2003

A Market Utility-Based Model for Capacity Scheduling in Mass Services

John C. Goodale
University of Oregon

Rohit Verma
Cornell University, rv54@cornell.edu

Madeleine E. Pullman
Cornell University

A Market Utility-Based Model for Capacity Scheduling in Mass Services

Abstract
Only a small set of employee scheduling articles have considered an objective of profit or contribution maximization, as opposed to the traditional objective of cost (including opportunity costs) minimization. In this article, we present one such formulation that is a market utility-based model for planning and scheduling in mass services (mums), mums is a holistic approach to market-based service capacity scheduling. The mums framework provides the structure for modeling the consequences of aligning competitive priorities and service attributes with an element of the firm's service infrastructure. We developed a new linear programming formulation for the shifts-scheduling problem that uses market share information generated by customer preferences for service attributes. The shift-scheduling formulation within the framework of mums provides a business-level model that predicts the economic impact of the employee schedule. We illustrated the shift-scheduling model with empirical data, and then compared its results with models using service standard and productivity standard approaches. The result of the empirical analysis provides further justification for the development of the market-based approach. Last, we discuss implications of this methodology for future research.

Keywords
service design, service attributes, staff scheduling, service economics

A Market Utility-Based Model for Capacity Scheduling in Mass Services

JOHN C. GOODALE, University of Oregon
ROHIT VERMA, University of Utah
MADELEINE E. PULLMAN, Cornell University

Only a small set of employee scheduling articles have considered an objective of profit or contribution maximization, as opposed to the traditional objective of cost (including opportunity costs) minimization. In this article, we present one such formulation that is a market utility-based model for planning and scheduling in mass services (mums), mums is a holistic approach to market-based service capacity scheduling. The mums framework provides the structure for modeling the consequences of aligning competitive priorities and service attributes with an element of the firm's service infrastructure. We developed a new linear programming formulation for the shifts-scheduling problem that uses market share information generated by customer preferences for service attributes. The shift-scheduling formulation within the framework of mums provides a business-level model that predicts the economic impact of the employee schedule. We illustrated the shift-scheduling model with empirical data, and then compared its results with models using service standard and productivity standard approaches. The result of the empirical analysis provides further justification for the development of the market-based approach. Last, we discuss implications of this methodology for future research.

Introduction

There are many strategies for planning and scheduling service capacity in response to customer demand projections including determining optimal pricing (Mendelson and Whang 1990), shared capacity (Charalambides 1984), increasing customer participation (Larrson and Bowen 1989), scheduling work shifts with prescribed periodic employee requirements (Jacobs and Bechtold 1993), and developing flexible capacity by cross-training employees and using part-time employees (Brusco and Johns 1998).

Traditional employee workforce or staff scheduling is a common method for scheduling front-line service providers to meet periodic demand in a cost (including opportunity cost) minimization approach (Dantzig 1954; Andrews and Parsons 1989; Aykin 1996). Using this framework, many models in the current literature address the effect of forecasted customer arrivals on the scheduling of service capacity. However, a confounding and often ignored effect is the benefit generated by service levels from a system where capacity exceeds demand (Thompson 1995). That is, overstaffing may cause customers to receive better than prescribed or expected service levels, and this may lead to increased repeat business from satisfied customers and positive effects on the firm's image. Only a few methods that we explore in the literature review explicitly consider the effect of the level of service capacity on future customer demand.

Relatively low degrees of interaction and customization with a relatively high degree of labor intensity characterize a large number of service operations. These operations are classified as mass services according to the service process matrix (Schmenner 1986). However, when interaction and customization are low in some labor intensive services (compared with professional services, e.g.), customer contact can still be high and reflective of a pure service environment (Chase 1978). Many processes fit the mass service or pure service categories including inbound telephone call centers, bank branches, and fast-food restaurants. Researchers have modeled these types of processes as flat operations [as described by Buzacott (1996)], and often employ queuing analysis to determine waiting-line characteristics and service levels. This is the relatively high-volume operational environment in which our market utility-based methods operate. Hereafter, we refer to this environment as simply *mass services.*

We narrow the set of service operations in this study to mass services systems in order to start with the tightest set of assumptions. We anticipate that our model can be extended to consider other service configurations; however, that will be outside the scope of this article. For our purpose, the necessary conditions for a mass service are as follows:

1. High labor intensity. High labor costs relative to capital costs indicate that scheduling and managing employees have a large impact on firm profitability.
2. Each customer interacts/transacts with a front-line service provider and then departs. Thus, the service operation can be modeled as a standard, single-phase queue.
3. The service is completely specified with multiple attributes. Customers know all attributes so that customer preferences and market share can be clearly identified.

4. Service attribute levels are set by managers' decisions. This condition ties market share (and thus, customer arrivals to the service) to managers' operations decisions.
5. Managerial objective is to maximize profit/operations contribution.

The waiting-line environments described by the necessary conditions are systems that can be modeled with multiserver Markovian queuing systems (M/M/S) and more specifically, Erlang's delay formula [see, e.g., Segal (1974); Andrews and Parsons (1989); Kolesar and Green (1998)|. Numerous articles in the literature have relied on queuing models to provide expected performance measures of mass service and pure service systems [see, e.g., Gaballa and Pearce (1979); Holloran and Byrn (1986); Davis (1991)]. Hence, expected waiting time is an important operations attribute that customers use to make decisions about whether or not to patronize a particular service firm in this environment.

The purpose of this article is to (1) present a holistic approach to planning and scheduling in market-based, mass services that supports the emerging service strategy literature where competitive capabilities are systematically linked to service operations strategy (Roth and van der Velde 1991; Soteriou and Zenious 1999; Menor, Roth, and Mason 2001); (2) present a market-based model that generates arrival rates (consumer demand) based on integrating staffing levels and consumer's expected waiting-time distributions; and (3) develop a new linear programming (i.p) formulation for the shift-scheduling problem that uses these arrival rates and illustrate the model with empirical data. The shift-scheduling model within the framework of the market utility-based approach provides a business-level model that predicts the economic impact from employee scheduling based on configurations of operations attributes.

Literature Review

In this section, we provide a review of relevant literature in three parts. We begin in Section 2.1 by examining papers that addressed service capacity planning in service operations. In particular, we focus on service scheduling models that specified an economic value other than strictly labor cost. We review research that addressed service operation attributes and infrastructure in Section 2.2. In Section 2.3, we explore work regarding the specific operations attribute of customer waiting, which has been a key service-level parameter of detailed capacity scheduling.

Service Capacity Scheduling

The interaction of service capacity and the market has received significant attention in the literature. Kalai, Kamien, and Rubinovitch (1992) examined effects on market share of two competing

servers when customers choose servers with faster service speeds. They showed how market share decreased as waiting time increased. Li and Lee (1994) considered a larger set of service attributes including price, quality, and speed of delivery. They found that firms with a higher processing rate enjoyed a larger market share.

Stidham's (1992) model linked service time to capacity and price in a single server queuing system with a design variable of arrival rate as a function of price and service rate. Duenyas and Hopp (1995) concluded that optimal customer service policies led the firm to increase customer expectations of waiting (via quoted lead times), thus decreasing arrivals, as queue length increases. Karmarkar and Pitbladdo (1995) presented models of buyer arrivals where buyer and supplier costs were modeled as functions of the time customer and service providers expended. Ittig (1994) provided a practical accounting for waiting time when considering service capacity that shed light on the misconceptions of pursuing an objective of high labor utilization and lean staffing to minimize operating costs. He showed in a supermarket example that decreasing labor utilization from 97% to 78% increased contribution by 18%.

Regarding service capacity in scheduling, Thompson (1998) observed that there are three basic approaches with which to determine required or desired periodic employee levels. The approaches are productivity standards, service standards, and economic standards. Productivity standards are calculated by dividing expected arrival rates by the productivity standard. However, this approach is based on many simplifying assumptions. The productivity standard approach usually assumes a constant value of labor shortage or surplus. Therefore, the value of incremental service capacity is the same when the system is serving relatively few customers as it is when there are many customers. Conventional wisdom suggests that for mass services with high customer contact, this approach is low in sophistication at the expense of realism and accuracy. The productivity standard approach is more likely to be used in quasi-manufacturing environments (Chase 1978) where the time the customer spends in contact with the server is relatively small when compared with the service time, and where waiting time in the queue is not a primary factor of customer satisfaction [e.g., check processing at large banks in Mabert (1979) and Krajewski and Ritzman (1980)].

The service standard approach (SSA), or determining required or desired periodic employee levels using a service level, is a multivariable concept that is commonly expressed in service scheduling problems as the percent of customers serviced in a specified amount of time (Segal 1974; Brusco et al. 1995; Kolesar and Green 1998). For example, Andrews and Parsons (1989) reported and used a rule of 85% of customers served within 20 seconds for scheduling customer service representatives at L.L.

Bean's call center. The primary limitation of the SSA is in determining the minimum acceptable or desired service level. Justification for the parameters in SSA is difficult because the multitude of variables and associated effects on the value of the service, i.e., the effects of abandonment, retrials, waiting (long or short), idle labor, and others. Toward a model combining service standards and scheduling costs, Thompson (1997) scheduled service capacity with knowledge of the Pareto frontier between service level and labor cost.

The third approach is economic standards. The economic standard approach (ESA) requires that monetary values be attached to key service operations parameters, variables, and outcome measures. Many researchers have estimated various costs of labor shortages and surpluses. These penalty costs were incorporated in i.p formulations of service capacity problems. Baker's (1976) observation that incremental labor costs depend on the labor already scheduled suggested that the value of service was not the summation of linear penalty costs of labor shortage and surplus, because the penalty costs were most likely nonlinear. Keith (1979), and, more recently, Thompson (1995) and Easton and Rossin (1996) incorporated this nonlinearity into employee scheduling models. Some of the limitations of the **SSA** also are present in the ESA. In addition, a key limitation of the ESA is the difficulty of determining the economic consequence of the various effects and accounting for them. A summary of previously proposed schemes for specifying economic value for service scheduling solutions is presented in Table 1.

TABLE 1

A Summary of Previously Proposed Schemes for Valuing Service-Scheduling Solutions (in Chronological Order)

Author	Comments
Baker (1976)	Observed that incremental labor costs depend on labor already scheduled. Proposed a deterministic goal program to account for linear penalty costs of labor slack and surplus.
Keith (1979)	Addressed the issue of incremental labor costs depending on labor already scheduled. Proposed a deterministic goal program to account for the linear penalty costs of bounded and unbounded labor slack and surplus.
Mabert (1979)	Scheduled employees in a quasi-manufacturing environment with a model that minimized wage costs and opportunity costs of unprocessed items at the end of the prescribed period.
Grassman (1988)	Proposed a profit-maximizing model from sales revenue generated by customers served less costs of customer waiting and scheduled servers.
Andrews & Parsons (1989, 1993)	Calculated the expected lost net profit from telephone orders with the expected number of calls per period, the expected percent of calls turning into orders, the average value of a lost order, the expected number of calls abandoned at the prescribed service level, and the expected percent of abandoned calls that come back as retrials.
Davis (1991)	Proposed total cost function that included the cost of waiting and the cost of service. The cost of waiting considered the levels of satisfaction customers experience with their wait in the queue.
Ittig (1994)	Maximized profit in determining the number of clerks in a retail queuing environment. Balanced the cost of additional servers against the revenues and profits arising from greater demand. Demand effects were modeled as linear, exponential, and inverse functions of expected waiting time in the queue. Expected waiting time was modeled using *M/M/s* queuing systems.
Thompson (1995)	Considered the impact of satisfaction with customer waiting time in the queue on net present value (NPV) of future customer transactions. Customer satisfaction levels with waiting in the queue were generated by methods adapted from Davis (1991). Employees were scheduled to maximize NPV.
Easton & Rossin (1996)	Accounted for shortage costs as opportunity costs of abandoned calls by the methods of Andrews and Parsons (1989). Model accommodates nonlinear penalty costs. Scheduled service capacity with a stochastic goal program that expressed each sufficient staff constraint as a distribution of employee requirements.
Goodale and Tunc (1997)	Adapted Thompson's (1995) model to schedule a workforce that exhibited dynamic service rates.

Service Design Attributes and Operations Infrastructure

For a service firm, managers make decisions concerning which attributes and what levels of attributes arc to be designed into its service operations based on market data, along with knowledge and intuition associated with the market. Managers must focus on certain operations attributes that are reflections of critical success factors that serve as the key link between operations and marketing in service organizations (Roth and van der Velde 1991). Roth and van der Velde (1991) derived the Customer/Account Base **(cab)** matrix to classify certain critical success factors in order to link systematically the competitive priorities to marketing strategy. They show this link in their empirical

investigation of the retail banking industry and categorize competitive priorities as customer/account winners and account qualifiers. Other important work in the emerging service strategy literature sheds light on how firms might link competitive capabilities to service strategy. Menor, Roth, and Mason (2001) identified strategic service groups whose operations strategies match the needs of target markets with respect to service concept (including key design attributes), resource competencies, strategic choices, and performance. Furthermore, Soteriou and Zenios' (1999) framework combines strategic benchmarking with simultaneous analysis of operations design, service quality, and profitability. They present a model showing how operations design decisions impact profitability via service quality. So designing and/or identifying appropriate service attributes to which managers apply market-based decision making is key in today's contemporary service strategy approaches. Chase (1978, 1981) observed that major design considerations for service operations are facility location, facility layout, product design, process design, scheduling, production planning, worker skills, quality control, time standards, wage payment, capacity planning, and forecasting. Chase drew distinctions between firms based on the degree of customer contact for each design consideration. Chase and Tansik (1983) presented a normative approach that considered the match between level of contact and service structure. However, in light of developments in the service process matrix (Schmenner 1986) and service structure literature (Shostack 1987), Chase's (1978, 1981) design considerations appear to reflect physical elements of the service structure and infrastructure (Roth and Jackson 1995) and not necessarily service attributes promoted to the customer as customer/account winners. Service attributes on which customers focus, e.g., might include availability, convenience, dependability, personalization, price, quality, reputation, safety, and speed (Fitzsimmons and Fitzsimmons 1998). In a similar vein, Parasura- man, Zeithaml, and Berry (1988) identified reliability, responsiveness, assurance, empathy, and tangibles as five dimensions of service that customers used to judge quality. However, Chase's (1981) construct of customer contact is clearly an element of the service structure that maps directly to a number of customer/account winners, e.g., availability, personalization, responsiveness, assurance, and empathy.

Conventional wisdom suggests that service winning, qualifying, and losing attributes of services are industry specific. However, the construct of the firm's service infrastructure (Roth and Jackson 1995), which supports specific service attributes, is generalizable across firms. We further delineate, here and in our proposed framework in the following section, that there are two basic categories of service infrastructure: hard and soft. The purpose of proposing this distinction is to more clearly identify

managerial decisions where (1) the impact on customers can be measured and (2) an accounting of the direct costs exists.

Hard infrastructure in service firms are those elements that have clearly defined conceptual bounds. For example, the employee schedule or waiting-time standards are hard infrastructure because the elements have clearly defined conceptual bounds—that which is identified has quantitative or binary measures that can accurately specify it. For example, an employee is scheduled at an exact time, or the waiting-time standard is met (or not). Accounting for staffing costs is standard accounting practice. Marketing and operations personnel have accounted for the cost of waiting (Andrews and Parsons 1989; Easton and Rossin 1996). On the other hand, soft infrastructure in the firm are those elements that are not exactly measured, and the degree of their effect is expressed in relative terms. For example, divergence (Shostack 1987) is soft because the element usually is expressed as "more (or less) divergent," and attempts to quantify divergence are attaching proxies to the relative position. Accounting for the direct costs of divergence is difficult, at best. Furthermore, soft infrastructure like the level of customer contact (Chase 1978, 1981) is indirectly a function of the manager's decisions about hard infrastructure, such as scheduling of service capacity. To maintain the greatest clarity regarding the environment that we study, we focus solely on decisions regarding hard infrastructure, where the economic impact of managerial decisions can be measured.

We view hard infrastructure as further distinction between the infrastructure choices in recent cross-functional work in service operations and marketing (Roth and van der Velde 1991; Roth and Jackson 1995). Similarly, we view it as further distinction between the managerial elements in Fitzsimmons and Fitzsimmons (1998). Given this hard/soft categorization scheme, a summary of previously proposed schemes for classifying services using hard infrastructure variables is provided in Table 2. The review of the articles in Table 2 reflects and underlines the applicability and strategic importance of a framework that values managerial decisions regarding hard infrastructure and integrates the decisions with the economics of the firm.

The Market Utility-based Model

This section presents the market utility-based model for service capacity planning and scheduling **(mums).** A service firm's market acuity (combined competitive strength of marketing and sales) is significant in explaining generic operations capabilities (Roth and Jackson 1995). Thus, a firm's level of understanding of what their customers prefer will help explain operations attributes that reflect their operations capabilities. It follows that this also applies to the level of operations attributes, which

will reflect details of their operations capabilities. In short, customer preferences of a firm's service attributes should influence managerial decisions regarding operations capabilities and attributes. A number of the articles in Table 1 show the use of market-based information [e.g., Grassman (1988); Andrews and Parsons (1989); Davis (1991)]. We develop the framework of **mums** in this section and then focus on a detailed service capacity scheduling application in Section 4. Figure I provides a graphical depiction of **mums.** Sections 3.1-3.3 specify construct definitions, relationships, and boundaries.

Supply Component

Three main constructs make up the supply component of **mums.** The managerial decisions construct addresses strategic, intermediate level (tactical), and operational decisions. We propose that the types of decisions accounted for in **mums** are a function of the hard infrastructure with associated costs that make up operating profit (Horngren and Sundem 1990; e.g., employee costs at the operational level) and the existence of service design attributes that have an impact on market share (e.g., the service level policy at the tactical level) that is determined in the demand component.

Managerial decisions in Figure 1 affect market share when changes tire reflected as service attributes to the market of potential customers. For example, staffing levels and employee schedules are key in determining the service infrastructure and the expected level of waiting time that customers experience. These staffing decisions are made in the context of the firm's competitive priorities (Roth and van der Velde 1991), which are developed with information from marketing regarding the preference structure for service attributes, and how much customers value the waiting time attribute. That is, Figure 1 has an arrow pointing from marketing toward the managerial decisions box indicating that this information is collected and synthesized in order to make service capacity planning decisions. Conversely, service attributes borne out of managerial decisions regarding competitive priorities are promoted to the market via marketing (arrow pointing from managerial decisions to marketing, and from marketing to the demand component). For example, the expected level of waiting can be advertised to or observed (as a waiting line) by the market via marketing.

Managerial decisions also specify expenditures on service structure and hard infrastructure (arrow pointing toward operations box). For example, the employee schedule is part of the specification of the front-line structure and hard infrastructure for servicing customers. The arrow pointing from operations to the economic component reflects costs expended on hard service infrastructure (e.g., corresponding labor costs); however, the decisions that specify the structure and infrastructure use information from operations (represented by the arrow to managerial decisions) regarding existing

structure and infrastructure strengths and weaknesses. In summary, **mums** provides a framework for the staff scheduling problem. That is, employees can be assigned to shifts (supply component), where labor costs are accounted for in the economic component and the demand component provides customer arrival information.

TABLE 2

A Summary of Previously Proposed Schemes for Classifying Service Operations Using Hard Infrastructure (in Chronological Order)

Author	Proposed Schemes and Comments
Chase (1978, 1981)	Proposed 12 major design considerations previously listed in Section 2.2. Service operations classified in each design area as a function of degree of customer contact (pure services, mixed services, and quasi-manufacturing services). Addressed hard infrastructure variables of service product design, process design, scheduling, production planning, worker skills, quality control, time standards, wage payment, capacity planning, and forecasting.
Thomas (1978)	Described service firms as falling in two basic categories. The two primary categories were equipment-based and people-based services. Subcategories for equipment-based type included infrastructure variables for automation, unskilled operators, and skilled operators. Subcategories for people-based type included the infrastructure variables of unskilled labor, skilled labor, and professionals.
Chase & Tansik (1983)	Described a normative model based on Chase's (1978) concept of customer contact (soft structure) that classified firms along the contact dimension. Developed 13 propositions that convey critical distinctions between high and low contact services. Hard infrastructure variables included setting capacity, production planning system, control system, and communication system.
Lovelock (1983)	Provided a summary of previously proposed schemes for classifying services from a soft structure perspective. Proposed five schemes for classifying services. One scheme was based on the hard infrastructure variable: method of service delivery.
Mills & Turk (1986)	Proposed that the customer-firm interface (soft structure) mediates the relationship between the hard infrastructure variables: task activities and information processing. Results indicate that the customer-firm interface mediates the relationship between information equivocality and task uncertainty in retail services. This particular result generally supports propositions from Chase and Tansik (1983) for high-contact retail services.
Huete & Roth (1988)	Validated a delivery system design matrix that was based on the customer contact model (Chase 1981) and the product-process design matrix (Hayes and Wheelwright 1979). This typology addressed structural elements such as the industrialization of the delivery (the extent of substituting technology and systems for people) and span (the number of distinct delivery channels). Addressed the hard infrastructure variable of delivery system design.
Roth & van der Velde (1991)	Empirically linked the competitive priorities of retail banks with operations strategy contents of structure, infrastructure, and integration choices. Addressed the hard infrastructure variables of information systems, human resources, performance management, and the quality management program.
Kellogg and Chase (1995)	Developed a measurement model for customer contact (soft structure). Classified service operations as high, medium, or low in contact. Hard infrastructure variables included communication time and reversibility of service.

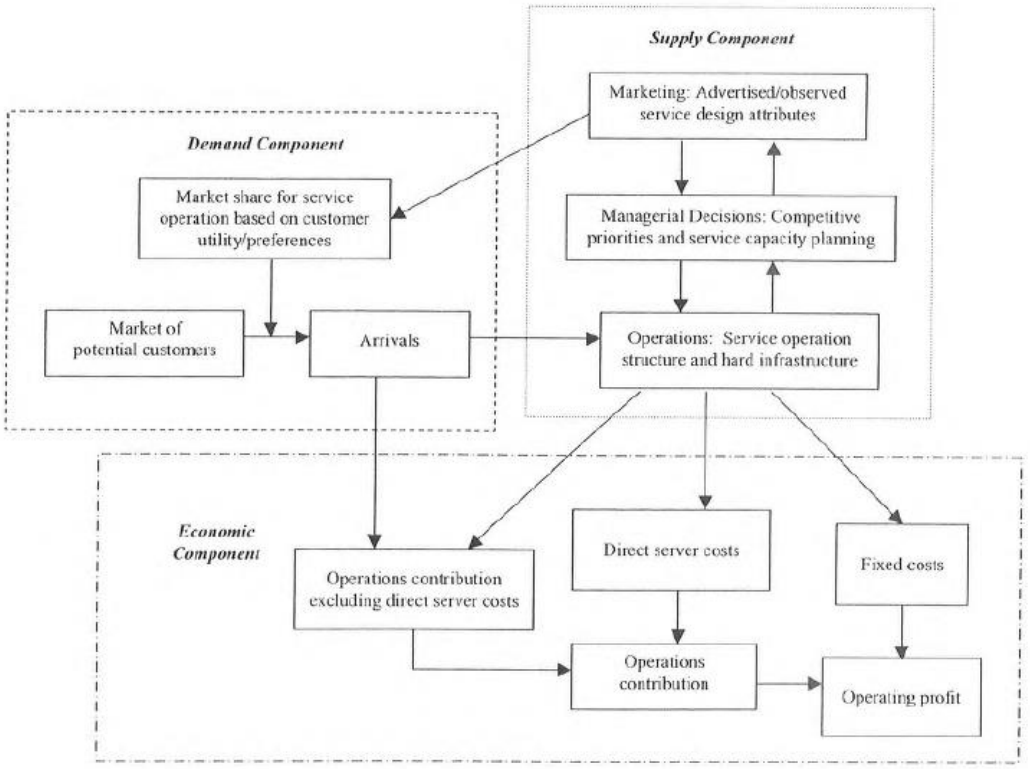

FIGURE 1. Illustration of the MUMS.

Demand Component

The Demand component is made up of three constructs. First, consumers interpret the determinant attributes of the service. The market of potential customers for a service firm is all of the customers that have that firm in their choice set. The actual customer arrival rate (arrivals) for the facility is based on market share, which is a function of the set of determinant attributes decided on by the manager of the service operation. A key point concerning the relationship between the supply and demand components of **mums** is that arrivals to the service operation will impact the resulting level of certain determinant attributes (e.g., expected waiting time in line). This is the basis for a very important fundamental insight, which is well known by managers and researchers (Easton and Pullman 2001), but warrants this special attention because it is traditionally ignored when scheduling service capacity, i.e.,

FUNDAMENTAL INSIGHT: The relationship between the supply and demand components of the service operation is dynamic. Changing one determinant attribute to affect market share (e.g., increasing variety) will impact levels of other determinant attributes (e.g., customer waiting in the queue or service lime) that are also used by consumers in making buying decisions.

There are two important ramifications of this fundamental insight on market-based scheduling of service capacity. First, a schedule is generated based on projected periodic employee requirements (usually hourly or 30-minute planning periods) that reflects the levels of determinant attributes (e.g., expected waiting in the queue and variety of services available). However, because we generally cannot treat each period as an independent epoch (Baker 1976; Thompson 1995; Easton and Rossin 1996) when scheduling employees in practice, we will likely have over- or understaffing in each individual period. If this is the case, then the consumers will experience service levels that are better or worse, respectively, and this in turn will change associated determinant attributes (e.g., customer waiting time in the queue).

The second important ramification is that changes in the levels of any determinant attribute (including waiting) will cause corresponding changes in market share and customer arrivals, which will then change determinant attributes related to waiting. Generally, waiting is considered a key determinant attribute (Gaballa and Pearce 1979; Holloran and Byrn 1986; Davis 1991). We will address both of these important ramifications and other scheduling issues in Section 4 where we schedule employees using the **mums** framework.

Economic Component

The Economic component of **mums** integrates the service operations structure and hard infrastructure from the supply component and the consumers' decisions in the demand component with a basic *contribution approach*—type framework (Horngren and Sundem 1990). Structure and hard infrastructure are as described in Section 2.2. Structure may generate variable costs but is usually associated with fixed costs. Decisions to change the layout or location of a service operation constitute strategic-level changes in structure that affect the fixed costs of the operation. On the other hand, a decision to distribute or change materials distributed to customers during a service transaction is also part of the structure and is clearly a change in variable cost, e.g., informative materials given out at an insurance firm or disposable equipment used by a dentist. Structure can map to advertised service attributes via managers' decisions.

Variable costs, such as materials (e.g., food to be sold in restaurants) can be combined with the revenue per customer (price) multiplied by market share to yield the contribution toward fixed costs from operations. Included in the variable costs are direct server costs, which are a function of the schedule. Hereafter, we will refer to variable costs in the operations contribution excluding direct server costs as "variable costs." We will refer to direct server costs as "employee costs." Operations

contribution (which includes employee costs) minus fixed costs yields operating profit. So, with this general framework provided by **mums,** we next specify a new contribution-maximizing staff scheduling model that uses the customer arrival information in order to determine economic implications of employee schedules.

Employee Scheduling in Mass Services

In this section, service capacity scheduling is explored within the framework of **mums, mums** and discrete choice analysis **(dca)** are used to identify parameters for the problem. Baker (1976) pointed out that treating each period as an individual epoch is a key limitation of setting periodic employee requirements in employee scheduling problems. Today, it is well known that unless employees can be scheduled for a single period at a lime, the solution to such a problem will likely be suboptimal. Still, employee scheduling researchers and practitioners have largely ignored this issue because of the added complexity of accounting for the dependence (Easton and Rossin 1996). Easton and Rossin's (1996) stochastic goal programming model overcomes this limitation by using distributions of expected employee requirements. Thompson (1995) overcomes this limitation by creating distributions of expected waiting time and relating the expected waiting time to NPV capitalize estimates of future customer transactions. In an attempt to contribute to this stream of literature, we specify an integer programming approach for the shift-scheduling problem within the **mums** framework that also overcomes this limitation, addresses the dynamic nature of attributes in service operations as described by the fundamental insight, and can be conveniently solved optimally with **up.**

The Shift Scheduling LP Model

In this section we present the new shift-scheduling model. Dantzig's (1954) first presented the basic lp formulation and then later Keith (1979) contributed a goal programming form. To address the limitations of setting periodic staffing constraints [found in Dantzig (1954) and Keith (1979)] described at the beginning of Section 4, we present the following integer programming formulation in the spirit of Baker (1976), Easton and Rossin (1996), and Thompson (1995):

(1)

$$\text{maximize} \qquad Z = P \sum_{t \in \mathbf{T}} \sum_{j=1}^{\mathbf{W}} \pi_{tj} y_{tj} - C \sum_{k \in \mathbf{K}} x_k$$

Subject to

(2)

$$\sum_{k \in \mathbf{K}} a_{tk} x_k - \sum_{j=1}^{\mathbf{W}} y_{tj} = 0, \qquad t \in \mathbf{T};$$

(3)

$$x_k \geq 0, \quad \text{and integer,} \quad k \in \mathbf{K}; \quad \text{and}$$

(4)

$$y_{tj} \geq 0, \quad y_{tj} \leq 1, \quad \text{and integer,} \quad t \in \mathbf{T}, \quad j \in \mathbf{W};$$

where the model parameters and decision variables are specified in Table 3. The first element of the objective function maximizes operations contribution (excluding employee costs), which is the product of the marginal operations contribution π_{tj} and arrivals for a given period (see Figure 1). Then, the total number of workers scheduled to work in a given period t is $\sum_{j=l}^{W} y_{tj}$. The second part of the objective function subtracts employee costs.

Equation (2) forces the number of workers scheduled in a given period *t* when accounting for operations contribution (excluding employee costs) $\sum_{k \in K} a_{tk} x_k$ to be equal to the number of workers scheduled in a given period *t* when accounting for employee costs $\sum_{j=l}^{W} y_{tj}$. Equations (3) and (4) provide nonnegativity and integer constraints for x_k and constrain y_{tj} to be 0-1 binary.

A key advantage and contribution of this new shift-scheduling approach is that the problem can be solved optimally with regular **lp.** However, to use **lp,** the model must satisfy four critical assumptions:

- **Assumption 1**: For a given period, expected waiting time in the queue monotonically decreases as the number of employees working *j* increases, and therefore arrival rate increases as *j* increases.
- **Assumption 2:** For a given period, incremental market share, and therefore incremental arrival rate, is monotonically decreasing as *j* increases (weakly decreasing returns to scale assumption).
- **Assumption 3**: Marginal contribution for a shift cannot equal zero.
- **Assumption 4:** Constraint coefficient matrix of a_{tk} is totally unimodular.

Assumption 1 generally is considered conventional wisdom. Davis (1991) observed that as waiting time decreased at a last-food restaurant beyond the point of dissatisfaction, satisfaction with the service increased at a decreasing rate (Assumption 2). We also would expect our empirical data to

show this property. A well-known property of queuing systems is the decreasing returns (change in expected waiting time in the queue) to scale tor increasing overall service rate (number of servers). The shift-scheduling problem requires Assumption 2, and we expect small discrepancies if the actual data do not satisfy the assumption. We checked for these small discrepancies and made minor transformations as was necessary for the illustration in Section 4.2.

TABLE 3

Model Parameters and Decision Variables

Model Parameters

- $\mathbf{T}$ = set of time periods for $t = 1, \ldots, \mathbf{T}$;
- $\mathbf{K}$ = set of work schedules for $k = 1, \ldots, \mathbf{K}$;
- $\mathbf{W}$ = set of employees available to work for $j = 1, \ldots, \mathbf{W}$;
- a_{tk} = 1 if t is a working period in schedule k, 0 otherwise; for $t = 1, \ldots, \mathbf{T}$; $k = 1, \ldots, \mathbf{K}$;
- π_{tj} = $\mathbf{T}x\mathbf{W}$ array of incremental arrival rate data; the incremental increase in arrival rate when the staff level is increased from $j - 1$ to j, for $t = 1, \ldots, \mathbf{T}$; $j = 1, \ldots, \mathbf{W}$;
- C = cost per server per work schedule;
- P = marginal operations contribution per period excluding employee costs.

Decision Variables

- x_k = number of employees working schedule k;
- y_{tj} = $\mathbf{T}x\mathbf{W}$ array of binary variables that are 1 if at least j employees are scheduled to work in period t, 0 otherwise.

The rationale for Assumptions 3 and 4 is inherent in the following theorem and proof.

Theorem 1. Given assumptions 1-4, the optimal solution to the i.p problem presented (I)-(4) without integer constraints will provide integer values for the y_{tj} variables.

Proof. Given Assumptions 1 and 2, if $\pi_{t1} \geq \pi_{t2} \geq \cdots \geq \pi_{tj}$ then lp will assign values $y_{t1} \geq y_{t2} \geq \cdots \geq y_{tj}$, where $0 \leq y_{tj} \leq 1$ and *j* = 1, 2,. . . , W. Each shift *k* reflected by a_{tk} for $t \in$ T can be evaluated at its marginal or incremental contribution 7*t,j* in each working period for *t* G T. Given the optimization of (1) and Assumption 3, a stall level variable y_{tj} will be equal to 1.0 when a related shift's incremental contribution is positive and will be equal to 0.0 when negative. Thus, an optimal solution will have no fractional values of y_{tj}.

Theorem 1 assures integer values for y_{tj}. If these values are moved to the right side of (2) and given Assumption 4, the x_k decision variables will be integer solutions because of the unimodular property [see Garfinkel and Nemhauser (1972)]. To have a total unimodular constraint matrix, there must be restrictions on the types of constraints one can place on employee schedules. The traditional form (Dantzig 1954) of the constraint matrix with 4-hour shifts (which is the form we used in this study) satisfies the unimodular property. However, if different categories of workers are to be scheduled, then a constraint cannot restrict the maximum number (percentage) of employees scheduled in each category |e.g., this was done for the tour formulation in Easton and Rossin, (1991)]. If this constraint is

included, then the unimodularity condition will be violated and integer solutions will not be guaranteed. Also, if workers are modeled with fractional values of a_{tk} [see, e.g., Li, Robinson, and Mabert (1991)] instead of binary 0-1 values, and then the total unimodularity condition may be violated and integer solutions may not be guaranteed. These additional constraints do not impact our study, because we locus on the employee scheduling problem as it has been traditionally modeled with homogeneous workers. Next, we illustrate the **mums** approach by applying it to a real-world service problem.

Optimizing Employee Schedules

A fast-food restaurant operates within an international terminal at a major airport in the United States. The name of the restaurant is disguised at the request of airport facility management—hereafter, we refer to it as The Restaurant. The Restaurant is open between 6:00 a.m. and 9:00 p.m. daily. The Restaurant is one of four food and beverage retail operations we examined in a food court located at the center of the terminal. In this subsection we applied the **mums** framework to The Restaurant's scheduling of front-line service providers.

Conjoint and **dca** have been used to model customer utility (preferences) for a service in response to experimentally designed profiles of service attributes (Louviere 1988). Recent studies indicate that market utility models developed from carefully conducted **dca** experiments could be used to effectively predict market share for various types of products and services (Louviere 1983; Louviere and Woodworth 1983; Ben-Akiva and Lerman 1991; Pullman and Moore 1999). Verma, Thompson, and Louviere (1999) provided a guideline for designing and conducting **dca** studies for services.

Discrete choice experiments involve careful design of service profiles (of specific service) and choice sets in which two or more service alternatives are offered to decision-makers and they are asked to evaluate the options and choose one (or none). Based on the experimental design, the decision-makers' choices (dependent variable) are a function of the attributes of each alternative, personal characteristics of the respondents, and unobserved effects captured by a random component (e.g., unobserved heterogeneity or omitted factors). To develop the market preference structure for The Restaurant, we performed the following steps: (1) identification of attributes, (2) specification of attribute levels, (3) experimental design, (4) presentation of alternatives to' respondents, and (5) estimation of choice model (Verma, Thompson and Louviere 1999). The choice model was used in the decision support system **(dca)** for determining customer arrival rates.

The initial stage of the data collection process was a survey of 100 travelers asking them to identify important attributes they used to choose a food vendor. We identified five service attributes

based on the methods of Griffin and Hauser (1993) and Verma, Thompson, and Louviere (1999). The selected service attributes were brand name, menu variety, wailing time before service, service time, and price. Based on managers' suggestions, we added menu language and a picture display of the food as additional service attributes, which made seven service attributes customers could use to make their purchasing decision. All seven factors had two or three levels. Fractional factorial experimental designs were used according to Hahn and Shapiro (1966) such that respondents considered 18 experimentally generated choice sets. Demographic questions also were included on the survey instrument.

Data were collected from 452 respondents randomly selected from travelers in the food court during a time period of June-October 1998. We used the data and the NTELOGIT program (IMS 1992) to estimate multinomial logit (MNL) customer choice models for all respondents and for the market as a whole. McFadden's *p2* and adjusted p2 (goodness-of-fit statistics) indicated a significant ($p < 0.05$) lit between the estimated model and observed empirical data (Ben-Akiva and Lerman 1991). The output of NTELOGIT specified the weights for the market utility function for each food vendor, and market share values were calculated for a given set of factor levels. For The Restaurant, the weight for the attribute of expected waiting in the queue was $b_{jjk} < 0$, i.e., market share $P(k\backslash A)$ decreased with an increase in level / of attribute *j,* expected waiting time in the queue. A proof that $P(k\backslash A)$ decreases as *i* increases when $b_{ijk} > 0$ is provided in Appendix A.1.

Another input into the **mums** framework was the expected within-day intertemporal variation for the entire market as shown in Table 4 for a given day, broken down into hourly planning periods. We derived this information from (1) a typical week's time-phased transaction data provided by The Restaurant's point-of-sale system and (2) The Restaurant's baseline market share projections. Next, we solved the new shift-scheduling problem.

For the purpose of this illustration, we used only part-time employees that were available for 4-hour shifts. First, a **dss** was developed using the customer choice model we developed. The **dss** considered design configurations of any combination of the seven attributes and estimated market share for each vendor. For this study, the seven design attributes for each food vendor initially were set to their current configuration. To examine the market's sensitivity to waiting in line at The Restaurant, all configurations in the **dss** were kept constant except the attribute of waiting (in the queue) before service at The Restaurant.

TABLE 4

Breakdown of Expected Customer Arrivals to the Food Court

Periodic Start Time														
6 am	7 am	8 am	9 am	10 am	11 am	12 pm	1 pm	2 pm	3 pm	4 pm	5 pm	6 pm	7 pm	8 pm
107.0	296.5	241.4	320.9	195.6	287.3	265.9	119.2	238.4	317.9	440.1	379.0	403.4	152.8	168.1

TABLE 5

Algorithm for Setting π_{tj} in the MUMS Shift-Scheduling Problem

Step

1 Start with an initial service design for The Restaurant (including, service time and expected wait in the queue W_q).

2 Begin with $j = 1$ employees scheduled for period t.

3 DSS: Use W_q as an attribute for the service operation and determine customers' aggregate utility for the design with the corresponding market share, and then calculate the consequent expected period arrival rate π_{tj}.

4 *M/M/s*: Calculate a new W_q using j servers and π_{tj} from step 3.

5 If W_q from step 3 is equal to W_q from step 4, continue to step 6; otherwise retain W_q from step 4 and go to step 3.

6 Record π_{tj}, set $j = j + 1$, and go to step 2. Continue for all possible $j \in \mathbf{W}$ and $t \in \mathbf{T}$.

Second, we developed a M/M/s model that reflected The Restaurant's front-line service providers. The M/M/s system generated the expected waiting time assuming all other parameters were constant. The M/M/s model satisfied Assumption I. In addition, there is rich literature where researchers have analyzed service systems using Markovian assumptions tor arrival and service rates [see, e.g., Segal (1974); Holloran and Byrn (1986); Andrews and Parsons (1989); Kolesar and Green (1998)]. However, the method of generating the operations contribution for the **mums** framework does not depend on any specific distribution or queuing system. The premise is that a positive relationship between arrivals and expected waiting time in the queue exists and that there is a procedure for determining the expected waiting-time distribution as a function of arrival rate and capacity.

By combining the **dss** and M/M/s models, we found periodic arrival rates π_{tj}; and captured the dynamic relationship between staffing levels and market share. The algorithm that solved for π_{tj} is presented in Table 5. A proof that a single market share equilibrium point exists is provided in Appendix A.2. The arrival rates π_{tj} that provided this equilibrium point are provided in Table 6 as a function of staffing level.

Multiplying the incremental values from Table 6 by the operations contribution per customer and subtracting the incremental employee cost provides the incremental operations contribution for a given number of employees scheduled in a particular planning period. The new shift-scheduling model

objective function [see (1)] has employee costs separate from the operations contribution for reflecting the conceptual model structure. Table 7 presents the incremental operations contribution net of employee costs, assuming a marginal operations contribution (excluding employee costs) of P = \$4/customer and a wage rate of \$6/hour.

TABLE 6

Arrival Rate as a Function of Staffing Level

Employees Scheduled	6 am	7 am	8 am	9 am	10 am	11 am	12 pm	1 pm	2 pm	3 pm	4 pm	5 pm	6 pm	7 pm	8 pm
1	20.7	24.7	24.2	24.9	23.5	24.6	24.4	21.3	24.1	24.9	25.5	25.2	25.3	22.5	22.9
2	34.1	49.5	48.0	50.0	46.0	49.3	48.7	36.8	47.9	49.9	51.5	50.8	51.1	42.3	44.0
3	36.9	72.8	68.5	73.9	61.5	72.4	70.9	40.8	68.2	73.8	77.0	75.8	76.3	51.1	55.3
4	37.3	90.9	79.8	94.1	66.7	89.4	85.4	41.5	79.0	93.7	101.7	99.2	100.4	52.9	58.0
5	37.4	99.7	83.1	106.1	68.0	97.1	90.8	41.7	82.1	105.3	124.0	117.2	120.4	53.3	58.6
6	37.4	102.4	84.1	110.2	68.3	99.4	92.3	41.7	83.0	109.2	140.5	127.0	133.1	53.4	58.8
7	37.4	103.3	84.4	111.6	68.4	100.2	92.8	41.7	83.3	110.6	148.8	130.6	138.2	53.5	58.8
8	37.4	103.6	84.4	112.0	68.4	100.4	93.0	41.7	83.4	111.0	152.0	131.9	140.0	53.5	58.8
9	37.4	103.7	84.5	112.2	68.4	100.5	93.0	41.7	83.4	111.1	153.2	132.3	140.7	53.5	58.8
10	37.4	103.7	84.5	112.3	68.4	100.5	93.0	41.7	83.4	111.2	153.7	132.5	141.0	53.5	58.8
11	37.4	103.7	84.5	112.3	68.4	100.5	93.0	41.7	83.4	111.2	153.9	132.6	141.1	53.5	58.8
12	37.4	103.7	84.5	112.3	68.4	100.5	93.0	41.7	83.4	111.2	153.9	132.6	141.1	53.5	58.8

TABLE 7

Operations Contribution per Incremental Employee

Employees Scheduled	6 am	7 am	8 am	9 am	10 am	11 am	12 pm	1 pm	2 pm	3 pm	4 pm	5 pm	6 pm	7 pm	8 pm
1	76.80	93.00[1]	90.80	94.00[1]	88.00	92.60[1]	91.60	79.20	90.40	93.80[1]	97.00[1]	95.60[1]	96.20[1]	84.00	85.60
2	47.60	93.00[1]	89.20	94.00[1]	84.00	92.60[1]	91.20	56.00	89.20	93.80[1]	97.00[1]	95.60[1]	96.20[1]	73.20	78.40
3	5.20	87.20	76.00	89.60	56.00	86.40	82.80	10.00	75.20	89.60	96.00	94.00	94.80	29.20	39.20
4	−4.40	66.40	39.20	74.80	14.80	62.00	52.00	−3.20	37.20	73.60	92.80	87.60	90.40	1.20	4.80
5	−5.60	29.20	7.20	42.00	−0.80	24.80	15.60	−5.20	6.40	40.40	83.20	66.00	74.00	−4.40	−3.60
6	−6.00	4.80	−2.00	10.40	−4.80	3.20	0.00	−6.00	−2.40	9.60	60.00	33.20	44.80	−5.60	−5.20
7	−6.00	−2.40	−4.80	−0.40	−5.60	−2.80	−4.00	−6.00	−4.80	−0.40	27.20	8.40	14.40	−5.60	−6.00
8	−6.00	−4.80	−6.00	−4.40	−6.00	−5.20	−5.20	−6.00	−5.60	−4.40	6.80	−0.80	1.20	−6.00	−6.00
9	−6.00	−5.60	−5.60	−5.20	−6.00	−5.60	−6.00	−6.00	−6.00	−5.60	−1.20	−4.40	−3.20	−6.00	−6.00
10	−6.00	−6.00	−6.00	−5.60	−6.00	−6.00	−6.00	−6.00	−6.00	−5.60	−4.00	−5.20	−4.80	−6.00	−6.00
11	−6.00	−6.00	−6.00	−6.00	−6.00	−6.00	−6.00	−6.00	−6.00	−6.00	−5.20	−5.60	−5.60	−6.00	−6.00
12	−6.00	−6.00	−6.00	−6.00	−6.00	−6.00	−6.00	−6.00	−6.00	−6.00	−6.00	−6.00	−6.00	−6.00	−6.00

[1] To satisfy the weakly decreasing returns to scale assumption for contribution per incremental employee, the value for one and two employees scheduled is their combined mean. The largest discrepancy was at 4:00 p.m. where scheduling one employee provided an operations contribution of \$96.00 and the second employee provided \$98.00.

Then, the new shift-scheduling problem with the four requisite assumptions was solved optimally for The Restaurant. For the optimal solution, the numbers of employees working each shift and the 4-hour shift patterns are shown in Table 8.

A TxW array of binary decision variables are allocated across the TxW array of operations contributions per incremental employee (Table 7) according to Table 9. The result is an optimal solution that provides an operations contribution (net of employee costs) of $4,913.20 by scheduling 21 employees. The solution procedure for the lp model is robust with respect to shift length. Any shift configuration could be used in Table 8 as long as Assumption 4 is satisfied.

TABLE 8

Optimal Schedule for Maximizing Operations Contribution

Shift Number k	Employees Working	a_{tk} Values: 1 if Period t Is a Working Period in Shift k; 0 Otherwise														
		6 am	7 am	8 am	9 am	10 am	11 am	12 pm	1 pm	2 pm	3 pm	4 pm	5 pm	6 pm	7 pm	8 pm
1	3	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
2	2	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0
3	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0
4	3	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
5	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0
6	2	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
7	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0
8	2	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0
9	1	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0
10	4	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0
11	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0
12	3	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
Total:	21															
Optimal Contribution: $4913.20																

TABLE 9

Array of **TxW** *Binary Decision Variables* y_{ij} *for Optimal Solution*

Employees Scheduled	6 am	7 am	8 am	9 am	10 am	11 am	12 pm	1 pm	2 pm	3 pm	4 pm	5 pm	6 pm	7 pm	8 pm
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
5	0	1	1	1	1	1	1	0	1	1	1	1	1	0	0
6	0	0	0	1	0	0	0	0	0	1	1	1	1	0	0
7	0	0	0	1	0	0	0	0	0	1	1	1	1	0	0
8	0	0	0	1	0	0	0	0	0	0	1	1	1	0	0
9	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

To address the issue of the significance, we examined the magnitude of the difference a manager might obtain with different scheduling objectives. We compared our optimal solution to the solution obtained with the **esa** and the **ssa.** The **esa** and **ssa** both identity a minimum level for the number of workers that must be satisfied in the optimal solution. The **esa** identified that minimum level by choosing the largest number of employees where the incremental operations contribution (Table 7) was positive. The **ssa** identified that minimum level by choosing the smallest number of employees that satisfied a statement of service level such as "initiate service for $\alpha\%$ of the customers in less than *¡3* seconds." The service level parameters α and β usually are specified by the manager based on industry standards and experience.

We solved the shift-scheduling problem with **esa** and **ssa.** For **ssa,** we varied the service level parameters over reasonable ranges of α and β for an M/M/s system. We examined *α* levels from 50% to 99%, and we varied *(3* levels from 0 to 120 seconds. Solutions were generated for all possible combinations of α and β. The solutions to the comparison methods were super-imposed on the operations contribution values in Table 7. The resulting optimal solutions for **esa** and **ssa** are presented in Table 10 along with their percent difference from the optimal solution to the new shift-scheduling model.

The results of the comparison methods reflect their performance compared with a true optimum found with the new **mums** shift-scheduling model. So, for this particular case, managers could most closely approximate the optimal solution based on the **mums** framework using **esa** (0.67% difference) or **ssa** (0.28% difference) with service level parameters α = 60% and β *= 0* seconds ("initiate service for 60%> of the customers with no wait").

In addition to the tangible schedule that is generated by the **mums** approach, the schedules for individual operations like The Restaurant yield insight into the effects of operations attributes on customer's patronizing behavior. That is, Table 10 indicates that no waiting (no lines) is highly valued in this setting. For this setting, the service levels that gave values for the operations contribution closest to the optimal values were achieved with service levels where β *= 0* seconds. Thus, **mums** can identify contribution-maximizing staffing levels that might be substantially different from levels specified by industry service standards (such as "serve 85% of the customers within 20 seconds or less").

TABLE 10

Comparison to Schedules Generated with Other Employee Scheduling Objectives

Comparison Method		Minimum Employees Scheduled	Mean[1] ($)	SD	% Difference from Optimal Solution
Economic Approach:		24	4880.40	3.85	0.67
Service Level Approach:					
β (sec.)	α (or $P\{W_q < t\}$)				
0	0.50	20	4868.84	17.14	0.90
0	0.60	21	4899.28	6.09	0.28
0	0.70	24	4873.00	2.89	0.82
0	0.80	24	4887.84	2.12	0.52
0	0.90	28	4818.40	1.12	1.93
0	0.99	37	4616.00	0.00	6.05
15	0.50	17	4690.56	32.28	4.53
15	0.60	17	4699.44	17.10	4.35
15	0.70	17	4694.80	0.00	4.45
15	0.80	17	4694.80	0.00	4.45
15	0.90	17	4715.52	45.58	4.02
15	0.99	18	4796.68	10.65	2.37
30	0.50	17	4699.20	33.11	4.36
30	0.60	17	4691.16	31.62	4.52
30	0.70	17	4694.80	0.00	4.45
30	0.80	17	4694.80	0.00	4.45
30	0.90	17	4694.80	0.00	4.45
30	0.99	17	4759.80	9.89	3.12
60	0.50	16	4616.00	0.00	6.05
60	0.60	16	4616.00	0.00	6.05
60	0.70	16	4616.00	0.00	6.05
60	0.80	17	4685.60	0.00	4.63
60	0.90	17	4685.60	0.00	4.63
60	0.99	17	4685.60	0.00	4.63
120	0.50	16	4615.40	34.84	6.06
120	0.60	16	4641.80	51.48	5.52
120	0.70	16	4625.36	50.70	5.86
120	0.80	16	4602.24	39.97	6.33
120	0.90	16	4621.80	43.62	5.93
120	0.99	17	4723.96	21.81	3.85

[1] Ten solutions generated with different initial corner point solutions ($n = 10$); t-tests found significant differences for all means from optimal solution at $p < 0.001$.

Discussion

Implications for Employee Scheduling in Service Systems

Only a small set of employee scheduling articles have considered an objective of profit or contribution maximization, as opposed to the traditional objective of cost (including opportunity costs) minimization. Thompson's (1995) net present value model of employee scheduling used Davis' (1991) benefit and cost structure. Goodale and Tunc (1997) used Thompson's model to schedule employees

with dynamic service rates. A substantial amount of research in employee scheduling has considered an objective of minimizing literal costs and also opportunity costs in a Dantzig-type (1954) formulation [see, e.g., Mabert (1979); Andrews and Parsons (1993)| or a goal programming formulation [see, e.g., Baker (1976); Keith (1979); Easton and Rossin (1996)]. The **mums** provides a new business-level framework for process decisions.

At a strategic level, the **mums** framework models the consequences of aligning competitive priorities and service attributes with an element of hard infrastructure in service firms. We model the inherent cyclical relationship between these purposefully ordered elements: (a) competitive priorities; (b) hard infrastructure; (c) service attributes; (d) market share, and the corresponding impact of customer arrivals on, again, (b) hard infrastructure. The Restaurant illustration showed how managerial decisions might align hard service infrastructure with competitive priorities. For example, firms may wish to establish policy in order to address the competitive priority of customer service. In most of the cases, the service-level approach to planning service capacity yielded inferior economic performance in The Restaurant illustration because it assumed an acceptable level of customer waiting in an environment where customer waiting was highly disagreeable. Thus, the **mums** framework provides insight with regard to how the relationship between competitive priorities and managerial decisions can be moderated by customer preferences.

At an operational level, this research extends the labor scheduling literature by addressing the dynamic nature of setting staffing levels. Goal programming formulations and Thompson's (1995) approach addressed the asymmetry and nonlinearity issues of employee shortage and surplus costs. However, the dynamic nature of setting staffing levels, which determine expected waiting time and the corresponding impact on customer arrivals (which affects the staffing decision) was still unresolved. Our new shift-scheduling model based on the **mums** framework accounts for this dynamic relationship by finding the single market share equilibrium point (i.e., an equilibrium arrival rate and expected customer waiting time in the queue). We presented this new model and an illustration of its use.

In a comparison of the new shift-scheduling model to other approaches, we found that parameter settings for other approaches closely approximated the true optimum value generated by the **mums** approach and the new model. However, those parameter settings that provided the highest operations contribution may not be what has been considered the industry standard. Additionally, we feel that other shift-scheduling approaches would still benefit from using the new **mums** framework for establishing optimal staffing levels.

Future Research

The **mums** approach to scheduling service operations will aid in examining a number of important theoretical issues. We suggest three such issues that arise from our research but have yet to be addressed in the service scheduling literature.

First, the **mums** approach is general and applicable in a variety of industries and/or market segments where employees are scheduled to work in order to meet uncertain demand. In particular, mass services are strong candidates for applying this approach because of the importance placed on expected waiting time as a service operations attribute. Thus, our first observation is

Observation 1. Based on market behavior, there exists unique optimal service standard (level) policies for different types of mass services.

Table 10 shows that that the optimal schedule from Table 8 most closely matched the optimal schedules generated with a customer service-level policy that had β = 0 (no waiting). Accepting lower levels of customer service decreased the operations contribution from the optimal level. So, the data indicated that customers valued minimal expected waiting limes significantly more than what we expected given the customer service-level policies reported in the literature and used in industry. Therefore, our second observation is the following.

Observation 2. For certain mass services, the utility for minimal expected waiting time in line increases radically, indicating that there exists a “no waiting" effect, i.e., the weight (in the utility function) allocated to expected waiting time in line should increase at an increasing rate as expected waiting time in line approaches zero.

Last, a full-scale experiment that examines the proposed superiority of **mums** should be undertaken in the tour (weekly schedules) environment. This is the topic for our third observation,

Observation 3. In general service classifications where operations can he modeled as multiserver queuing systems, tour schedules generated with the mums approach to staff-scheduling will yield significant economic advantages over the productivity standard, service standard, and traditional esa.

In addressing Observation 3, one needs to address the stationary independent period by period **(sipp)** modeling issue (Green, Kolesar, and Soares 2003) when generating expected waiting time distributions. This would allow a more accurate comparison of the approaches. Without accounting for lags between peaks of demand and system congestion, staffing levels may be inaccurate and result in excessive customer waiting after the arrival peaks. In summary, we feel that these observations offer a

natural extension of the staff-scheduling literature using the new framework of the **mums** approach to capacity scheduling in mass services.

Concluding Remarks

As service managers look for new methods for designing and managing service operations, increasingly we should expect to see market-based information incorporated into decision-making tools. Scheduling is no exception. Treating operations as a competitive or strategic priority precludes a cost-center orientation. Toward this end, new methods must embrace tools from other functional areas, e.g., **dca** from marketing. Incorporating these tools into service planning and control systems offers very exciting challenges.

Appendix

A.1. Proof that $P(k|A)$ Decreases as i Increases when $b_{ijk} \leq 0$

The probability of choosing brand k from choice set A is given as

(A1)

$$P(k|A) = P_k = \frac{e^{U_k}}{\sum_{s \in A} e^{U_s}}.$$

To isolate the effect of changing a particular attribute level in brand *k,* we can separate the brand *k* elements in the denominator and express the ratio as follows:

(A2)

$$P_k = \frac{e^{U_k}}{\left[\sum_{s \in A, s \neq k} e^{U_s}\right] + e^{U_k}}.$$

To simplify the denominator, let the constant C_1 be expressed as follows:

(A3)

$$C_1 = \sum_{s \in A, s \neq k} e^{U_s}.$$

Thus,

(A4)

$$P_k = \frac{e^{U_k}}{C_1 + e^{U_k}}.$$

We want to isolate the effect of changes of one attribute level in brand *k* on the market share for brand *k.* Given that the utility U_k is given as

(A5)

$$U_k = \sum_{j=1}^{n} b_{ijk} X_{ijk} + \epsilon_k,$$

where *i* ∈ m levels of attribute *j,* we can express the constant C_2 as a sum of the partial utility for all *n* attributes except when *j* is equal to the qth attribute, i.e.,

(A6)

$$C_2 = \sum_{j \in n, j \neq q} b_{ijk} X_{ijk} + \epsilon_k,$$

Where *i*∈*m.* Thus,

(A7)

$$U_k = C_2 + b_{iqk} X_{iqk},$$

where *i* ∈ m. Substituting (A7) into the exponents of (A4) creates the following expression that isolates the effect of attribute *q* in P_k*:*

(A8)

$$P_k = \frac{e^{C_2 + b_{iqk} X_{iqk}}}{C_1 + e^{C_2 + b_{iqk} X_{iqk}}},$$

where *i* ∈ m. So, it is easily seen that P_k decreases as the level of attribute *q* increases when $b_{iqk} < 0$ and $C_1 > 0$. If attribute *q* is expected waiting time in the queue, then the result is that market share (P_k) decreases with an increase in customer waiting time (increase in level *i* of attribute *q).*

A.2. Proof that a Single Market Share Equilibrium Point Exists

One can make the following statements of M/M/S queuing systems

$$\lim_{\lambda \to 0} W_q = 0 \qquad \text{and} \qquad \lim_{\lambda \to s\mu} W_q = +\infty,$$

where $s\mu$ is the system completion rate for s servers, each with service rate μ. In addition, the state $\lambda \geq s\mu$, is infeasible for M/M/s queuing systems. Therefore, any decreasing function of A must intersect the expected waiting time function at a point (λ^*, W_q*) in the feasible state $\lambda < s\mu$.* The proof in Section A. 1 shows that P_k is a decreasing function of W_q when $b_{ijk} < 0$. Specifying arrivals to be $\lambda = P_k M$ for market share P_k and a total market of M customers, and then it follows that there must be a single market share equilibrium point *($P_k M^*$, W_q*).*

References

Andrews, B. H. and H. L. Parsons (1989), "L. L. Bean Chooses a Telephone Agent Scheduling System," *Interfaces*, 19. 6, 1-9.

——— AND ——— (1993), "Establishing Telephone-Agent Staffing Levels Through Economic Optimization," *Interfaces,* 23, 2, 14-20.

Aykin, T. (1996), "Optimal Shift Scheduling with Multiple Break Windows," *Management Science,* 42, 4, 391-602.

Baker, K. R. (1976), "Workforce Allocation in Cyclical Scheduling Problems: A Survey," *Operational Research Quarterly,* 27, 1, 133-167.

Ben-Akiva, M, AND S. R. Lehman (1991), *Discrete Choice Analysis,* MIT Press, Cambridge, MA.

Brusco, M. J. and T. R. Johns (1998), "Staffing a Multiskilled Workforce with Varying Levels of Productivity: An Analysis of Cross-Training Policies," *Decision Sciences,* 29, 2, 499-515.

———, L. W. Jacobs, R. J. Bonoiorno, D. V. Lyons, and B. Tang (1995), "Improving Personnel Scheduling at Airline Stations," *Operations Research,* 43, 5, 741-751.

Buzacott, J. A. (1996), "Commonalities in Reengineered Business Processes: Models and Issues," *Management Science,* 42, 5, 768-782.

Charailambides, L. C. (1984). "Shared Capacity Resource Reallocation in a Decentralized Service System," *Journal of Operations Management,* 5, 1, 57-74.

Chase, R. B. (1978), "Where Does the Customer Fit in a Service Operation?" *Harvard Business Review,* 56, 6, 137-142.

———— (1981), “The Customer Contact Approach to Services: Theoretical Bases and Practical Expansions,"

Operations Research, 29. 4, 698-706.

———and D. A. Tansik (1983), "The Customer Contact Model for Organization Design,” *Management Science.* 29, 9, 1037-1050.

Dantzig, G. B. (1954), "A Comment on Edie's ‘Traffic Delays at Toll Booths,' ” *Operations Research,* 2, 3, 339-341.

Davis, M. M. (1991), “How Long Should a Customer Wait for Service,” *Decision Sciences,* 22, 2, 421-434.

Duhnyas, I. and W. J. Hopp (1995), “Quoting Customer Lead Times,” *Management Science,* 41, I, 43-57.

Easton, E. F. and M. E. Pullman (2001), “Optimizing Service Attributes: The Seller’s Utility Problem,” *Decision Sciences,* 32, 2, 251-275.

———— and D. F. Ross IN (1991), “Sufficient Working Subsets for the Tour Scheduling Problem,” *Management*

Science, 37, 1 1, 1441-1451. '

——— AND ————(1996), “A Stochastic Goal Program for Employee Scheduling,” *Decision Sciences,* 27, 3, 541-568.

Fitzsimmons, .1. A. and M. J. Fitzsimmons (1998), *Service Management: Operations, Strategy, and Information Technology,* 2nd Ed., Irwin McGraw-Hill, New York.

Caballa, A. and W. Plargl (1979), “Telephone Sales Manpower Planning at Qantas,” *Interfaces,* 9, 3, 1-9.

Garlinkll, R. S. and G. L. Nlmiiaushr (1972), *Integer Programming,* John Wiley and Sons, Inc., New York.

Goooal.l, J. C. and E. Tunc- (1997), “Tour Scheduling with Dynamic Service Rates,” *International Journal of Service Industry Management,* 9, 3, 226-247.

Grassman, W. K. (1988), “Finding the Right Number of Servers in Real-World Queuing Systems,” *Interfaces,* 18, 2, 94-104. '

Green, L. V., P. J. Kolesar, and .1. Soares (2003), “An Improved Heuristic for Staffing Telephone Call Centers with Limited Operating Hours,” *Production and Operations Management,* 12, 1, 1-16.

Griffin, A. and J. R. Hauser (1993), “The Voice of the Customer,” *Marketing Science,* 12, 1, 1-27.

Hahn, G. J. and S. S. Shapiro (1966), *A Catalog and Computer Program for the Design and Analysis of Orthogonal Symmetric and Asymmetric Fractional Factorial experiments.* Report no. 66-C-I65, The General Electric Company, Schenectady, NY.

Holloran, T. J. and J. E. Byrn (1986), “United Airlines Station Manpower Planning System,” *Interfaces,* 16, 1, 39-50. '

Horngren, C. T. and G. L. Sundem (1990), *Introduction to Management Accounting,* 8th ltd., Prentice-Hall, Inc., Englewood Cliffs, NJ.

Huete, L. M. and A. V. Roth (1988), "The Industrialization and Span of Retail Banks' Delivery Systems," *International Journal of Operations and Production Management,* 8, 3, 46-66.

IMS (1992), *NTELOGIT, Software and User's Manual,* Intelligent Marketing Systems, Edmonton, Canada.

Intel, P. T. (1994), "Planning Service Capacity When Demand Is Sensitive to Delay," *Decision Sciences,* 25, 4, 541-559.

Jacobs, L. W. and S. E. Bechtold (1993), "Labor Utilization Effects of Labor Scheduling Flexibility Alternatives in a Tour Scheduling Environment," *Decision Sciences,* 24, I, 148-166.

Kalai, E., M. I. Kamien, and M. Rijbinovitcii (1992), "Optimal Service Speeds in a Competitive Environment," *Management Science,* 38, 8, 1 154-1 163.

Karmarkar, U. S. and R. Pitbladdo (1995), "Service Markets and Competition," *Journal of Operations Management,* 12, 3-4, 397-411.

Keith, E. G. (1979), "Operator Scheduling," *AIIF* 7 1*ransaetions*, II, 1, 37-41.

Kellogg, D. L. and R. B. Chase; (1995), "Constructing an Empirically Derived Measure for Customer Contact," *Management Science,* 41, II, 1734-1749.

Kolesar, P. J. and L. V. Green (1998), "Insights on Service System Design from a Normal Approximation to Erlang's Delay Formula," *Production and Operations Management, 1,* 3, 282-293.

Krajewski, L. J. and L. P. Ritzman (1980), "Shift Scheduling in Banking Operations: A Case Application," *Interfaces*, 10, 2, 1-6.

Larrson, R. and D. E. Bowen (1989), "Organization and Customer: Managing Design and Coordination," *Academy of Management Review,* 14, 2, 213-233.

Li, C, E. P. Robinson, and V. A. Mabert (1991), "An Evaluation of Tour Scheduling Heuristics with Differences in Employee Productivity and Cost," *Decision Sciences, 22, 4,* 700-718.

Li, L. and Y. S. Lia; (1994), "Pricing and Delivery-Time Performance in a Competitive Environment, " *Management Science,* 40, 5, 633-646.

Louvilrh, J. J. (1983), "Integrating Conjoint and Functional Measurement with Discrete Choice Theory: An Experimental Design Approach," *Advances in Consumer Research* 10, 151-156.

————(1988), *Analyzing Decision Making: Metric Conjoint Analysis,* SAGE Publications, Newbury Park, CA.

———— and G. G. Woodworth (1983), "Design and Analysis of Simulated Consumer Choice or Allocation

Experiments: An Approach Based on Aggregate Data, ” *Journal of Marketing Research,* 20, 4, 350-367.

Lovelock, C. H. (1983), “Classifying Services to Gain Strategic Marketing Insights,” *Journal of Marketing,* 47, 3, 9-20. ' ‘

Mabert, V. A. (1979), “A Case Study of Encoder Shift Scheduling Under Uncertainty,” *Management Science, 25,* 7, 623-631.

Mendelson, FI. and S. Whang (1990), “Optimal Incentive-Compatible Priority Pricing for the M/M/l Queue,” *Operations Research,* 38, 5, 870-883.