planning period to another. Consider, for example, a quick-service restaurant using 15-minute planning periods, where the mean service duration is three minutes. If the current planning period ends at 1:00 PM, a customer who arrives at 12:50 will be served in the current period. For a customer who arrives between 12:57 and 1:00, however, a portion of the service is performed in the next period. The closer a customer arrives to the end of the current period, the greater the demand the customer places for service in the next planning period.

The effect of that spillover is to lag demand. When demand is increasing across periods, the spillover effect reduces the number of staff required in a period, as the service

---

### Exhibit 9
**Accounting for waiting effects on future business**

<table>
<thead>
<tr>
<th>Number of servers</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) P(W ≤ 0.15)</td>
<td>0.2606</td>
<td>0.6446</td>
<td>0.8033</td>
<td>0.8968</td>
<td>0.9713</td>
<td>0.9750</td>
</tr>
<tr>
<td>(2) P(0.15 &lt; W ≤ 3)</td>
<td>0.3250</td>
<td>0.2776</td>
<td>0.1765</td>
<td>0.0983</td>
<td>0.0284</td>
<td>0.0239</td>
</tr>
<tr>
<td>(3) P(3 &lt; W ≤ 5)</td>
<td>0.1180</td>
<td>0.0510</td>
<td>0.0161</td>
<td>0.0043</td>
<td>0.0003</td>
<td>0.0002</td>
</tr>
<tr>
<td>(4) P(5 &lt; W ≤ 10)</td>
<td>0.1233</td>
<td>0.0249</td>
<td>0.0040</td>
<td>0.0006</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>(5) P(10 &lt; W)</td>
<td>0.0441</td>
<td>0.0019</td>
<td>0.0001</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Transactions lost</td>
<td>20.819</td>
<td>3.231</td>
<td>0.644</td>
<td>0.137</td>
<td>0.029</td>
<td>0.006</td>
</tr>
<tr>
<td>Transactions gained</td>
<td>21.817</td>
<td>36.100</td>
<td>44.987</td>
<td>50.221</td>
<td>53.129</td>
<td>54.651</td>
</tr>
<tr>
<td>Net transactions</td>
<td>112.998</td>
<td>144.870</td>
<td>158.345</td>
<td>162.084</td>
<td>165.010</td>
<td>166.645</td>
</tr>
<tr>
<td>Total transaction value</td>
<td>$894.99</td>
<td>$634.35</td>
<td>$681.71</td>
<td>$700.42</td>
<td>$705.50</td>
<td>$703.22</td>
</tr>
<tr>
<td>Labor cost</td>
<td>$80.00</td>
<td>$90.00</td>
<td>$100.00</td>
<td>$110.00</td>
<td>$120.00</td>
<td>$130.00</td>
</tr>
<tr>
<td>Net benefit</td>
<td>$484.99</td>
<td>$634.35</td>
<td>$681.71</td>
<td>$700.42</td>
<td>$705.50</td>
<td>$703.22</td>
</tr>
</tbody>
</table>

aProportion of customers waiting 0.15 minutes or less (from the queuing-model results).
bProportion of customers waiting between 0.15 and 3 minutes (from the queuing-model results).
cProportion of customers waiting between 3 and 5 minutes (from the queuing-model results).
dProportion of customers waiting between 5 and 10 minutes (from the queuing-model results).
eProportion of customers waiting more than 10 minutes (from the queuing-model results).

| | (row 2 x 0.00) | (row 3 x 0.00) | (row 4 x 0.00) + (row 5 x 0.00) | (row 5 x 0.00) + (row 6 x 2.00) | all lost transactions (due to poor service).
| | (row 1 x 0.50) | (row 2 x 0.00) | (row 3 x 0.00) | (row 2 x 0.00) | (row 3 x 0.00) | (row 4 x 0.00) | (row 5 x 0.00) | all additional transactions (from good service).
| | (row 1 x 0.50) | (row 2 x 0.00) | (row 3 x 0.00) | (row 2 x 0.00) | (row 3 x 0.00) | (row 4 x 0.00) | (row 5 x 0.00) | all additional transactions (from good service).
| | (row 1 x 0.50) | (row 2 x 0.00) | (row 3 x 0.00) | (row 2 x 0.00) | (row 3 x 0.00) | (row 4 x 0.00) | (row 5 x 0.00) | all additional transactions (from good service).
| | (row 1 x 0.50) | (row 2 x 0.00) | (row 3 x 0.00) | (row 2 x 0.00) | (row 3 x 0.00) | (row 4 x 0.00) | (row 5 x 0.00) | all additional transactions (from good service).
| | (row 1 x 0.50) | (row 2 x 0.00) | (row 3 x 0.00) | (row 2 x 0.00) | (row 3 x 0.00) | (row 4 x 0.00) | (row 5 x 0.00) | all additional transactions (from good service).

f Equal to the number of customers expected (112) times the expected number of transactions lost per customer; that is, (row 2 x 0.00) + (row 3 x 0.00) + (row 4 x 0.00) + (row 5 x 0.00) + (row 6 x 2.00) = all lost transactions (due to poor service).

**Equal to the number of customers expected (112) times the expected number of additional transactions per customer; that is, (row 2 x 0.00) + (row 3 x 0.00) + (row 4 x 0.00) + (row 5 x 0.00) + (row 6 x 2.00) = all lost transactions (due to poor service).

**Equal to the number of transactions lost, plus the number of transactions gained.

**Equal to net number of transactions times the transaction contribution (assumed to be $5.00 in this example).

**Equal to the number of servers times the hourly labor cost per employee ($10).

**Equal to the total transaction value minus the labor cost.
Exhibit 10
Ideal number of employees under increasing levels of forecast inaccuracy

<table>
<thead>
<tr>
<th>Coefficient of variation of forecast error*</th>
<th>Number of employees</th>
<th>Service-delivery cost per hour**</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>10</td>
<td>$111.05</td>
</tr>
<tr>
<td>0.05</td>
<td>10</td>
<td>$112.81</td>
</tr>
<tr>
<td>0.10</td>
<td>11</td>
<td>$117.28</td>
</tr>
<tr>
<td>0.15</td>
<td>12</td>
<td>$125.25</td>
</tr>
<tr>
<td>0.20</td>
<td>13</td>
<td>$134.46</td>
</tr>
<tr>
<td>0.25</td>
<td>14</td>
<td>$144.30</td>
</tr>
<tr>
<td>0.30</td>
<td>15</td>
<td>$154.70</td>
</tr>
</tbody>
</table>

Note: Figures are based on the assumptions used in scenario one.


** The source of the service-delivery cost in the table above is wages plus costs for customers waiting.

spilling over from preceding periods is less than the service spilling over into the following periods. When demand is decreasing across periods, the spillover effect increases the number of staff members needed. An earlier paper presents a method of calculating the spillover effect.*

Deviations from the ideal.
Forecasts of demand are translated into employee requirements on a period-by-period basis. When the actual schedule is developed, however, it is often difficult to match the number of employees scheduled to the number of employees needed. There are a number of reasons why this is true, some of which will be covered in Part 3 of this series of labor-scheduling articles.

If the matching cannot be done exactly, then it helps to know what happens when one deviates from the ideal staff size. That is, in addition to determining the number of employees that should be scheduled, the translation process should also provide information about the cost of deviations from that ideal staff size.

Exhibit 11 provides an example of such information for an economic standard, based on the assumptions of scenario one: a labor cost of $10 an hour, including benefits; the ability of employees to serve 16 customers an hour at 100-percent capacity; and the same customer-waiting cost of $13.46 an hour used earlier, in scenario one. In the first period the ideal staff size is five employees. Overstaffing by one employee increases costs $5.27 over the ideal. That is less than the $10 increase in labor costs arising from the additional employee because of the improved customer service and hence lower costs for customers waiting.

If one has to be overstaffed by one employee in one of the periods, it is best to do so in the third period, where the net cost of overstaffing is the lowest. Similarly, if one must be understaffed by one employee in one of the periods, the second period would be the choice.

For controllable work, the economic effect of deviating from the ideal staff size need not be calculated on a period-by-period basis, as controllable work is not attributed to specific periods. Rather, the economic information can be calculated based on deviations over several periods. For example, if one needs a total of 100 hours for housekeeping on a particular day, what is the cost if 99 hours or 101 hours are scheduled? Such informa-
tion can help one make decisions about how to deploy staff across days and across jobs.

**Deploying staff across jobs.** Faced with simultaneously staffing two or more jobs, managers often cross-train their employees. The rationale is that it may be possible to use employees more effectively by assigning them to different jobs at different times of the day.

Economic standards offer the only means of directly determining where to allocate staff across jobs.

Consider the case where a manager must staff two jobs using a pool of cross-trained employees but has insufficient employees to staff both jobs at their ideal levels. Neither a productivity standard nor a service standard can offer guidance about which job to short-staff. An economic standard, however, by identifying the economic effect of short-staffing for each job, allows one to make that determination.

**Feedback.** The four tasks of workforce scheduling are closely linked. In particular, the task of translating the demand forecasts into employee requirements is closely linked to the task of developing the workforce schedule and, in fact, should operate with feedback from the scheduling task. A hotel chain’s telephone reservation center that I visited can illustrate that point.

In the reservation center, which operates around the clock, the manager used the average productivity of employees in translating demand forecasts into employee requirements. Using average employee productivity is valid if the mix of employees working in all periods is similar. However, that was not the case at this center. The manager developed a generic work schedule and then allowed the employees, in order of their rank, to choose when they wanted to work.

<table>
<thead>
<tr>
<th>Period</th>
<th>Customer-arrival rate</th>
<th>Ideal staff size</th>
<th>Number of employees above or below the &quot;ideal staff size*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50.8</td>
<td>5</td>
<td>NA*</td>
</tr>
<tr>
<td>2</td>
<td>74.4</td>
<td>7</td>
<td>$122.04</td>
</tr>
<tr>
<td>3</td>
<td>118.2</td>
<td>10</td>
<td>$93.62</td>
</tr>
</tbody>
</table>

*It would not be reasonable to have only three employees on duty, because 3 employees x 16 customers per hour = 48 customers per hour maximum (which is less than the minimum customer-arrival rate shown).

The more senior and more productive employees chose first, and most of them selected day shifts, leaving the less productive employees to work the night shifts. Since the translation process did not make use of information about the varying productivity of the employees working at different times of the day, staffing levels were not set appropriately. The center could easily have corrected the situation by applying different productivity levels at different times of the operating day.

**Absenteeism.** A manager should consider absenteeism when developing the employee requirements. It is important to track absenteeism by time of the day, day of the week, and specific job; it does not occur consistently. (In the long run, of course, one should address the causes of systematic absenteeism.)

The effect of absenteeism is similar to that of forecast inaccuracy. The higher the absenteeism, the higher the staffing levels should be. Ideal staffing levels should be adjusted upward to reflect the likelihood of absenteeism. I will discuss absenteeism more thoroughly in the fourth paper in this series.

**The Best Approach**

This paper covered three approaches to translating demand forecasts into employee requirements. Of the three approaches, economic standards are the most appropriate for both controllable and uncontrollable work. For uncontrollable work, the translation process identifies the ideal number of employees to have working in each of the planning periods in the scheduling horizon. For controllable work, the translation process identifies the labor-hour requirement for the work and the window in which the work can be performed. For both kinds of work an economic standard can provide information about the economic consequences of deviating from the ideal staff size.

The paper described how forecast inaccuracy and absenteeism both increase the number of employees necessary. There is a clear incentive, then, to develop accurate forecasts and to control absenteeism.

The outputs of the translation process become inputs to the third step—the development of the labor schedule. The next paper in this series addresses that task. **CO**