PRODUCTION AND OPERATIONS MANAGEMENT
Vol. 12, No. 2, Summer 2003
Printed in U.S.A.

APPOINTMENT POLICIES IN SERVICE OPERATIONS: A CRITICAL ANALYSIS OF THE ECONOMIC FRAMEWORK*

SUSANA V. MONDSCHEIN AND GABRIEL Y. WEINTRAUB
Department of Industrial Engineering, U. of Chile, Yale School of Management, 135 *Prospect Street, P.O. Box* 208200, *New Haven, Connecticut* 06520-8200, *Management Science and Engineering Department, Stanford University, Terman Engineering Center,* 3*rd Floor Stanford University, Stanford, California* 94305-4026

In this paper we review the literature on appointment policies, specifically in terms of the objective function commonly used and the assumptions made about the behavior of demand. First, we provide an economic framework to analyze the problem. Based on this framework we make a critical analysis of the objective functions used in the literature. We also question the validity of the assumption made throughout the literature that demand is exogenous and independent of customers' waiting times. We conclude that the objective functions used in the literature are appropriate only in the case of a central planner facing a demand that is unresponsive to waiting time. For other scenarios, such as a private server facing a demand that does react to waiting time, these objective functions are only shortcuts for the real objective functions that must be used. A more general model is then proposed that fits these scenarios well. Finally, we determine the impact of using the literature's objective functions on optimal appointment policies.
(SERVICE OPERATIONS; APPOINTMENT POLICIES; PRIVATE SERVER AND CENTRAL PLANNER)

1. Introduction

Nowadays many service companies operate in highly competitive markets and face increasingly demanding customers. Competition is waged not only in terms of price, but also through the quality of service provided. One of the aspects of quality that has become important in recent years is the speed with which services are delivered: customers do not like having to waste their time waiting to be served or attended to. As a result, long waiting times cause customers to become dissatisfied with the service received, and this undermines the competitiveness of the company concerned. Taylor (1994) and Katz, Larson, and Larson (1991), among other authors, have drawn attention to the negative impact that being forced to wait has on customers' overall satisfaction with the service they receive. The first of these papers reports survey data and the second reports personal interviews. Both find that customers' assessment of the service worsens if waiting time increases.

Related to speed of service, there are two characteristics that have to be considered: (*i*) the customer's perception of waiting time, and (*ii*) the actual waiting time itself. Clearly, a

* Received January 2001; revisions received July 2001 and April 2002; accepted May 2002.

1059-1478/03/1202/266$1.25

waiting period that can be kept short will directly produce a lower perception of waiting time. If the waiting time cannot be controlled, however, it may be possible to reduce customers' perception of it. Baker and Cameron (1996) present an integrative review of customers' perception of waiting time. They make propositions to manage the service environment in order to reduce customers' perception of waiting time and increase their overall satisfaction. For example, they recommend that the service environment should be as comfortable as possible, regarding lighting, temperature, music, and waiting room furniture. Besides, a sensation that the system of attending to customers is "socially fair" (first-in first-out) helps to reduce the perception of waiting time.

A decisive factor in controlling customer waiting time is the appointment system used by the provider of the service, if such a system exists. Several papers in the literature discuss the main trade-off considered when deciding on an appointment policy. For example, Bailey (1952) points out, "Again, the congestion that arises in the hospitals' waiting rooms means that an undue amount of hospital accommodation, which is also in short supply, is devoted merely to sheltering a large crowd of people, many of whom could have been given appointments at more suitable times. On the other hand, it is important that the time of the consultant in charge of the clinic is used to the best advantage. In practice, the requirement that the consultant be kept fully occupied is usually regarded as an over-riding consideration: large queues of patients are often allowed to build up in order to avoid the possibility of the consultant ever having to wait for a patient." This trade-off can be summarized as follows: if the time between appointments is short compared to average service time, then expected customer waiting time will be long and expected server idle time will be short.

In brief, the appointment policy used has a direct effect on customer waiting time and consequently on satisfaction with the service. Accordingly, the design of the appointment policy may be fundamental in explaining the success or failure of the service-providing enterprise.

Bailey (1952) was one of the first to draw attention to the importance of a well-designed appointment policy, addressing the topic from a quantitative operations management perspective. He showed that the use of quantitative tools can improve the performance of an appointment-based system of attention, in terms of controlling both customer waiting time and server idle time. Since then, a number of articles have been published which approach the topic from different points of view. Several different appointment policies have been studied under different scenarios (different service time distributions, service companies, and patterns of customer behavior toward appointments). A brief literature review is presented in Section 2.

This paper makes a critical analysis of the literature on appointment policies, focusing on the objective function commonly used and the assumptions made about the behavior of demand. In this analysis it is crucial to distinguish between a private server and a central planner, and therefore, different subsections are devoted to analyze the economic formulation in these two cases. We define a private server as an agent that maximizes his/her own utility. On the other hand, a central planner is defined as an agent, commonly a public entity, whose goal is to maximize the social welfare that includes the utility of the server plus the utility of the customers or users.

We observe in practice that public services do not always behave as central planners; it is common to find public institutions with poor services from the customers' point of view: long waiting times and little concern for customers' satisfaction. However, there are examples of public entities that, in fact, act as central planners. For example, in Chile the public health and the IRS are public services that act, in some dimensions, as central planners. Although they, in theory, can provide a low quality service in terms of the customers' waiting time, they actually incorporate the customers' utility in their objective functions. In fact, during the last four years the Chilean IRS has reduced the customers' waiting time significantly through the implementations of operations management tools (e.g., design of the optimal number of

cashiers and elimination of paperwork through Internet services, see *The Economist* (2000)). Also, the secretary of health of the current government received a mandate from the president to reduce the waiting time in public hospitals to those common in the private system within the first three months of her period (*El Mercurio* (2000)).

The rest of this paper is organized as follows. First, in Section 2 we present a brief literature review, describing the specific problems that have been addressed in the literature. In Section 3 we analyze the economic framework of the problem of designing an appointment policy and make a critical analysis of the objective functions used in the literature, discussing their economic interpretation. This is crucial for understanding the economic scenarios under which the results obtained are valid. We also question the assumption made throughout the literature that the amount of the service demanded is exogenous, i.e., demand is assumed to be independent of appointment-policy decisions. A more realistic model is proposed that includes customers' sensitivity to waiting times and adequately represents the objective function of a central planner or private server. In Section 4 we solve the new model proposed and compare the optimal appointment policies derived from it with those obtained using the formulations from the literature. Finally, in Section 5 we present conclusions and recommendations for future research.

2. Literature Review

In this section we present a brief literature review on appointment policies, describing the problems addressed and the main conclusions obtained in the literature.

Jansson (1966), Soriano (1966), Mercer (1973), Geiszler (1981), Sabria and Daganzo (1989) assume that the system reaches steady state, thereby making it possible to use results from queuing theory which provide analytical expressions to describe the long-run evolutionary behavior of the system. In Geiszler (1981) and Sabria and Daganzo (1989) this assumption is direct and realistic in the context considered; they study systems that operate continuously in time (productive processes and cargo-handling in ports, respectively). In the other papers assuming the system achieves steady state is merely a simplifying assumption.

In practice, it is common, however, to find services where the system never reaches steady state, because they operate for finite time periods of only a few hours each. Thus, the papers discussed in what follows assume that the planning horizon is finite and steady state is never reached.

Fries and Marathe (1981), Liao, Pegden, and Rosenshine (1993), Liu and Liu (1998a, 1998b), and Vanden Bosch, Dietz, and Simeoni (1999) address the problem of scheduling N people in K predetermined instants of time. Thus, the problem is to decide the number of customers to schedule in the K instants of time $(n_1, n_2, \ldots, n_K)$, such that $\sum_{i=1}^{K} n_i = N$. The length of the intervals between appointments are considered constant and of equal size, and there is no freedom to change them. They conclude that making appointments in variable-sized blocks may be better than other more common appointment systems like scheduling customers in equal blocks.

Pegden and Rosenshine (1990), Healy (1992), Wang (1993, 1997), and Stein and Cote (1994) remove the assumption that appointment instants are predetermined. They consider more general appointment policies, where appointment instants are decision variables. Therefore, these papers find the optimal vector $X = (x_1, x_2, \ldots, x_N)$, where x_i is the time of the appointment of customer i, given that there are N customers to be scheduled. We remark that, in this case, the set of feasible appointment policies considers all possible alternatives. For particular cases, where customers are punctual and do not fail to show up, service times are i.i.d. and have exponential or phase-type distributions, and there is a single server, it is possible to derive mathematical expressions to determine the optimal appointment policy. They conclude that the commonly used equally spaced appointment policy is not necessarily optimal.

Ho and Lau (1992, 1999) and Yang, Lau, and Quek (1998) consider specific sets of feasible appointment policies and analyze a large number of scenarios using simulation techniques. These scenarios consider different service time distributions, percentages of absenteeism, and numbers of customers to be served. They conclude that there is no universal appointment rule; what is best depends on the scenario being considered. For example, if customers' time has a high valuation compared to server's time in the objective function, then the optimal policy will lead to short customer waiting time. On the other hand, if the coefficient of variation of service time increases, customers will wait longer and the server will be idle for longer periods of time.

Charnetsky (1984), Weiss (1990), Klassen and Rohleder (1996), Wang (1999), and Rohleder and Klassen (2000) study appointment policies for customers that come from populations with different service time distributions. Given that service times are not i.i.d., it is relevant to decide the instant at which appointments are made and their order. This problem is common in scheduling interventions in an operating room, where different surgical procedures have different time duration distributions (Charnetsky (1984) and Weiss (1990)). Most of these papers (except Wang (1999) that uses a nonlinear optimization approach) use simulation to solve the problem. They conclude that the best policy is to make appointments in increasing order of service time variance.

Finally, Jackson, Welch, and Fry (1964), Vissers and Wijngaard (1979), Rising, Baron, and Averill (1973), Cox, Birchall, and Wong (1985), O'Keefe (1985), Babes and Sarma (1991), Brahimi and Worthington (1991), and Bennett and Worthington (1998) provide more general and qualitative recommendations for appointment policy design. Most of them are based on the use of operations research to improve service quality in medical clinics. In particular, Jackson, Welch, and Fry (1964) observe that punctuality of patients and doctors, together with the variability of service times, are key variables in running an appointment system. Bennett and Worthington (1998) suggest a flexible and open approach to deal with this type of problem, which includes qualitative and quantitative tools.

3. Economic Framework

In this section we analyze the general economic framework of the appointment policy design problem. In Subsection 3.1 we show how the problem is usually formulated in the literature in objective function terms, and we discuss the assumptions made about the behavior of demand. Then, in Subsection 3.2 we propose a new model to describe the optimization problem faced by a service company that has to choose an optimal appointment policy. This new model has more realistic customer behavior than what is usually assumed throughout the literature. Finally, in Subsection 3.3 we give an economic interpretation of the objective functions used in the literature and identify the assumptions under which these objective functions may adequately represent the problem solved by a central planner or a private server.

3.1. *Analysis of the Objective Functions Used in the Literature*

Several of the papers that study appointment policies consider an objective function equivalent to minimizing a linear combination of expected total customer waiting time and server completion time. The latter is defined as the total time the server spends in the system from the beginning of the session until the last customer scheduled for the period has been served. For example, if a physician begins to work at 8:00 AM and in a particular day he finishes serving the last customer at 12:15 PM, the completion time for that particular day would be 4 hours and 15 minutes. Furthermore, during the time the server spends in the system (the completion time), he/she is either busy serving customers or idle when the system is empty. Hence, completion time is equal to the sum of customer service time plus total idle time. This statement is valid for any customer arrival process (i.e., any type of appointment

policy and pattern of customer delays), and any service time distribution even if service times are not identically distributed.

Alternatively, some papers use total expected customer time in the system, instead of total expected customer waiting time, which only differs by a constant equal to the total expected service time. As this is equivalent from the optimization point of view, in this paper we use total expected customer waiting time to refer to both cases. The objective function mentioned above is denoted as A1 and is used in Pegden and Rosenshine (1990), Healy (1992), Wang (1993), Stein and Cote (1994), and Wang (1997, 1999).

An alternative objective function is used in Bailey (1952), Jansson (1966), Vissers and Wijngaard (1979), Charnetsky (1984), Weiss (1990), Ho and Lau (1992), Klassen and Rohleder (1996), Yang, Lau, and Quek (1998), Liu and Liu (1998b), Ho and Lau (1999), and Rohleder and Klassen (2000). These papers minimize a linear combination of expected server idle time (instead of completion time) and total expected customer waiting time; this is denoted as A2. The factors used to multiply the expected times are usually described as the "economic values of the server's and clients' times" (see Yang, Lau, and Quek (1998)).

All the studies assume that the number of customers to be attended is exogenous and independent of the problem's decision variables. This assumption implies that demand is independent of customer waiting time, which is directly affected by the appointment policy imposed. We denote by N the fixed and exogenous demand considered in the literature, which corresponds to the total number of customers to be served in a period of time. Thus, the problem solved in the papers mentioned above can be written as follows:

$$(A1) \qquad \min_{s \in S_N} \alpha N \bar{W}(s) + \beta E(t_c(s)) \qquad \text{or} \qquad \min_{s \in S_N} \alpha N \bar{W}(s) + \beta E(t_I(s)), \qquad (A2)$$

where $\bar{W}(s)$ is the average customer waiting time, which depends on the appointment policy implemented (s), α is the unit cost of customer waiting time, and β is the unit value of server time in the objective function. In addition, t_c is server completion time and t_I is server idle time. $E(X)$ is the expected value of the random variable X. The literature only considers policies in which the N customers are given scheduled appointments; we denote this set by S_N. In this case, the only decision to be made is how those N appointments are given (for example, schedule customers at intervals equal to average service time or schedule customers in blocks of two at intervals equal to twice the average service time, etc.).

As mentioned above, for a fixed number of customers, N, expected server completion time is equal to expected server idle time plus total expected customer service time, i.e.,

$$E(t_c) = E(t_I) + \sum_{i=0}^{N} E(t_s^i), \qquad (1)$$

where t_I is total server idle time and t_s^i is the time required to serve customer i. Using equation (1) and the fact that total expected customer service time ($\sum_{i=0}^{N} E(t_s^i)$) is constant over all appointment policies, we observe that the two objective functions (A1 and A2) are equivalent except for a constant term. Thus, minimizing a linear combination of total expected customer waiting time and expected server completion time (A1) is equivalent to minimizing a linear combination of total expected customer waiting time and expected server idle time (A2). Accordingly,

$$(A1) \quad \min \alpha N \bar{W} + \beta E(t_c) \Leftrightarrow \min \alpha N \bar{W} + \beta E(t_I) + \beta \sum_{i=0}^{N} E(t_s^i) \Leftrightarrow \min \alpha N \bar{W} + \beta E(t_I) \quad (A2)$$

Furthermore, the cost of server idle time (opportunity cost of idle server time) is equal to

the cost of server time. Given these equivalences, throughout the rest of the paper we use objective function A1, unless the contrary is explicitly stated.

We note that there are other papers that use alternative objective functions. Fries and Marathe (1981) and Liu and Liu (1998a) consider an objective function that includes the cost of customer expected waiting time, the cost of server idle time, and the cost associated with the expected overtime, defined as the extra time the server works beyond a predefined period. Finally, Liao, Pegden, and Rosenshine (1993) and Vanden Bosch, Dietz, and Simeoni (1999) only consider the cost associated with customer waiting time and the cost associated with overtime in the objective function. However, the analysis we make throughout the paper using objective function A1 can be easily extended to consider these different objective functions. For this purpose, it is necessary to use different costs associated with the server's operation.

The optimal appointment policy depends on the relative weights of expected server's completion time (or idle time) and expected customers' waiting time in the objective function (β/α). If this ratio is low, then the optimal policy will be one where average customers' waiting time is low. This is because customers' waiting time is more valuable compared to server's idle time in the objective function. For example, customers are scheduled individually at increasing intervals of time, greater than the average service time. On the other hand, if the ratio β/α is high, the optimal appointment policy will be one where customers suffer longer waits and the server will be occupied most of the time. For example, a big block of customers is scheduled at the beginning of the day. Therefore, in practice, a crucial factor to determine the optimal policy is to have a good estimation of the ratio β/α.

Some papers analyze the objective function and its parameters in terms of its economic interpretation. Bailey (1952) mentions that objective function A2 represents a social benefit function that takes account of the time spent by both customers and server. Referring to the costs associated with customer waiting time, Fries and Marathe (1981) point out: ". . . in private medical practice, the value of waiting time is more amorphous, including aspects of goodwill, loss of return business, etc. In the public sector, and in particular in public hospitals, it becomes yet harder to identify the aspects of the cost of waiting. . . . [in this case,] issues of goodwill, and "cost of society" place a value on the time patients wait." Finally, Wang (1999) defines α as the cost "for the system to handle the waiting client" and β as the cost "per hour for hiring a server."

The main goal of the literature on appointment systems has been to find an optimal appointment policy that minimizes the corresponding objective function, which incorporates the fundamental trade-off between customers' waiting time and server's idle time. We believe, however, that the literature lacks analysis in terms of defining the implicit economic framework in the objective functions considered and giving a correct interpretation of the fundamental parameters of the model (β/α). When analyzing and implementing appointment policies derived from models, it is crucial to know whether the objective function describes a private server or a central planner, and what revenues and costs are involved.

In the next subsection we propose a new model to address the problem of choosing an appointment policy, which includes more realistic assumptions about the economic behavior of both server and customers (they are sensitive to waiting time). We formulate the model for both a central planner and a private server. The model is also useful for analyzing the economic interpretation of the objective functions commonly used in the literature. The model has some similar elements to the one used in Stenbacka and Tombak (1995), where they analyze the effects of privatizing a service company, in particular, the effect on service speed.

3.2. *New Model*

We assume the existence of a population of risk-neutral customers that potentially want to receive the service. Utility in each case is given by the value the service has for the customer

concerned, which is defined as the maximum amount of money he or she is willing to pay for the service when there is no waiting time (*the reservation price*) minus the price of the service and the cost associated with waiting time. We denote the utility of customer i by u_i, which can be written as follows:

$$u_i = r_i - p - aw,$$

where r_i is the customer's reservation price, p is the price of the service, w is the customer's waiting time, and a is the unit cost of waiting time for the client. Customers make their demands for the service effective if the expected utility of doing so is positive, i.e.,

$$E(u_i) = r_i - p - a\bar{W} > 0, \tag{2}$$

where $\bar{W}$ is the expected waiting time ($E(w)$). We note that the expected utility of a potential customer that does not receive the service is equal to zero.

A similar customer utility function has also been used in papers dealing with the problem of finding pricing mechanisms that induce socially optimal customer queuing behavior, considering that a customer produces a negative externality when joining a queue, as all customers behind him have to wait longer (see, for example, Mendelson and Whang (1990) and Hassin (1995)). This utility function assumes customers are risk-neutral with respect to waiting time. An alternative utility function for risk-averse customers might consider not only expected waiting time but also its variance. For example, a customer might prefer a system in which there is a certain wait of 10 minutes, rather than a service where half the time there is no wait at all and in the other half waiting time is 20 minutes long. In fact, Leclerc, Schmitt, and Dubé (1995) show through empirical studies that individuals tend to be risk-averse in decisions that involve the use of their time. One possible explanation is that they want to avoid risk in the time domain so as to plan better.

We assume that the potential customer population is of size M, i.e., M is the number of customers that would demand the service if waiting time and the price were both zero. In a more general model this potential demand could be described by a stochastic process, for example, a Poisson process. Although every customer knows his or her own reservation price, from the server's (or decision-maker's) point of view these are independent random variables. Without loss of generality, we assume that customers come from a homogeneous population, i.e., their reservation prices are identically distributed. We define R as the random variable that represents the reservation price of a customer from the server's perspective and $f_R(r)$ as its pdf. $f_R(r)$ is positive in the interval $[\underline{R}, \bar{R}]$ and zero elsewhere ($r < \underline{R}$ and $r > \bar{R}$ with $0 \leq \underline{R} < \bar{R}$). We note that the results presented below can also be obtained by defining several market segments with different reservation-price probability density functions.

Given an appointment policy s and a price p, the number of customers who demand the service is a binomial random variable with parameters M and $\Pr[R > p + a\bar{W}(s, p)]$. Then,

$$\Pr[G(s, p) = i] = \binom{M}{i} (\Pr[R > p + a\bar{W}(s, p)])^i (1 - \Pr[R > p + a\bar{W}(s, p)])^{M-i},$$

$$i = 0, \ldots, M, \tag{3}$$

where $G(s, p)$ is the demand for the service. Thus the expected number of customers who demand the service is equal to

$$E(G(s, p)) = M \cdot \Pr[R > p + a\bar{W}(s, p)] = M \int_{p + a\bar{W}(s,p)}^{\bar{R}} f_R(r)dr. \tag{4}$$

If the waiting time and/or the price increase, the quantity of the service demanded decreases. We assume that customers have *rational expectations* (Muth (1961)) in estimating their

expected waiting time ($\bar{W}(s, p)$). This means that the expectations customers form about their average waiting time is the same for all customers and proves to be correct; i.e., it is exactly the value this variable does in fact take. Consequently, given the appointment policy s defined by the server, and a price p, customers estimate a value $\bar{W}(s, p)$, which they use to make their demand decisions. This demand (equation (3)) generates an average waiting time that coincides exactly with their initial estimate. When assuming that customers have rational expectations in estimating their average waiting time, we implicitly assume that they learn from experience. Using data gathered in previous visits to the service and information provided by other customers, they are able to make a correct estimate of average waiting time, which is quite close to reality. Even though customers of a service are not able to estimate the complete distribution of the waiting time, they might be capable of estimating its average. This is a widely used assumption in the economics literature.

In the model defined above, an appointment policy that generates long customer waiting times leads to a reduction in the quantity demanded, which concurs with practical observations in most services. Empirical studies by Taylor (1994) and by Katz, Larson, and Larson (1991) find that waiting time has a negative effect on customers' overall assessment of the service. Moreover, Boulding, Kalra, Staelin, and Zeithaml (1993) conclude that the more satisfied customers are with service quality, the more likely their behavior will benefit the company's profitability (by demanding the service again, recommending it, etc.). Accordingly, we can claim that demand is sensitive to waiting time in most service industries; the longer the waiting time, the more customers will turn to the competition or simply they may decide not to seek the service.

Let K ($K \leq M$) be the number of available slots in an appointment policy, i.e., K is the maximum number of customers that can be accommodated. Therefore, in this model an appointment policy is defined not only in terms of how the appointments are given, but also in the maximum number of customers to be scheduled. Thus, the model also allows for capacity decisions, which gives more flexibility when defining an appointment policy compared to the models in the literature. For example, overbooking policies can be implemented when the percentage of absenteeism is high.

Given an appointment policy s, with K available slots, and a price p, the expected number of customers attended is equal to

$$E(Q(s, p)) = E[\min(K, G(s, p))]$$

$$= \sum_{i=0}^{K} i \cdot \Pr[G(s, p) = i] + K \sum_{i=K+1}^{M} \Pr[G(s, p) = i], \tag{5}$$

where $Q(s, p)$ corresponds to the number of customers attended to.

In the case of a private server whose goal is to maximize profits, the server's expected utility is given by the revenue obtained from serving customers minus the expected cost. We assume this cost to be associated with the server's completion time. Thus, defining the server's expected utility as $E(\pi)$, the problem faced by a private server when choosing an appointment policy is equal to

$$\max_{s \in S, p} E(\pi(s, p)) = pE(Q(s, p)) - bE(t_c(s, p)), \tag{6}$$

where $t_c(s, p)$ is the completion time, b is the unit value of server time, and S is the set of feasible policies. As we previously mentioned, an appointment policy s is defined by the number of slots available to serve customers in a given session, and the way in which those appointments are scheduled. In the general case, the server also has to decide the price of the service. Note that we explicitly include only costs that depend directly on the appointment

policy used. Other costs are invariant from the optimization point of view (e.g., electricity and payroll).

We now consider the problem faced by a central planner whose goal is to maximize expected social welfare, defined as the sum of total expected consumer utility plus the expected utility of the server. We define $E(U_T(s, p))$ as total expected customer utility, namely, the sum of the expected individual utilities of customers obtaining the service. As discussed at the beginning of this subsection, the expected utility of customers that do not receive the service is zero. Given an appointment policy s, with K available slots and a price p, total expected customer utility is equal to

$$E(U_T(s, p)) = \sum_{i=0}^{K} i \cdot E(u|R > p + a\bar{W}(s, p)) \cdot \Pr[G(s, p) = i] + K \sum_{i=K+1}^{M} E(u|R > p + a\bar{W}(s, p)) \cdot \Pr[G(s, p) = i], \quad (7)$$

where $E(u|R > p + a\bar{W}(s, p))$ is the expected utility of a single customer who obtains the service, and

$$E(u|R > p + a\bar{W}(s, p)) = \frac{1}{\Pr[R > p + a\bar{W}(s, p)]} \int_{p+a\bar{W}(s,p)}^{\bar{R}} (r - p - a\bar{W}(s, p)) f_R(r)dr. \quad (8)$$

Considering that $E(u|R > p + a\bar{W}(s, p))$ is constant over all i, replacing equation (8) in (7) and recalling equation (5) we have

$$E(U_T(s, p)) = \frac{1}{\Pr[R > p + a\bar{W}(s, p)]} \int_{p+a\bar{W}(s,p)}^{\bar{R}} (r - p - a\bar{W}(s, p)) f_R(r)dr \cdot E(Q(s, p)). \quad (9)$$

Defining social benefit by B^s, the problem faced by a central planner is given by:

$$\max_{s \in S, p} E(B^s(s, p)) = E(U_T(s, p)) + E(\pi(s, p)), \quad (10)$$

where $E(U_T(s, p))$ is given by equation (9) and $E(\pi(s, p))$ by equation (6).

Thus, equation (6) and equation (10) define the optimization problems faced by a private server and a central planner, respectively.

In the next section we determine the relationship between the objective functions used in the literature and those corresponding to our new model, which recognizes customers' sensitivity to waiting times.

3.3. *Economic Interpretation of the Objective Functions Used in the Literature*

Throughout the literature it is implicitly assumed that the price p is fixed and exogenous, so we rule out any dependence thereon. To derive an economic interpretation of the objective functions used in the literature, we also limit the feasible policy space to the set S_M, i.e., the set containing policies in which there are available slots for all potential customers (M).

In the literature, it is also assumed that the number of customers to be attended is exogenous and does not depend on the problem's decision variables. This implies a fixed demand (N) for all appointment policies. Thus, we also consider that the fixed number of customers to be attended, assumed in the literature, is equal to the size of the potential customer population, and therefore, $M = N$.

In this case, recalling equation (4), the expected number of customers served by an appointment policy $s \in S_N$ is equal to

$$E(Q(s)) = E[\min(N, G(s))] = E(G(s)) = N \int_{p+a\bar{W}(s)}^{\bar{R}} f_R(r)dr. \tag{11}$$

The private server's expected utility is given by (replacing equation (11) in equation (6)):

$$E(\pi(s)) = pN \int_{p+a\bar{W}(s)}^{\bar{R}} f_R(r)dr - bE(t_c(s)). \tag{12}$$

The assumptions that potential demand is equal to N and the space of feasible policies is the set S_N leads to (replacing equation (11) in equation (9)):

$$E(U_T(s)) = N \int_{p+a\bar{W}(s)}^{\bar{R}} (r - p - a\bar{W}(s)) f_R(r)dr. \tag{13}$$

Thus, using equations (12) and (13), expected social welfare in this particular case is given by:

$$E(B^s(s)) = N \int_{p+a\bar{W}(s)}^{\bar{R}} (r - a\bar{W}(s)) f_R(r)dr - bE(t_c(s)), \tag{14}$$

since $pN \int_{p+a\bar{W}(s)}^{\bar{R}} f_R(r)dr$ is only a transfer from the customers to the server.

In the following subsections we determine how well the objective functions used in the literature represent the problem solved by a central planner or a private server. In doing so, we use our new model including the assumptions mentioned in this subsection, i.e., equations (12) and (14).

3.3.1. Central planner. Throughout the literature it is assumed that the number of customers to be attended is exogenous; i.e., independent of the waiting time. In our model, this assumption is equivalent to assuming that $p + a\bar{W}(s) < \underline{R}$, if demand for the service is N, $\forall s \in S_N$. In other words, for all feasible appointment policies, the value customers place on the service is sufficiently high for everyone to demand it, no matter how long their expected waiting time may be. Building this assumption into our model (14) and keeping in mind that $f_R(r)$ is zero if $r < \underline{R}$, the social benefit maximization problem can be written as

$$\max_{s \in S_N} E(B^s(s)) = N \int_{\underline{R}}^{\bar{R}} r f_R(r)dr - aN\bar{W}(s) \int_{\underline{R}}^{\bar{R}} f_R(r)dr - bE(t_c(s)). \tag{15}$$

Given that the first term is equal to $NE(r)$ and is constant for all appointment policies, and that the second integral is equal to one, then the central planner case problem corresponds to $\max_{s \in S_N} - aN\bar{W}(s) - bE(t_c(s))$ or, equivalently, to $\min_{s \in S_N} aN\bar{W}(s) + bE(t_c(s))$, which is the objective function (A1) used in Pegden and Rosenshine (1990), Healy (1992), Wang (1993), Stein and Cote (1994), and Wang (1997, 1999), considering $\alpha = a$ and $\beta = b$. Thus, given the assumptions made in these papers (fixed price and exogenous demand), the objective function does a good job of describing a central planner whose goal is to maximize social welfare. Another important implication of the above derivation relates to the direct interpretation of the parameters involved in the objective function, which is useful for estimation purposes. In this particular case, the parameter α represents the unit cost of waiting time for the customer, and β corresponds to the unit operating cost for the server.

The common assumption in the literature, that demand is exogenous, is equivalent to assuming that it is not sensitive to the waiting time. This might be close to reality in some special cases, for example, where the service company sells a "prime necessity" service (or

product)—for example, a public hospital or the IRS. In these cases, customers place a high value on the service so they are willing to obtain it regardless of the waiting time, and the service can nearly always fill all the appointment times. Other examples include the case of manufacturing companies where "customers" are products or parts instead of people. Nowadays, however, customers do business with the company that delivers the "best" service and where speed is a fundamental attribute (see, for example, Taylor (1994) and Katz, Larson, and Larson (1991).) In these cases the problem to be solved is to maximize the expected social benefit, considering that the quantity of the service demanded depends on waiting time, which in turn depends on the appointment policy imposed, and is therefore endogenously determined in the model. In addition, the maximization problem is carried out on a feasible set containing policies that have from one to N available slots and is not restricted to policies with exactly N available slots (set S_N).

It should be noted that the expected customers' waiting time appears in the objective function (10) not only because customers' utility in $E(B^s)$ is considered explicitly, but also because it determines the quantity demanded. The longer the waiting time, the fewer the number of customers served and the lower the social benefit. In these cases, objective function A1 is only a shortcut to the true objective function of the central planner. To make the right decisions with these objective functions, it is crucial to correctly choose the ratio β/α.

Thus, in the general case, we assume that the server faces a positive demand that reacts to waiting time, i.e., $\underline{R} < p + a\bar{W}(s) < \bar{R}$, $\forall s \in S_N$. For example, when assuming that R is uniformly distributed between 0 and $\bar{R}$, maximizing $E(B^s)$ is equivalent to maximizing (recalling equation (14)):

$$\frac{Na^2}{2\bar{R}}\,\bar{W}^2 - aN\bar{W} - bE(t_c),$$

which is a nonlinear function of $\bar{W}$. Additionally, $\bar{W}$ is the expected waiting time corresponding to the rational expectations equilibrium. We remark that this expected waiting time is different to the one obtained when demand is fixed (N) as assumed in the literature (A1). Although these two quantities increase or decrease together according to the implemented appointment policy, they do not necessarily take the same value. Therefore, a single parameter β/α in A1 is, in fact, a complex function of several fundamental quantities (a, b, and the reservation price R). Thus, the calibration of the ratio β/α commonly used in the literature is not an easy task. In Section 4 we develop some numerical examples to compare the optimal policies obtained when maximizing $E(B^s)$ to those obtained when using objective functions A1 or A2 for different values of β/α. We also do numerical examples for the case of a private server.

3.3.2. Private server. In the case of a profit-maximizing private server, it is not possible to give any direct interpretation of the objective functions commonly used in the literature. They can only be interpreted as a shortcut to a private objective function. Thus, minimizing these objective functions would lead to "reasonable" appointment policies from a private point of view. However, we show that using this shortcut is inconsistent with the widely held assumption that the number of customers to be served is independent of waiting time.

Without loss of generality, we assume that the relevant cost for the server is that associated with completion time. As in the analysis of the central planner case, we assume a fixed potential demand equal to N and set of feasible policies equal to S_N. Thus, from equation (12), the objective function for the private server can be written as

$$\max_{s\in S_N} pN \int_{p+a\bar{W}(s)}^{\bar{R}} f_R(r)dr - bE(t_c(s)). \tag{16}$$

Despite the fact that consumers' utility is not considered in the private server's objective function, the latter does internalize the waiting time imposed customers through the demand function; accordingly there are incentives to provide a good service. Indeed, if waiting time is short, a higher price can be charged since customers' willingness to pay for the service increases. The private server also perceives a cost by making customers wait as a result of market coverage. The longer the waiting time, the lower the demand, so the server earns lower sales revenue. Therefore, average waiting time appears in the objective function of a private service provider if it faces a positive demand that reacts to waiting time. As we previously mentioned, one way of introducing this fact in the model is to assume that $\underline{R} < p + a\bar{W}(s) < \bar{R}$, $\forall s \in S_N$. Moreover, to see the direct relationship with the objective function A1, we assume that $f_R(r)$ is uniformly distributed in $[\underline{R}, \bar{R}]$. Given this, the private objective function (16) can be written as

$$\max_{s \in S_N} pN \frac{(\bar{R} - (p + a\bar{W}(s)))}{\bar{R} - \underline{R}} - bE(t_c(s)).$$

The above problem is equivalent to

$$\min_{s \in S_N} \frac{pa}{\bar{R} - \underline{R}} N\bar{W}(s) + bE(t_c(s)). \tag{17}$$

This objective function is similar to A1, except for the factors that multiply $\bar{W}(s)$ and $E(t_c(s))$. We also notice that the expected waiting time and completion time considered in our objective function correspond to the rational expectations equilibrium, which do not necessarily correspond to those obtained using the literature's objective functions. In the latter case, the expected waiting time and completion time considered are the ones reached when demand is fixed and equal to N.

In some cases, when the ratio β/α is correctly estimated, the use of objective function A1 would lead to the optimal appointment policy according to the model proposed in this paper. However, this is not an easy task, considering that the values of $\bar{W}(s)$ and $E(t_c)$ are different from the ones obtained using the objective function developed in this paper. Also, the ratio β/α has to summarize a series of effects (p, a, b, and R) in a complex way.

As we previously mentioned, a flaw in the equivalence stated above is that in the literature the number of customers to be served during a session is assumed known, so demand remains fixed for all appointment policies, or equivalently, it does not react to waiting time. However, it is clear that different appointment policies will lead to different waiting times, and, as we assumed initially to derive equation (17), to different numbers of customers served. In fact, average waiting time appears in the private server's objective function because it determines the quantity demanded, a fact that contradicts the assumption made in the literature that demand is fixed. Thus, in industries where demand reacts to waiting time, a model to determine the optimal appointment policy from the private standpoint should maximize revenue and explicitly consider that the number of customers to be served during a given period of time is endogenously determined in the model. Thus, one possible formulation is described by our model (equation (6)). For other cost structures, the conclusions are analogous.

4. Impact of Different Objective Functions on Optimal Appointment Policies

In this section, we numerically determine the impact of using the literature's objective functions on optimal appointment policies. For example, we quantify the revenue losses that incur a private firm that faces a waiting time sensitive demand function (objective function (12)), but determines its appointment policy using the literature's objective functions, i.e., it uses a fixed demand that does not react to waiting time. We make a similar analysis for the

case of a central planner (objective function (14)). In this analysis, we keep the assumptions made in Subsection 3.3: the feasible policy space is the set containing policies in which slots are available for all potential customers. We also consider that the fixed number of customers to be attended, assumed in the literature, is equal to the size of the potential customers' population.

To assess the analysis described above, we use the framework proposed by Ho and Lau (1992). In that paper, the authors present a study of appointment policies under a broad range of managerial scenarios. They study the performance of eight different appointment rules (selected out of 50) under different scenarios defined by the coefficient of variation of service time (CV), the number of customers to be attended (N), and the percentage of absenteeism (γ).

They consider 27 different scenarios and choose the optimal appointment policy for each of them, which in turn depends on the ratio β/α. As described in the previous section, this ratio corresponds to the relative importance given to server's idle time and customers' waiting time in the objective function A2.

We define A_i as the appointment time for customer i, μ as the expected value, and σ as the standard deviation of the service times, respectively. In Table 1 we describe the eight rules used in Ho and Lau (1992) (we keep the same order used in the paper).

The appointment rules are listed in a decreasing order of average customers' waiting time. Thus, appointment rules 5 and 8 lead to the highest and lowest average customers' waiting times, respectively.

4.1. *Model's Resolution*

We briefly describe the algorithm used to solve the mathematical model proposed in this paper for the private server and the central planner, respectively.

For the case of a private server, we maximize the objective function (12) and find the optimal appointment policy among the eight rules proposed by Ho and Lau (1992), for each managerial scenario. For this purpose, we apply the following algorithm:

0. **STEP 0:** Initialization.
   - Set the values of CV, N, γ, a, b, p, and a pdf for R.
   - $S_N = s_1, s_2, \ldots, s_8$, where s_i corresponds to the ith appointment rule proposed by Ho and Lau.
   - $k = 1, \quad s = s_k$.
1. **STEP 1**: Find the rational expectation equilibrium for the customers' waiting time: $\bar{W}^*(s)$, solving the following fixed point equation:

$$\bar{W}(s) = T(\bar{W}(s), s), \tag{18}$$

TABLE 1

The Eight Appointment Rules Used in Ho and Lau (1992)

Rule Number	Description
5	$A_1 = A_2 = A_3 = A_4 = 0$; for $i > 4$, $A_i = A_{i-1} + \mu$
3	$A_1 = 0$, $A_2 = 0.3$, $A_3 = 0.6$, $A_4 = 0.9$; for $i > 4$, $A_i = A_{i-1} + \mu$
4	$A_1 = 0$, $A_2 = 0.5$, $A_3 = 1.0$, $A_4 = 1.5$; for $i > 4$, $A_i = A_{i-1} + \mu$
2	$A_1 = 0$, $A_2 = 0.2$, $A_3 = 0.6$; for $i > 3$, $A_i = A_{i-1} + \mu$
1	$A_1 = A_2 = 0$; for $i > 2$, set $A_i = A_{i-1} + \mu$
6	$A_i = (i - 1)\mu - k\sigma$, $(k = 0.1)$
7	$A_i = (i - 1)\mu$; for $i \leq 5$, $A_i = A_i - 0.15(5 - i)\sigma$ for $i > 5$, $A_i = A_i - 0.3(5 - i)\sigma$
8	$A_i = (i - 1)\mu$; for $i \leq 5$, $A_i = A_i - 0.25(5 - i)\sigma$ for $i > 5$, $A_i = A_i - 0.5(5 - i)\sigma$

where $T(\bar{W}(s), s)$ = average customer waiting time given $\bar{W}(s)$ and s. We note that for a given value of the customers' waiting time, $\bar{W}(s)$, a stochastic demand for the service, is generated, which has a binomial distribution with parameters N and $\Pr[R > p + a\bar{W}(s)]$. This demand, in turn, generates an average waiting time ($T(\bar{W}(s), s)$) which is computed using simulation (we use a precision of 1% with a level of confidence of 95%, as in Ho and Lau (1992).) To find the fixed point, we solve equation (18) with the Van Wijngaarden, Dekker, and Brent method (Press, Flannery, Teukolsky, and Vetterling (1990)).

2. **STEP 2**: Once $\bar{W}^*(s)$, corresponding to the rational expectations equilibrium, is known, determine $E(\pi(s)) = pN \int_{p+a\bar{W}(s)}^{\bar{R}} f_R(r)dr - bE(t_c(s))$ using simulation.
3. **STEP 3**: If $k = 8$, go to Step 4. Otherwise, $k = k + 1$, $s = s_k$ and return to STEP 1.
4. **STEP 4**: Determine $\max_{s \in S_N} E(\pi(s))$ and assign the value of the optimal appointment rule and the maximum expected utility to s^* and $E(\pi(s^*))$, respectively.

A similar procedure is used to find the optimal appointment rule for the case of a central planner.

We programmed the algorithm in *C* and ran it in a Pentium III PC. For each scenario, the algorithm took 5 minutes on average to find the optimal appointment policy.

4.2. *Numerical Examples*

In this subsection we compare the performance of the optimal policies obtained using the literature's objective function against those obtained using the economic framework proposed in this paper. In order to do this, for a given managerial scenario, we compute the optimal policy for the private server and the central planner, respectively, and compare the resulting optimal objective function against the one obtained, if the optimal policy proposed by Ho and Lau (1992) was implemented.

For the numerical experiments, we consider a set of managerial scenarios consisting of different values of the parameters involved in the problem. Similar to Ho and Lau (1992), we consider different values of the coefficient of variation of the service times ($CV = 0.2$, 0.5, and 1), of the total number of customers to be served ($N = 10$, 20, and 30) and of the no-show probability ($\gamma = 0$, 0.1, and 0.2). Additionally, different values for the reservation price expected value were used ($E(R) = 1$, 2, and 4) in order to include services that are more and less valuable for the customers. Finally, different values of the unit value of the server time were used ($b = 0.3$ and 0.7) to incorporate services with low- and high-operation costs. We also set the price (p) equal to 1 and the unit cost of waiting time for the customer (a) equal to 0.3.

In what follows we present the results obtained in a specific managerial scenario; we remark, however, that similar results were obtained in the other cases. The scenario is given by a uniform service time distribution with an expected value of 1 and a coefficient of variation of 0.5 ($CV = 0.5$), a total number of customers to be served equal to 20 ($N = 20$), no absenteeism ($\gamma = 0$), a service price equal to one ($p = 1$), a Weibull distribution for the reservation price with a coefficient of variation of 0.5, and a unit cost of waiting time for the customer of 0.3 ($a = 0.3$). In the experiments, we use different values for the reservation price expected value ($E(R)$) and for the unit value of the server time (b). Table 2 shows the optimal policy obtained by Ho and Lau (1992) for the scenario when $N = 20$, $CV = 0.5$, and $\gamma = 0$, for different ranges of the ratio β/α. The objective function used by these authors is given by $\min_{s \in S_N} \alpha N\bar{W}(s) + \beta E(t_I(s))$ (*A2*). The first column shows the different ranges for β/α and the second column contains the optimal appointment rule obtained using objective function A2, within each range for the ratio β/α. We note that a higher value of β/α leads to an appointment rule with higher average customers' waiting time.

4.2.1. Private server.

Table 3 presents a comparison of the performance of the optimal appointment rules obtained with our model against those obtained with the Ho and Lau

TABLE 2

Results Obtained by Ho and Lau (1992): *Optimal Appointment Rule for Different Values of* β/α ($N = 20$, $CV = 0.5$, $\gamma = 0$)

Range of β/α	No. of Optimal Rule
>79.5	5
37.4–79.5	3
29.1–37.4	4
19.2–29.1	2
15.1–19.2	1
6.7–15.1	6
2.6–6.7	7
<2.6	8

model. The first two columns contain different expected value levels for the customers' reservation price (the value that the service has for the customer) and for the unit value of the server time, respectively. Columns 3, 4, and 5 contain the optimal appointment rule (s^*), which is independent of β/α, the expected customers' waiting time obtained (corresponding to the rational expectations equilibrium) when implementing the optimal policy (W^*), and the maximum expected private utility ($E(\pi)^*$), obtained when solving the problem developed in this paper.

Finally, in the last eight columns, we show the percentage difference of the optimal expected utility obtained using our model and the one obtained using the optimal appointment rule recommended by Ho and Lau (1992), for different ranges of the ratio β/α. In order to do this, for each range of the ratio β/α, we evaluate the objective function for a private server, proposed in this paper (equation (12)), using the optimal appointment rule obtained by Ho and Lau (1992) (Table 2). We define this objective function by $E(\pi)_{Ho-Lau}$. Finally, we compute the percentage difference as

$$\Delta E(\pi) = 100 \cdot \frac{E(\pi)^* - E(\pi)_{Ho-Lau}}{E(\pi)^*}.$$

For example, when $E(R) = 2$ and $b = 0.3$, the optimal appointment policy is 7, which leads to an expected waiting time of 0.5 units and to an expected utility of \$10.2. When using Ho and Lau's formulation, the appointment rule obtained coincides with the optimal appointment policy only when $\beta/\alpha \in [2.6, 6.7]$. For any other estimation of this ratio, there

TABLE 3

Comparison of Appointment Rules for Different Objectives Functions for a Private Server

Parameters		Optimal Appointment Policy			$\Delta E(\pi)$ for Different Ranges of β/α							
E(R)	b	s^*	W^*	$E(\pi)^*$	>79.5	37.4–79.5	29.1–37.4	19.2–29.1	15.1–19.2	6.7–15.1	2.6–6.7	<2.6
1	0.3	8	0.4	5	41.4	31.4	22.9	24.8	19.4	8.3	0.1	0.0
	0.7	6	0.6	1.7	27.3	16.2	8.3	9.4	5.7	0.0	6.9	11.0
2	0.3	7	0.5	10.2	24.0	14.2	8.9	8.4	5.4	2.6	0.0	0.7
	0.7	1	1.2	3.8	13.6	4.2	1.0	0.6	0.0	1.7	17.4	31.2
4	0.3	6	1	12.6	5.2	2.3	0.9	0.8	0.3	0.0	1.8	4.0
	0.7	3	2	5.1	1.6	0.0	1.5	1.2	3.0	6.3	26.0	40.9

would be a loss in the private server' utility. Furthermore, as long as this estimation departs from this range, the utility loss increases until it reaches a maximum value of 24%.

From Table 3, we observe that higher values of b lead to higher customers' waiting times in the optimal appointment policy. In these cases, the unit value of the server time (b) is more expensive, and therefore, the optimal policies reduce the idle time at expenses of the customers' waiting time. Similarly, when the expected customers' reservation price $E(R)$ increases (a higher customer's willingness to pay for the service), the optimal appointment rules lead to higher customers' waiting times. Given that the price is fixed, a higher value of $E(R)$ implies a higher customers' willingness to wait to be served.

In the computational experiments, we also observed that if a (the unit cost of customers' waiting time) is low, the optimal appointment policies lead to high customers' waiting times.

Thus, in service systems where b and $E(R)$ are high and a is low, the optimal appointment rules lead to high customers' waiting times. Therefore, when using the objective function proposed by Ho and Lau (1992), it is necessary to use a high value for the ratio β/α in order to obtain appointment rules close to the optimal ones. It is not an easy task to know the exact value for this ratio. As it was discussed earlier in the paper, the ratio corresponds to a complex function of the customers' behavior fundamental parameters as well as the structural model's parameters.

On the other hand, several papers in the literature suggest to choose the parameters α and β in objective function proposed by Ho and Lau (1992) as the customers' waiting cost per unit time and the server's idle cost per unit time, respectively (see Yang, Lau, and Quek (1998)). This implies that $\alpha = a$ and $\beta = b$.

In most of these cases, there are significant losses in the server's utility. These losses are shown in the last column of Table 3. In these experiments, we use a value of $a = 0.3$, and therefore, Table 3 considers ratios of $\beta/\alpha = b/a$ equal to 1 and 2.3. We observe that there are important utility losses for the higher values of this ratio; for example, there is a 41% difference in the utility when $E(R) = 4$ and $b = 0.7$, because of the implementation of a suboptimal policy. When this ratio increases (over 1) the optimal policies lead to higher customers' waiting times. However, the appointment rule chosen in Ho and Lau (1992) corresponds to 8, which keeps customers' waiting time to a low level.

4.2.2. Central planner. Table 4 is similar to Table 3 and shows the results for the central planner case. We observe that the maximization of the social welfare leads to optimal appointment policies with lower expected customers' waiting times compared to those obtained with the private server's objective function (fourth column in Tables 3 and 4). Thus,

TABLE 4

Comparison of Appointment Rules for Different Objectives Functions for a Central Planner

Parameters		Optimal Appointment Policy			$\Delta E(B^s)$ for Different Ranges of β/α							
E(R)	b	s^*	W^*	$E(B^s)^*$	>79.5	37.4–79.5	29.1–37.4	19.2–29.1	15.1–19.2	6.7–15.1	2.6–6.7	<2.6
1	0.3	8	0.4	8.3	46.0	35.9	27.1	29.1	23.3	11.0	0.7	**0.0**
	0.7	8	0.4	4.8	41.9	31.0	21.7	23.6	18.0	6.3	0.0	**0.0**
	1.5	5	1.4	0.0	0.0	16.3	39.9	29.0	54.8	119.9	**257.8**	291.8
2	0.3	8	0.3	29.7	36.5	25.6	18.6	18.2	14.0	9.2	1.9	**0.0**
	0.7	8	0.3	22.2	35.1	23.3	15.5	15.4	10.8	6.4	0.2	**0.0**
	1.5	6	0.9	8.5	34.5	17.9	9.0	8.7	4.0	0.0	**4.5**	15.2
4	0.3	8	0.3	70.5	18.2	12.3	9.0	8.6	6.8	4.7	0.7	**0.0**
	0.7	8	0.3	61.4	17.1	10.9	7.5	7.1	5.3	3.3	0.0	**0.0**
	1.5	7	0.5	44.2	16.1	8.7	4.9	4.8	2.8	1.3	**0.0**	1.9

the expected number of customers attended is higher in the case of a central planner than in the case of a private server. For example, for $E(R) = 2$ and $b = 0.7$, the customers' expected waiting time is four times higher for the case of a private server than for the one obtained in the central planner case (1.2 units of time compared with 0.3). This is due to the fact that a central planner explicitly considers customers' utility in its objective function.

For the scenario analyzed above ($E(R) = 2$ and $b = 0.7$), only in the case when $\beta/\alpha < 2.6$, the formulation A2 used in the literature leads to the optimal appointment policy. In any other case, there is an utility loss that increases as long as the ratio β/α also increases. The maximum loss is obtained when $\beta/\alpha > 79.5$, with a loss of 35.1%.

If the parameters α and β in objective function used in Ho and Lau (1992) are chosen as the customers' waiting cost per unit time and the server's idle cost per unit time, respectively (i.e., $\alpha = a$ and $\beta = b$), it is possible to obtain significant utility losses. The boldface numbers in Table 4 show the percentage social welfare loss when appointment policies are chosen using the objective function in Ho and Lau (1992) and the ratio β/α is chosen using the parameters described above. We observe that for the cases where the ratio is 5 ($\alpha = a = 0.3$ and $\beta = b = 1.5$), we obtain utility losses from 0 to 257.8%.

From Table 4, we observe that for small values of b, the optimal appointment policy leads to the lowest customers' expected waiting time. However, when b increases, this is not valid any longer. For example, when $E(R) = 2$ and $b = 1.5$, the optimal appointment policy is 6, which is not the most convenient rule from the customers' point of view.

On the other hand, for higher values of $E(R)$, the optimal appointment rules lead to lower customers' waiting times (see the case $b = 1.5$). Thus, in the case of a central planner, when customers assign a higher value for the service, they also receive a faster service, i.e., the central planner does not take advantage from the higher disposition of customers to buy the service. This is due to the fact that the central planner incorporates the customers' utility in its objective function; lower waiting times will lead to a larger demand and a higher total customers' utility and, therefore, result in a higher social welfare. This effect is more important as $E(R)$ increases. The statement above is not valid for the cases where the customers' valuation of the service is extremely high, and therefore, all customers demand the service, independently of the waiting time. We notice that the effect of $E(R)$ over the optimal appointment policy described above is the opposite for a private server.

In contrast, in the computational experiments we observed, if a decreases, the optimal appointment policy makes customers wait more, similarly to the private server case.

4.3. *Managerial Implications*

The mathematical formulation commonly used in literature assumes a fixed demand that is insensitive to the customers' waiting time. Thus, in the general, an important feature of customers' behavior is not considered in the models. However, in some instances, where the ratio β/α is "adequately" chosen, the optimal appointment policies that consider a demand function that reacts to customers' waiting time can be selected. In order to choose this ratio adequately, it is crucial to distinguish between a private server and a central planner. Furthermore, as we analyzed in Section 3, these parameters are a function of several fundamental parameters (e.g., customers' willingness to pay for the service, customers' value of waiting time, and server's value of his/her time). Therefore, an adequate estimation of this ratio is not direct. If the estimation belongs to a reasonable range, then the optimal policy is chosen. Otherwise, it is possible to incur in significant revenue losses.

The model developed in this paper proposes an economic framework that allows the understanding of the economic insights of the problem. For example, in Chile we observe that some public services do not deliver a high-quality service and customers wait for long periods of time before being served. Why? An immediate answer to this question is that the public server is not acting as a central planner, i.e., it does not consider the customers' utility in its objective function. If this were not the case, the public server would select an

appointment policy where customers would wait much less. Furthermore, if the service is highly valued by customers (a high value for $E(R)$), then the service should be even better. The only case where it is optimal for a central planner to generate high customers' waiting times is when the unit cost of waiting time for the client is small and the unit value of the server time is high, for example, the case of very specialized medical equipment.

As we mentioned before, similar results to those presented in Subsection 4.2 were obtained in several computational experiments, with a wide variety of managerial scenarios. Depending on the scenario chosen, the losses in the private server and the central planners' expected utility are more or less sensitive to the correct estimation of the ratio β/α. For example, a very sensitive scenario for a private server is $N = 20$, $CV = 0.5$, $\gamma = 0.2$, $p = 1$, $a = 0.3$, $E(R) = 2$, and $b = 0.7$. In this scenario, the optimal appointment rule for a private server is 5. If β/α is greater than 9.6, then the appointment rule chosen by Ho and Lau (1992) is the optimal one. However, if β/α is between 6.3 and 9.6, the utility loss is 9% and it starts growing as β/α decreases. Consequently, if β/α is less than 1.6, the utility loss is 83%. On the other hand, a scenario which is not so sensitive is $N = 10$, $CV = 0.5$, $\gamma = 0$, $p = 1$, $a = 0.3$, $E(R) = 4$, and $b = 0.3$. In this case, the optimal appointment rule for a private server is 1 and the worst appointment rule, in terms of expected utility, is 5. However the utility loss is only 4%.

We notice that we would have observed similar effects when using the literature's objective function on optimal appointment policies, if we would have considered other framework different from the one proposed in Ho and Lau (1992) (for example, another set of feasible appointment rules).

It is important to mention that the model developed in this paper involves several parameters that require estimation when applying it to practice. Depending on the nature of the parameter, different methodologies are available in the literature. To estimate the unit cost of waiting time for a client, a, standard techniques in transportation economics can be used, which relate the parameter to customers' salaries. The percentage of absenteeism, γ, can be estimated using historical data collected in the service company. To estimate N, the size of the potential demand, we can assume that all available slots could be filled if the price and the waiting time were zero. Therefore N could be equal to the length of the service day divided by the average service time.

Finally, for an estimation of the reservation price distribution (random variable R), methodologies from marketing can be used (based, for example, on surveys and historical data). Regarding the latter, it is interesting to note that we performed numerical experiments showing that, for unimodal distributions, the optimal appointment rules depend only on the mean and standard deviation of the reservation price.

5. Conclusions and Recommendations for Future Research

The problem of appointment policy has been widely addressed in literature, by studying the performance of different appointment policies under a variety of scenarios. Studies have come up with recommendations that have made it possible to improve the performance of appointment-based systems and have shown that the use of quantitative tools in this area can be extremely useful. We believe the work of Ho and Lau (1992) and Yang, Lau, and Quek (1998) are particularly interesting as they study different appointment policies under various scenarios. Moreover, Yang, Lau, and Quek (1998) construct a general appointment rule that can be used under various scenarios without the need for complicated additional calculations or simulations. It would be important to extend their work to a larger number of scenarios and consider other variables such as a larger number of servers or walk-ins.

On the other hand, we believe it is essential that assumptions about the behavior of demand be made more realistic: demand must depend on waiting time. The objective function used

should be appropriate to the case, taking into account whether the server is a central planner or a private server. It should consider the benefits and costs involved.

The objective functions used in the literature correctly represent the situation of a central planner facing a demand that does not react to waiting time. A few decades ago this situation was the context of many services, especially medical services, most of which were state-owned monopolies. Nowadays, however, the vast majority of services are private (including medical services), and they face more competitive environments with customers who do not want to wait; this fact needs to be taken into consideration in the models. The objective functions used in the literature are only shortcuts for the real objective functions that must be used in these scenarios. In these cases a single parameter (the ratio between the weights of expected server idle time—or completion time—and expected customer waiting time in the objective function) has to include a series of effects given by the customers' behavior fundamental parameters and structural model's parameters. This is not a trivial task and if this single parameter is not correctly estimated, the decision maker can have important losses in utility.

A new model that fits well in these more general scenarios is proposed in this paper. The model can be used in other managerial frameworks where other decisions related to service operations are made. For example, with this model it would also be possible to address the more general problem in which the firm has to jointly decide on its appointment policy and the price to charge for the service. In addition we could analyze capacity decisions, such as overbooking policies.

Although more realistic models that include these aspects are more difficult to solve than those traditionally used in literature, we believe they could be very useful, especially in studies of a more theoretical nature. In this type of study it is essential to use models that are as close to reality as possible, so as to obtain results with greater validity and gain correct insights to the problem.[1]

[1] The authors thank Fondecyt (Chile) Grant 1010457 for financial support.

References

BABES, M. AND G. SARMA (1991), "Out-patient Queues at the Ibn-Rochd Health Care," *Journal of the Operational Research Society,* 42, 10, 845–855.

BAILEY, N. (1952), "A Study of Queues and Appointment Systems in Hospital Outpatient Departments with Special Reference to Waiting-Times," *Journal of the Royal Statistical Society,* 14, 2, 185–199.

BAKER, J. AND M. CAMERON (1996), "The Effects of the Service Environment on Affect and Consumer Perception of Waiting Time: An Integrative Review and Research Propositions," *Journal of the Academy of Marketing Science,* 24, 4, 338–349.

BENNETT, J. AND D. J. WORTHINGTON (1998), "An Example of a Good but Partially Succesful OR Engagement: Improving Outpatient Clinic Operations," *Interfaces,* 28, 5, 56–69.

BOULDING, W., A. KALRA, R. STAELIN, AND V. ZEITHAML (1993), "A Dynamic Process Model of Service Quality: From Expectations to Behavioral Intentions," *Journal of Marketing Research,* 30, 1, 7–27.

BRAHAMI, M. AND D. J. WORTHINGTON (1991), "Queuing Models for Outpatient Appointment Systems—A Case Study," *Journal of the Operational Research Society,* 42, 9, 733–746.

CHARNETSKY, J. R. (1984), "Scheduling Operating Room Surgical Procedure with Early and Late Completion Penalty Costs," *Journal of Operations Management,* 5, 1, 91–102.

COX, T., J. BIRCHALL, AND H. WONG (1985), "Optimising the Queuing System for an Ear, Nose and Throat Outpatient Clinic," *Journal of Applied Statistics,* 12, 2, 113–126.

EL MERCURIO (2000), "Metas y Plazos Exige Lagos a los Ministros," March 14, Santiago, Chile.

FRIES, B. E. AND V. P. MARATHE (1981), "Determination of Optimal Variable-Sized Multiple-Block Appointment Systems," *Operations Research,* 29, 2, 324–345.

GEISZLER, C. (1981), "A Numerical Procedure for the Selection of the Constant Interarrival Time to a Single-Server Queue," *Computers and Mathematics with Applications,* 7, 6, 537–546.

HASSIN, R. (1995), "Decentralized Regulation of a Queue," *Management Science,* 41, 1, 163–173.

HEALY, K. (1992), "Scheduling Arrivals to a Stochastic Service System Mechanism," *Queueing Systems,* 12, 3–4, 257–272.

HO, C. AND H. LAU (1992), "Minimizing Total Cost in Scheduling Outpatient Appointments," *Management Science,* 38, 12, 1750–1764.

——— AND ——— (1999), "Evaluating the Impact of Operating Conditions on the Performance of Appointment Scheduling Rules in Service Systems," *European Journal of Operational Research,* 112, 3, 542–553.

JACKSON, R. R. P., J. D. WELCH, AND J. FRY (1964), "Appointment Systems in Hospitals and General Practice," *Operational Research Quarterly,* 15, 3, 219–237.

JANSSON, B. (1966), "Choosing a Good Appointment System—A Study of Queues of the Type (D, M, 1)," *Operations Research,* 14, 2, 292–312.

KATZ, K., B. LARSON, AND R. LARSON (1991), "Prescription for the Waiting-in-Line Blues: Entertain, Enlighten and Engage," *Sloan Management Review,* 32, 2, 44–53.

KLASSEN, K. AND T. ROHLEDER (1996), "Scheduling Outpatient Appointments in a Dynamic Environment," *Journal of Operations Management,* 14, 2, 83–101.

LAW, A. AND D. KELTON (2000), "Simulation Modeling and Analysis," McGraw-Hill, New York.

LECLERC, F., B. SCHMITT, AND L. DUBÉ (1995), "Waiting Time and Decision Making: Is Time Like Money?," *Journal of Consumer Research,* 22, 1, 110–119.

LIAO, C., D. PEGDEN, AND M. ROSENSHINE (1993), "Planning Timely Arrivals to a Stochastic Production or Service System," *IIE Transactions,* 25, 5, 63–73.

LIU, L. AND X. LIU (1998a), "Dynamic and Static Job Allocation for Multi-Server Systems," *IIE Transactions,* 30, 9, 845–854.

——— AND ——— (1998b), "Block Appointment Systems for Outpatient Clinics with Multiple Doctors," *Journal of the Operational Research Society,* 49, 12, 1254–1259.

MENDELSON, H. AND S. WHANG (1990), "Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue," *Operations Research,* 38, 5, 870–883.

MERCER, A. (1973), "Queues with Scheduled Arrivals. A Correction, Simplification and Extention," *Journal of the Royal Statistical Society,* B35, 1, 104–116.

MUTH, J. (1961), "Rational Expectations and the Theory of Price Movements," *Econometrica,* 29, 3, 315–335.

O'KEEFE, R. M. (1985), "Investigating Outpatient Departments: Implementable Policies and Qualitative Approaches," *Journal of the Operational Research Society,* 36, 8, 705–712.

PEGDEN, C. AND M. ROSENSHINE (1990), "Scheduling Arrivals to Queues," *Computers and Operations Research,* 17, 4, 343–348.

PRESS, W., B. FLANNERY, S. TEUKOLSKY, AND W. VETTERLING (1990), "Numerical Recipes in C, The Art of Scientific Computing," Cambridge University Press, New York.

RISING, E. J., R. BARON AND B. AVERILL (1973), "A System Analysis of a University Health Service Outpatient Clinic," *Operations Research,* 21, 5, 1030–1047.

ROHLEDER, T. AND K. KLASSEN (2000), "Using Client-Variance Information to Improve Dynamic Appointment Scheduling Performance," *Omega,* 28, 3, 293–302.

SABRIA, F. AND C. F. DAGANZO (1989), "Approximate Expressions for Queuing Systems with Scheduled Arrivals and Established Order," *Transportation Science,* 23, 3, 159–165.

SORIANO, A. (1966), "Comparison of Two Scheduling Systems," *Operations Research,* 14, 3, 388–397.

STEIN, W. AND M. COTE (1994), "Scheduling Arrivals to a Queue," *Computers and Operations Research,* 22, 8, 607–614.

STENBACKA, R. AND M. M. TOMBAK (1995), "Time-Based Competition and the Privatization of Services," *The Journal of Industrial Economics,* 63, 4, 435–455.

TAYLOR, S. (1994), "Waiting for Service: the Relationship Between Delays and the Evaluation of Service," *Journal of Marketing,* 58, 2, 56–69.

The Economist (2000), "Government and the Internet: the Next Revolution," June 24.

VANDEN BOSCH, P. M., D. C. DIETZ, AND J. R. SIMEONI (1999), "Scheduling Client Arrivals to a Stochastic Service System," *Naval Research Logistics,* 46, 3, 549–559.

VISSERS, J. AND J. WIJNGAARD (1979), "The Outpatient Appointment System: Design of a Simulation Study," *European Journal of Operational Research,* 3, 6, 459–463.

WANG, P. (1993), "Static and Dynamic Scheduling of Costumer Arrivals to a Single-Server System," *Naval Research Logistics,* 40, 3, 345–360.

——— (1997), "Optimally Scheduling N Client Arrival Times for a Single-Server System," *Computers and Operations Research,* 24, 8, 703–716.

——— (1999), "Sequencing and Scheduling N Customers for a Stochastic Server," *European Journal of Operational Research,* 119, 3, 729–738.

WEISS, E. N. (1990), "Models for Determining Estimated Start Times and Case Ordering in Hospital Operating Rooms," *IIE Transactions,* 22, 2, 143–150.

YANG, K. K., M. L. LAU, AND S. A. QUEK (1998), "A New Appointment Rule for a Single-Server, Multiple-Client Service System," *Naval Research Logistics,* 45, 3, 313–326.

Susana V. Mondschein is an Associate Professor of Operations Management at the Industrial Engineering Department at the University of Chile and a Visiting Professor at the Yale School of Management. Her research interests include the study of optimal decision making under uncertainty in production and service operations. She has worked on applications to pollution control, retailing, catalog sales, hotel reservations, appointment systems, and urban transportation.

Gabriel Y. Weintraub is a second year Ph.D. student at the Management Science and Engineering Department, Stanford University. He studied Industrial Engineering in University of Chile. After graduating he worked in the same engineering department for two years as a lecturer and doing research. He was also a member of the team that was awarded with the IFORS for Operational Research in Development Prize for the implementation of the auction that assigns Chile's school system catering contracts. His research interests include dynamic programming, stochastic modeling and computational game theory.