

An Appointment Overbooking Model to Improve Client Access and Provider Productivity

Linda R. LaGanga
Director of Quality Systems
Mental Health Center of Denver
4141 East Dickenson Place
Denver, CO 80222
Linda.LaGanga@Colorado.edu
Phone: 303.504.6665
FAX: 303-757-5245

Stephen R. Lawrence †
Associate Professor of Operations Management
Leeds School of Business
University of Colorado at Boulder
419 UCB
Boulder, CO 80309-0419
Stephen.Lawrence@Colorado.edu
Phone: 303.492.4351
FAX: 303-492-5962

† Corresponding author.

An Appointment Overbooking Model to Improve Client Access and Provider Productivity

ABSTRACT

Yield and revenue management have been extensively investigated for transportation and hospitality industries, but there has been relatively little study of these topics for appointment services where customers are scheduled to arrive at prearranged times. Such settings included health care clinics; law offices and clinics; government offices; retail services such as tax preparation, auto repair, and salons; counseling centers; and admissions offices, among many others. The problem of *no-shows* (customers who do not arrive for scheduled appointments) is significant for appointment services, with reported no-show rates varying widely from 3 to 80%. No-shows reduce revenues and provider productivity, increase costs, and limit the ability of the provider to service its customer population by reducing effective capacity. In this paper, we develop an analytic appointment scheduling model that balances the benefits of increased revenues and service with the expected costs of customer waiting and provider overtime. Our results demonstrate that effective appointment overbooking can significantly improve customer service and operations revenues while balancing the potential costs of customer waiting and server overtime.

Keywords: Appointment scheduling, overbooking, service operations, scheduling policies

Introduction

An important class of service operations is where customers or clients schedule appointments with service providers prior to service, as opposed to those operations where customers randomly arrive for immediate service according to their own volition. Appointment services are common in modern economies and range across a wide range of service operations such as health care clinics, law offices, tax preparation stores, personal care salons, auto repair garages, portrait studios, professional consulting, and many others. To generalize these service offerings, we refer to them as *appointment services*.

Appointment services are often plagued by no-shows – clients who make appointments for service but then fail to appear when scheduled. Client no shows cause a decline in the performance of the affected service operation by reducing revenues, preventing other clients from obtaining timely service, decreasing office productivity, and causing fixed resources to stand idle. Appointment overbooking provides one means of mitigating the negative impact of no-shows by booking appointments in excess of available capacity (LaGanga and Lawrence 2007).

To investigate appointment overbooking, we develop an analytic model and employ a heuristic solution methodology to obtain good solutions for a wide range of problem settings. Our results indicate that overbooking can provide substantial benefits for appointment services across a wide range of service environments and costs structures. However, we show that patterns of overbooking vary widely across problems and that it is not possible to draw general conclusions regarding how overbooked schedules should be constructed; each appointment overbooking situation needs to be carefully studied and evaluated in order to obtain the best possible overbooking policy.

The contributions of our research are several. We believe that we are the first to model appointment overbooking as an analytic optimization problem. A novel aspect of our model is the exact calculation of probability vectors of the number of clients waiting for service throughout an office session. We also are the first to investigate quadratic client waiting and overtime costs in the context of appointment scheduling, which is arguably a more realistic representation of service operations practice. Results of our computational experiments serve to integrate the results of prior appointment scheduling research by showing that previously proposed appointment scheduling rules (*e.g.*, double-booking, wave scheduling) are in fact special cases of our more general model.

The remainder of the paper is organized as follows. The next section presents the context and background of appointment scheduling, and includes a review of relevant literature. The following section develops our analytic appointment overbooking model in some detail and presents a heuristic solution procedure that generates good appointment schedules. The fourth section reports the results of a computational study where 180 appointment scheduling problems were created and solved across a wide range of practical problem settings. We conclude with a summary of our results and ideas for future research.

Background

Appointment scheduling has been formally studied for more than half a century, much of it addressing issues in healthcare scheduling. Bailey (1952) and Welch and Bailey (1952) established the importance of developing effective appointment scheduling systems to protect the interests of healthcare customers (*i.e.*, patients), who, in previous delivery systems, were all told to arrive at the start of the provider's workday. While such scheduling policies minimized the

idle time of providers, customers were forced to bear the costs of inconvenience and long waits for providers.

More recent research in healthcare service scheduling considers additional complexities in appointment scheduling, such as varying levels of service time variability, fluctuating demand loads, and variable-interval schedule rules (Ho and Lau, 1992; Rohleder and Klassen, 2002). Other work evaluates block schedules that schedule more than one patient into the same appointment time (Blanco White and Pike, 1964; Soriano, 1966; Fries & Marathe, 1981). In these approaches, providers may build up an “inventory” of extra clients to reduce the expected time the provider waits for clients to arrive, but they do not explicitly overbook to attempt to mitigate the lost productivity caused by patients or customers who fail to show up for appointments. Several clinical appointment scheduling researchers, such as Vissers (1979), and Blanco White and Pike (1964), make general recommendations about how to overbook an appointment schedule that are useful in initiating analysis of overbooked systems. However, they do not attempt to test the performance of overbooking across a wide range of possible no-show rates, such as the no-show range of 3-80% reported by Rust, Gallups, Clark, Jones, and Wilcox (1995).

Revenue management, extensively in the transportation industry, typically includes extensive analysis of overbooking policies to balance the benefits of increased service capacity utilization versus overbooking costs such as customer dissatisfaction and compensation (Rothstein, 1971; Smith, Leimkuhler, & Darrow, 1992; Hillier and Lieberman, 2001). Whereas overbooking is useful to minimize the number of assets that perish unused because of no-shows (Weatherford & Bodily, 1992; Toh & Raven, 2003), appointment overbooking is very different from transportation services overbooking, because appointment no-shows are spread over time,

while transportation no-shows all occur at a single point in time. This difference in problem structure requires rather different solution approaches to the problem of mitigating the effects of no-shows (LaGanga & Lawrence, 2007b).

Operations management and statistical methods have been used in previous appointment scheduling studies that measure performance as the weighted sum of patient wait time and provider idle-time costs (Bailey, 1952; Welch & Bailey, 1952; Ho & Lau, 1992). These articles include no-shows as a significant factor in schedule performance and measure some of their effects, but do not focus on how to handle no-shows or reduce their negative impact in the scheduling system. Out of 36 articles categorized in a review of outpatient scheduling literature by Cayirli and Veral (2003), only 11 include the possibility of no-shows and only four include policies to mitigate the effects of no-show behavior.

Articles that consider no-shows include Blanco White and Pike (1964), who focus on the punctuality of patient arrivals as it impacts clinic performance, but only consider patient no-shows on a limited basis; Fetter and Thompson (1966), who investigate the impact of walk-ins as a counterbalance to no-shows, but do not consider appointment overbooking; Vissers and Wijngaard (1979), who adjust the mean and variance of service times to compensate for both no-shows and walk-ins, but do not directly study overbooking; and Vissers (1979), who recommends that the interval between scheduled appointments be reduced to compensate for no-shows, but provides no analysis or data to support this recommendation. In more recent work, LaGanga and Lawrence (2007a) use simulation analysis to develop and test the performance of scheduling rules that are designed specifically to accommodate excess overbooked appointments.

In summary, while research into appointment scheduling has been ongoing for some time, much of it has been *ad hoc* or has been based on simulation studies of appointment

scheduling systems. There has been relatively little work to develop analytic models of appointment scheduling, and little investigation of the potential benefits of overbooking as a means of mitigating the negative effects of client no-shows.

Appointment Scheduling Model

We model an appointment services operation that works with clients or customers on an appointment basis during a service session of duration D time units (see Table 1 for notation). A service session is a period of time (*e.g.*, a morning, a day) during which the office is in continuous operation, after which operations cease for a second period of time (*e.g.*, a lunch break, end of day, etc.). Each appointment during a session is of fixed duration d and the length of a session is designed to be an integer multiple of d so that $D=Nd$, where N is the number of appointment “slots” scheduled in a session. Without loss of generality, we set the duration of each appointment to duration $d=1$ time unit so that the length of an office session is $D=N$.

Clients are scheduled to arrive at the start of an appointment slot and are assumed to be punctual as is often the case in practice (Soriano 1966). However, some clients do not appear for their appointments (are no-shows) with frequency ρ ($0 \leq \rho \leq 1$). Since clients must either arrive for an appointment or are no-shows, the show rate is $\sigma = 1 - \rho$ ($0 \leq \sigma \leq 1$). Given show rate σ , the office may choose to overbook additional appointments during an office session so that the total number of appointments scheduled S is greater than or equal to the number of appointment slots ($S \geq N$).

Further assumptions of our model are that while an office may have multiple service providers, appointments are made for individual servers who do not share clients so that each server can be considered in isolation from the rest of the office. Clients are seen in the order that they are scheduled and arriving clients neither “jump the queue” nor preempt clients currently

being served. If clients remain waiting for service at the end of an office session, the office will work overtime until all clients have been served – clients are not sent away without service.

A schedule \mathbf{S} for an office session is a vector of the number of clients s_j scheduled to arrive at the start of each appointment slot j ($1 \leq j \leq N$). Schedule \mathbf{S} is feasible if it meets all side constraints that might be imposed by a particular office (*e.g.*, limits on the number of clients waiting due to a capacity-constrained waiting room, or limits on the number of clients scheduled for a single appointment slot due to capacity constraints at the check-in counter, *etc.*). The problem of the office is to create a feasible schedule \mathbf{S} of client appointments for each office session so that utility is maximized, as defined below.

Probability of k Clients Waiting

Central to our model of appointment scheduling is the calculation of the probability of the number of clients waiting for service at the start of an appointment slot, including new arrivals. We assume that the number of arriving clients a_j for slot j is binomially distributed and is a function of the number of patients scheduled s_j and their show rate σ . Let α_{jk} be the probability that k clients actually arrive in slot j . As shown in Appendix 1, the probability $\theta_{j+1,k}$ of k clients waiting at the start of period $j+1$ can be found using the recursive relationship:

$$\theta_{j+1,k} = \theta_{j,0} \alpha_{j+1,k} + \sum_{i=0}^{k-1} \theta_{j,i+1} \alpha_{j+1,k-i} \quad (1.1)$$

The first term of this series is the joint probability that there are no waiting clients at the start of slot j and that k clients arrive at the start of the slot $j+1$. Subsequent terms represent the joint probabilities of some clients waiting from the previous slot and new arrivals in the current period, such that the number of waiting and arriving clients sum to k . If $k > 0$, then one client will be serviced during slot j ; otherwise $k=0$ and the provider will be idle during that period.

Note that if one or more clients are waiting at the start of slot j , then one client will receive service during the slot, thus reducing the number of waiting clients by one in the subsequent slot. The normal initial condition at the start of an office session is that no clients are waiting for service prior to arrivals for the first appointment slot, and the normal terminal condition is that no additional clients are scheduled after the final appointment slot N of the office session.

Service Utility Objective Function

We adopt a general objective function for the appointment overbooking problem that works to balance the interests of the office with those of clients and of service providers (LaGanga and Lawrence 2007). This objective trades off the benefit of servicing additional clients with the costs or penalties for keeping some clients waiting for service plus the expected cost of office overtime incurred when all scheduled clients cannot be seen during an office session. We address each of these elements for the current model.

Service Benefits Accrued. For every client serviced, some net benefit is generated for the office, or for its goals and objectives. This benefit can represent net financial profit (after variable costs), community service delivered (in the case of not-for-profit organizations), accrued goodwill or some combination of these and other benefits. The gross benefit of servicing a client is reduced by variable service costs such as the costs of materials, external services, and other costs that vary directly with the number of clients served. Note that the costs of service providers are generally not included in benefit calculations since these costs are usually sunk and do not vary directly with the scheduling policy employed. The aggregate net benefit $\Pi(\mathbf{S})$ earned by schedule \mathbf{S} is a function of the number of clients A serviced during an office session. In this paper, we assume a linear benefit function:

$$\hat{\Pi}(\mathbf{S}) = \pi \hat{A} = \pi \sigma S \quad (1.2)$$

where π is the marginal net benefit of each client serviced and \hat{A} is the expected number of patients that arrive during an office session given that a total of S patients are assigned to appointment slots during schedule \mathbf{S} .

Client Waiting Costs

When appointments are overbooked during an office session, clients may have to wait for service depending on realized show rates and patterns (LaGanga and Lawrence 2007). However, overbooking extra clients beyond the number of appointment slots in a session is not without penalty. Client waiting costs may include expenses for larger waiting rooms and additional staff, lost client goodwill and reduced satisfaction, lost business, increased future no-show rates (Dyer 2005, Lowes 2005), and other behaviors and expenses that are detrimental to the effective and efficient operation of the office.

In this paper, we employ two waiting cost functions, one linear and the second quadratic. In the first case, waiting costs grow linearly with parameter ω , so that if a client waits t time units before the start of service, a penalty of ωt is incurred. However, linear waiting costs do not capture the familiar phenomena that long waits are disproportionately more annoying and disruptive than are shorter waits (Maister 1984). We represent this disproportionate penalty as a quadratic function where waiting costs grow as the square of client waiting time, so that a penalty of ωt^2 is incurred when a client waits t time units.

Define $\hat{\Omega}(\mathbf{S})$ as the cost function describing expected client waiting penalties incurred for schedule \mathbf{S} . As developed in Appendix 2, linear waiting penalty function $\hat{\Omega}^L(\mathbf{S})$ and

quadratic penalty function $\hat{\Omega}^Q(\mathbf{S})$ across all appointment slots and client waiting probabilities are calculated as

$$\hat{\Omega}^L(\mathbf{S}) = \frac{\omega}{\hat{A}} \left(\sum_{j=1}^N \sum_k k \theta_{jk} + \sum_k \sum_{i=1}^k (i-1) \theta_{N+1,k} \right) \quad (1.3)$$

$$\hat{\Omega}^Q(\mathbf{S}) = \frac{\omega}{\hat{A}} \left(\sum_{j=1}^N \sum_k (2k-1) \theta_{jk} + \sum_k \sum_{i=1}^k (i-1)^2 \theta_{N+1,k} \right) \quad (1.4)$$

Office Overtime Costs

Overbooking appointments can also result in some clients remaining unserved at the end of an office session (LaGanga and Lawrence 2007), causing the office to work overtime and to incur additional expenses. Office overtime costs can include lost client and staff goodwill, overtime wages paid to office staff, increased staff turnover, and other expenses and penalties that are detrimental to the performance of the office. Denote expected overtime costs for schedule \mathbf{S} as $\hat{T}(\mathbf{S})$. Again, we consider both linear overtime costs $\hat{T}^L(\mathbf{S})$ and quadratic costs $\hat{T}^Q(\mathbf{S})$, the latter of which recognizes that the implicit costs of overtime grow disproportionately large with the duration of overtime. Expected office overtime costs are

$$\hat{T}^L(\mathbf{S}) = \tau \sum_k k \theta_{N+1,k} \quad (1.5)$$

$$\hat{T}^Q(\mathbf{S}) = \tau \sum_k k^2 \theta_{N+1,k} \quad (1.6)$$

Details are included in Appendix 2.

Net Office Utility

Combining benefit and cost function equations yields the following linear and quadratic net utility functions:

$$\hat{U}(\mathbf{S}) = \hat{\Pi}(\mathbf{S}) - \hat{\Omega}(\mathbf{S}) - \hat{T}(\mathbf{S}) \quad (1.7)$$

$$\hat{U}^L(\mathbf{S}) = \pi \hat{A} - \frac{\omega}{\hat{A}} \left(\sum_{j=1}^N \sum_k k \theta_{jk} + \sum_k \sum_{i=1}^k (i-1) \theta_{N+1,k} \right) - \tau \sum_k k \theta_{N+1,k} \quad (1.8)$$

$$\hat{U}^Q(\mathbf{S}) = \pi \hat{A} - \frac{\omega}{\hat{A}} \left(\sum_{j=1}^N \sum_k (2k-1) \theta_{jk} + \sum_k \sum_{i=1}^k (i-1)^2 \theta_{N+1,k} \right) - \tau \sum_k k^2 \theta_{N+1,k} \quad (1.9)$$

Note that this formulation allows mixed linear and quadratic cost functions (for example, linear client waiting costs and quadratic overtime costs), as well as second order polynomial functions, but we do not examine such objective functions in this paper in the interests of parsimony.

Problem Solution Techniques

We investigated two methods of finding solutions to the appointment services overbooking problem: dynamic programming and heuristic search.

Dynamic Programming Formulation

The problem of determining an optimal appointment schedule \mathbf{S}^* can be formulated as a dynamic program. Each appointment slot j represents a stage in the dynamic program, and the decision of how many appointments s_j to schedule in appointment slot j is completely determined by state variable \mathbf{W}_j – the vector of probabilities of the number of clients waiting for service at the start of the slot before new client arrivals.

$$\begin{aligned} f_j^*(\mathbf{W}_j) &= \max_{s_j} \left\{ \hat{\Pi}(\hat{A}_j) - \hat{\Omega}(\mathbf{W}_j) + f_{j+1}^*(\mathbf{W}_{j+1}) \right\} \\ &= \max_{s_j} \left\{ \pi \sigma s_j - \omega \sum_{k=0}^K k^n \theta_{jk} + f_{j+1}^*(\mathbf{W}_{j+1}) \right\} \end{aligned} \quad (1.10)$$

where costs for the terminal $J+1$ slot are calculated as:

$$f_{N+1}(\mathbf{W}_{N+1}) = - \left[\sum_{k=0}^K k^n (\tau + \omega) \theta_{N+1,k} \right] \quad (1.11)$$

and where $n=1$ for linear and $n=2$ for quadratic waiting and overtime costs.

While it is straightforward to model the service appointment problem as a dynamic program, it is difficult to solve most problems of practical size using standard recursive techniques. This difficulty arises because the number of feasible states \mathbf{W}_j can be very large indeed. For example, a problem with N appointment slots and for which the maximum number of clients that might be waiting for service is K , the number of feasible states for \mathbf{W}_j in appointment slot N can be as large as K^N . Further, since there are both negative costs and positive benefits involved in the calculation of utility, it is difficult to develop effective bounds on the solution space. Consequently, backward recursion effectively becomes complete enumeration of feasible solutions, which means that only very small problems can be solved to optimality. Because of these difficulties, we did not employ dynamic programming to search for optimal solutions to the appointment overbooking problem.

Heuristic Search

Given the difficulty in finding provably optimal solutions using dynamic programming, we developed a heuristic search algorithm to find good solutions: specifically, a simple gradient search algorithm followed by pairwise interchange. The algorithm starts with a feasible schedule, typically where one client is scheduled to arrive for each appointment slot. The algorithm then proceeds to both add one and subtract one appointment for each appointment slot in turn. The single schedule perturbation (addition or removal of an appointment) that results in the greatest improvement to schedule utility is retained, and the algorithm repeats. When no addition or deletion improves utility, the gradient search algorithm terminates. Following gradient search, a pairwise interchange algorithm was employed to further improve the schedule. Pairs of appointment slots were systematically selected and the numbers of clients scheduled in each we exchanged. During each iteration, all possible pairwise swaps were examined and the

single swap that provided the largest improvement in schedule utility was retained. Pairwise interchange then continues until no swaps improved utility and the algorithm terminated.

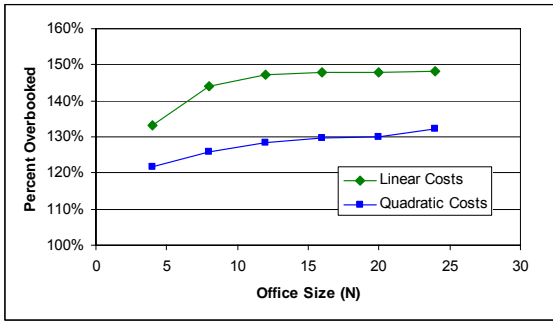
Computational Results

To test our model formulation and solution method, we created a number of representative appointment overbooking problems. These problems spanned a range of office sizes with $N = \{4, 8, 12, 16, 20, 24\}$, which might correspond to a 4-hour office session divided into appointment slots of duration 60, 30, 20, 15, 12, and 10 minutes respectively. Problems with show rates of $\sigma = \{90\%, 80\%, 70\%, 60\%, 50\%\}$ were included. For all problems, the benefit of seeing an additional client was set to $\pi=1$ without loss of generality since it is the ratios of cost and benefit parameters that matter, not their absolute value. Three wait-cost overtime-cost parameter pairs (ω, τ) were examined with values of (1.0, 1.0), (0.5, 1.5), and (1.5, 1.5). These three pairs respectively represent cases where operating costs are comparable to the marginal benefit of additional clients, where costs straddle benefits, and where costs exceed benefits. Note that the case of overtime costs less than marginal benefits is not included since this leads to scheduling large numbers of appointments at the end of an office session – an unrealistic result. In such cases, we anticipate that an appointment services office would simply extend the duration of its operations rather than aggressively overbook its last appointment slot. Finally, all problems were solved using both linear and quadratic objective functions. Taken together, a total of $6 \times 5 \times 3 \times 2 = 180$ problems were examined.

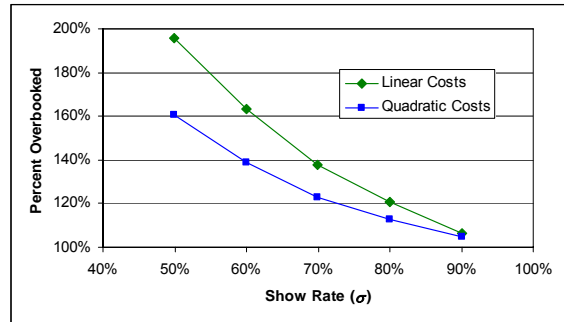
Comparison of Appointment Schedules

The results of our computation experiments demonstrate a number of important patterns regarding appointment overbooking the results of which are summarized in Figures 1 through 6.

Figure 1. Percentage of Overbooked Appointments



1A. Overbooking vs. office size N

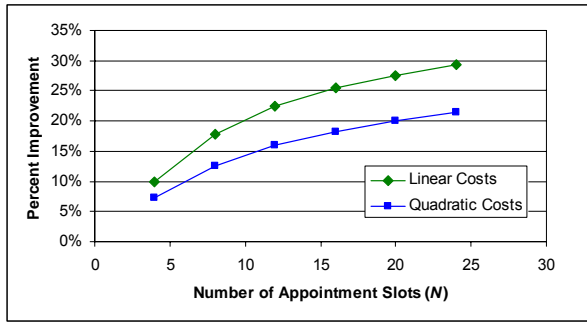


1B. Overbooking vs. show rate σ

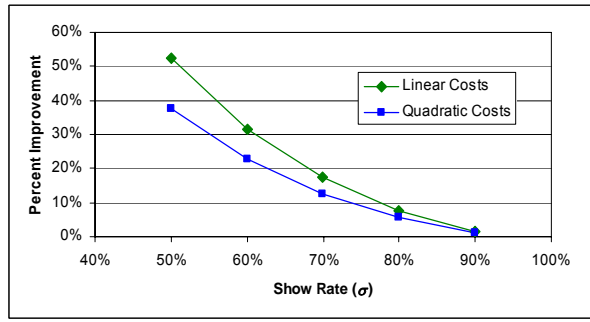
Degree of Overbooking. Figure 1 shows the percentage of overbooked appointments scheduled relative to the number of available appointment slots. Figure 1A demonstrates that overbooking is beneficial over a range of office sizes, and that the percentage of overbooked appointments increases with office size. This occurs because of the risk-pooling characteristics of larger offices, where more appointment slots provide a greater opportunity for overbooking to smooth the random impacts of client no-shows. However, risk-pooling benefits attenuate beyond an office size of $N=12$, so the percentage of overbooked appointments flattens. Figure 1B shows that overbooking increases as the show rate σ declines (or as the no-show rate ρ increases), which is not surprising since the need for overbooking arises because of the occurrence of no shows. Figures 1 and 6A show that overbooking can be beneficial even for small offices with moderate no-show rates.

Utility Gains from Overbooking. Figure 2 shows that appointment overbooking provides net utility gains across the examined range of office sizes and show rates. Overbooking is most beneficial for large offices with high no-show rates, but can provide some positive benefits even for small offices and for low no-show rates. However, our results do indicate that care should be exercised when evaluating small offices or problems with low no-show rates because of the relatively small benefit that overbooking provides.

Figure 2 – Net utility improvement with overbooking



2A. Utility improvement vs. office size N

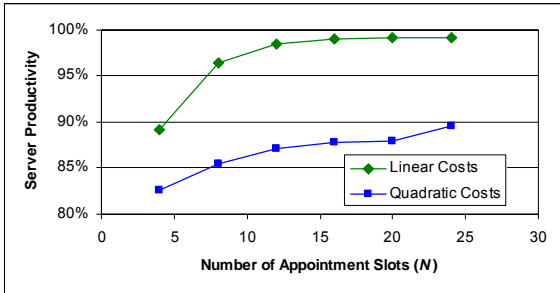


2B. Utility improvement vs. show rate σ

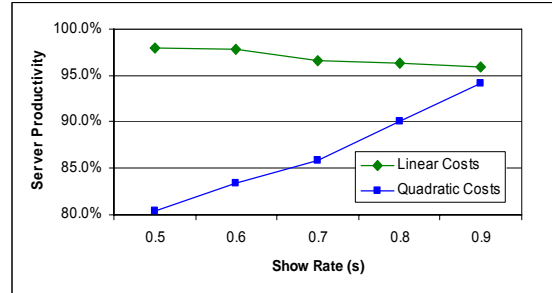
Server Productivity and Client Access. Figures 3A and 3B illustrate provider productivity (utilization) as a function of office size N and show rate σ , respectively. Provider utilization is an important measure of service operations performance and is widely used as an *ad hoc* measure of office effectiveness (Sweeney, 1996). Provider productivity is also a direct proxy for client access – the number of clients serviced relative to the number of appointment slots provides an identical measure of provider productivity. Note that without overbooking, provider utilization is simply equal to the client show rate σ , which should serve as the base reference for interpreting Figure 4. The results illustrated in Figure 3 indicate that appointment overbooking is a powerful tool for dramatically increasing office utilization and for improving client access.

Client Waiting and Office Overtime. The use of appointment overbooking is not without penalties however, as shown in Figure 4. Figures 4A and 4B show how appointment overbooking causes expected client waiting times increase as office size N increases and as show rate σ declines. Similarly, Figures 4C and 4D show how expected office overtime increases with office size and no-show rate. Waiting time and overtime increase with office size and no-show rate because the degree of overbooking increases (Figure 1), which increases the chances for

Figure 3. Server productivity (utilization).
Without overbooking, provider productivity is equal to the show rate σ .



3A. Productivity vs. office size N



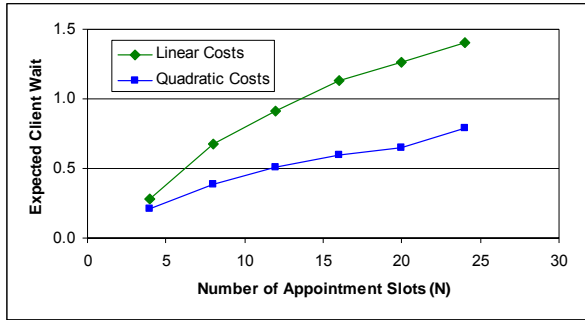
3B. Productivity vs. show rate σ

queues to build and for un-served clients to be waiting at the end of a session (LaGanga and Lawrence 2007).

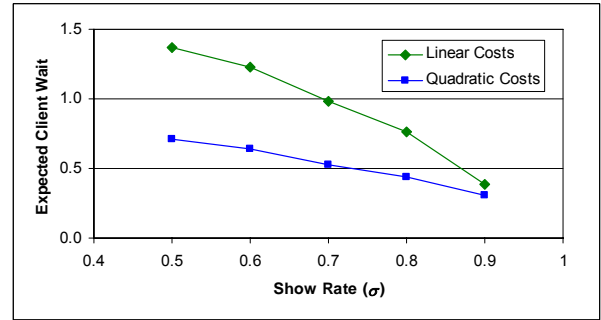
Linear vs. Quadratic Cost Functions. Figures 1-4 also compare the outcomes of overbooking using both linear and quadratic costs functions. In all cases, the use of quadratic costs tends to suppress the impact of overbooking. Relative to linear cost results, quadratic costs reduce the number of overbooked appointments (Figure 1), reduce patient waiting and office overtime (Figure 4), reduce provider productivity and client access (Figure 3), and attenuates net utility from overbooking (Figure 2). However, Figure 2 also demonstrates that appointment overbooking continues to provide substantial net utility gains even with the use of quadratic costs. The use of linear or quadratic cost functions will depend on the characteristics of a particular office, but Figure 2 shows that overbooking can provide significant benefits regardless of the cost function chosen.

Front Loading. Figure 5 shows the prevalence of overbooked appointments by schedule quartile and cost structure across all problems examined. A dominant pattern shown in the figure is the prevalence of front loading the first quartile with overbooked appointments regardless of cost structure. Front loading serves to “prime the pump” for an office session, insuring that there

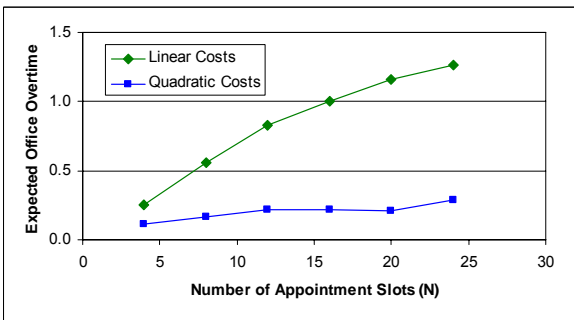
Figure 4. Expected client waiting time and expected clinic overtime. Times are expressed in multiples of the duration of a single appointment.



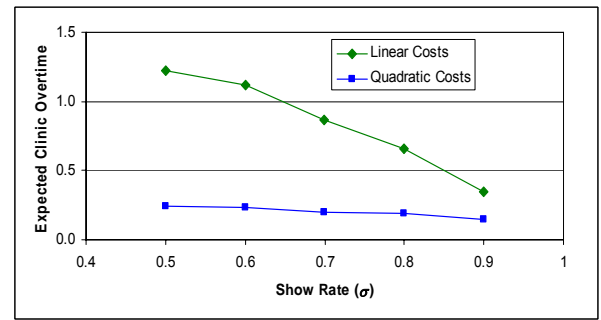
4A. Expected waiting vs. office size N



4B. Expected waiting vs. show rate σ



4C. Expected overtime vs. office size N

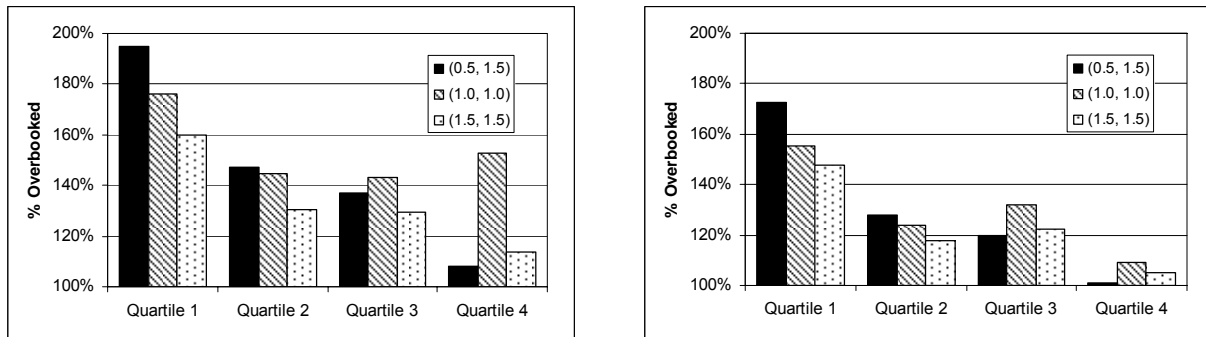


4D. Expected overtime vs. show rate σ

is a sufficient reservoir of waiting clients throughout the session. The number of overbooked appointments declines in each quarter except in the case of moderate linear overtime costs ($\tau = 1.0$), where the number of overbooked appointments increases significantly in the fourth quarter. This occurs because the cost of overtime and the benefit of additional clients are comparable, making it advantageous to schedule additional clients toward the end of a session despite the likelihood that they will need to be served using overtime. In contrast, fourth quarter overbooking is almost zero for quadratic costs regardless of cost structure since the cost of overtime rapidly escalates with the length of the overtime queue.

Appointment Patterns. Finally, Figure 6 illustrates several distinctive patterns that are representative of the schedules generated from our problem set. Figure 6A shows a schedule in which appointments are front-loaded in the service session, with fewer appointments scheduled

Figure 5. Overbooking by office session quartile for several cost patterns (ω , τ).



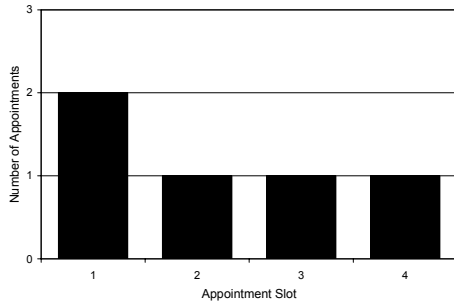
5A. Linear costs

5B. Quadratic costs

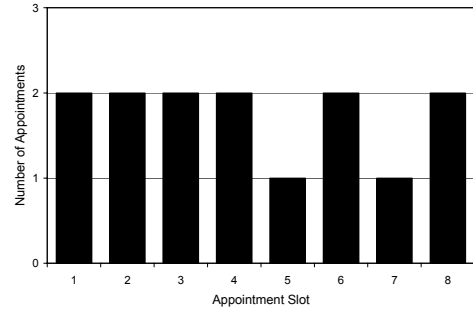
in subsequent sessions. This pattern is similar to the scheduling policy recommended by Bailey (1952) in his early appointment scheduling paper. Figure 6B shows an appointment schedule where two appointments are scheduled for several appointment slots, with other sessions scheduled with only one appointment. This pattern is identical to the “double-booking” suggested by Welch and Bailey (1952). Figure 6C illustrates an appointment schedule in which multiple clients are booked periodically during the office session with periods between where fewer clients are booked. This pattern is suggestive of the “wave scheduling” policy previously proposed by Baum (2001). Figure 6D shows a schedule with both front-loading and double booking, and 6E shows a pattern similar to wave-scheduling, but where the number of slots between waves increases throughout the session. Finally, Figure 6E illustrates a schedule that combines front-loading, double-booking, and erratic wave scheduling that does not fit any scheduling policy previously proposed in the literature.

Figure 6 serves to reconcile our model to scheduling policies previously reported in the literature by other researchers and practitioners. These previously proposed scheduling rules can be interpreted as special cases of our more general model. While these rules may provide excellent results for some office sizes, show rates, and cost structures, none of them will work

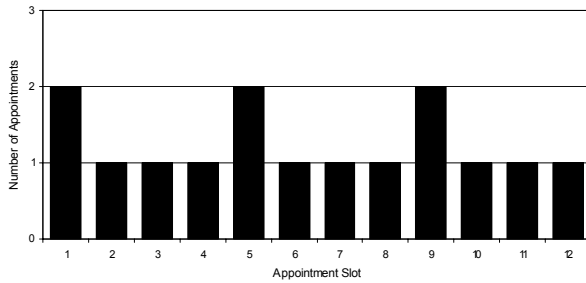
Figure 6 – Various Appointment Overbooking Patterns



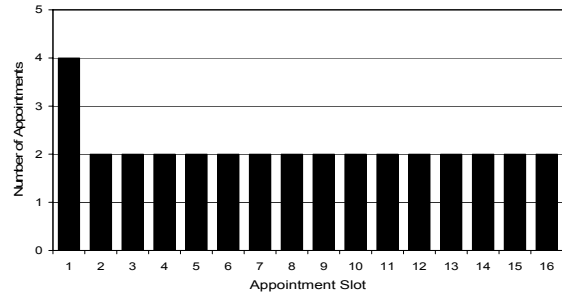
6A. $N=4$, $\sigma=0.8$, $(\omega, \tau) = (0.5, 1.5)$ quadratic



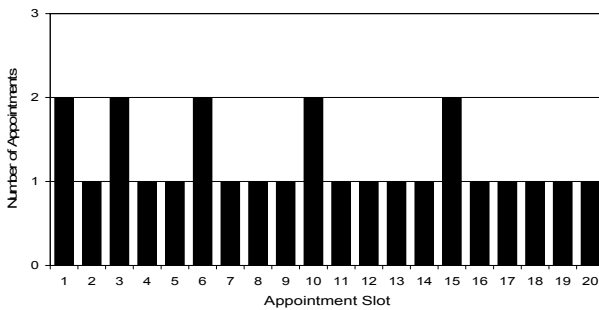
6B. $N=8$, $\sigma=0.5$, $(\omega, \tau) = (1.0, 1.0)$ linear



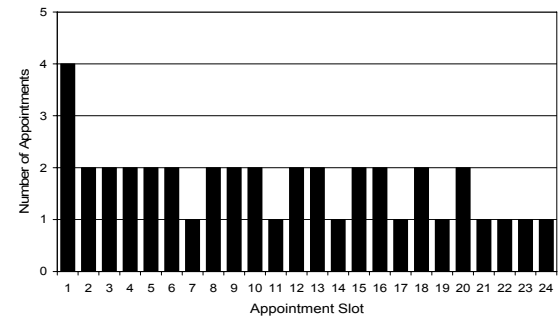
6C. $N=12$, $\sigma=0.7$, $(\omega, \tau) = (1.0, 1.0)$ quadratic



6D. $N=16$, $\sigma=0.5$, $(\omega, \tau) = 1.0, 1.0)$ linear



6E. $N=20$, $\sigma=0.8$, $(\omega, \tau) = (0.5, 1.5)$ linear



6F. $N=24$, $\sigma=0.5$, $(\omega, \tau) = (0.5, 1.5)$ quadratic

well across a wider range of problem settings. Our results confirm that it is difficult to create general scheduling policies that will work well across many problem settings (Ho and Lau 1992) and suggest that solutions obtained for specific problem characteristics will often dominate generic scheduling policies.

Summary and Future Directions

In this paper we have investigated the use of appointment overbooking in service operations where clients are seen by appointment. These “appointment services” are ubiquitous in modern economies and range from health care clinics to professional service offices to personal care salons. Appointment services are often plagued by no-shows – clients who make appointments for service but then fail to appear when scheduled. Client no shows cause a decline in the performance of the affected service operation by reducing revenues, preventing other clients from obtaining timely service, decreasing office productivity, and causing fixed resources to stand idle. Appointment overbooking provides one means of mitigating the negative impact of no-shows by booking appointments in excess of available capacity.

We developed an analytic model of appointment overbooking and employed a heuristic solution methodology to obtain good solutions for a wide range of problem settings. Our results indicate that overbooking can provide substantial benefits for appointment services across a wide range of service environments and costs structures. However, we show that patterns of overbooking vary widely across problems and that it is not possible to draw general conclusions regarding how overbooked schedules should be constructed; each appointment overbooking situation needs to be carefully studied and evaluated in order to obtain the best possible overbooking policy.

The research makes several contributions. First, we model appointment overbooking as an analytic optimization problem and provide an exact calculation of the probability vector of the number of clients waiting for service throughout an office session. We introduce quadratic client waiting and overtime costs in the context of appointment scheduling, a more realistic representation of service operations practice. Computational experiments serve to integrate our

results with prior appointment scheduling research and show that previously proposed appointment scheduling rules (*e.g.*, double-booking, wave scheduling) are in fact special cases of our more general model.

While the complexity of the appointment overbooking problem makes the identification of optimal solutions difficult, we developed a fast and effective heuristic solution procedure that provides good appointment schedules for comparison and evaluation. Our results are easily applied in practice since our model and its solution procedure can be implemented using common spreadsheet software with scripting, and require minimal computational resources. Critical to the successful use of our model is the identification of cost parameters for client waiting (w) and office overtime (τ), and the selection of either linear or quadratic representations for both.

This paper suggests several potentially fruitful avenues for future research. First, our current model assumes that client service durations are fixed and are equal to the duration of an appointment. While this assumption is appropriate for many appointment services, there are many others where service durations are moderately to highly stochastic. We are currently working to extend our model to accommodate uncertain service times. A second interesting extension of our research would be to incorporate walk-in traffic. Many service operations accommodate both clients with appointments and walk-in clients raising important questions about the appropriate mix of the two. Finally, we assume in this paper that service providers do not share clients. While this assumption is appropriate in many offices, other offices send arriving clients to any available provider. Our model might be extended to address these types of service operations as well.

References

- Bailey, N. T. (1952). A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. *Journal of the Royal Statistical Society, Series B, 14(2)*, 185-199.
- Barnhart, C., Belobaba, P., & Odoni, A. R. (2003). Applications of operations research in the air transport industry. *Transportation Science, 37(4)*, 368-391.
- Baum, N. H. (2001). Control your scheduling to ensure patient satisfaction. *Urology Times, 29(3)*, 38-43.
- Cayirli, T., & Veral, E. (2003). Outpatient scheduling in health care: A review of literature. *Production and Operations Management, 12(4)*, 519-549.
- Chung, M. K. (2002). Tuning up your patient schedule. *Family Practice Management, 9(1)*, 41-48.
- Dyer, O. (2005). Sick of getting stood up? No-shows say it's because they need a little respect. *National Review of Medicine, 2(1)*, 1/15.
- Fetter, R. B., & Thompson, J. D. (1966). Patients' waiting time and doctors' idle time in the outpatient setting. *Health Services Research, 1(1)*, 66-90.
- Fries, B. E., & Marathe, V. P. (1981). Determination of optimal variable-sized multiple-block appointment systems. *Operations Research, 29(2)*, 324-345.
- Hillier, F.S. & Lieberman, G.J. (2001). *Introduction to operations research (7th ed.)* New York, NY: McGraw-Hill.
- Ho, C., & Lau, H. (1992). Minimizing total cost in scheduling outpatient appointments. *Management Science, 38(12)*, 1750-1763.
- LaGanga, L. R. (2006). *An examination of clinical appointment scheduling with no-shows and overbooking*. Doctoral dissertation, University of Colorado, Boulder, CO.
- LaGanga, L.R., & Lawrence, S.R. (2007a). Appointment scheduling with overbooking to mitigate productivity loss from no-shows. Conference proceedings of Decision Sciences Institute Annual Conference, Phoenix, Arizona, November 17-20, 2007.
- LaGanga, L. R. & Lawrence, S. R. (2007b). Clinic overbooking to improve patient access and increase provider productivity. *Decision Sciences, 38(2)*.

- Lieberman, W. (2004). Revenue management trends and opportunities. *Journal of Revenue and Pricing Management*, 4(1), 91-99.
- Lieberman, W. (2005). How times have changed for the revenue management professional! *Journal of Revenue and Pricing Management*, 4(20), 109-110.
- Lowes, R. (2005). Practice pointers: How to handle no-shows. *Medical Economics*, 82(8), 62-65.
- Maister, D. (1984). *The Psychology of Waiting Lines*. HBS Teaching Note #9-684-064. Harvard Business School Press.
- Rohleder, T. R., & Klassen, K. J. (2002). Rolling horizon appointment scheduling: A simulation study. *Health Care Management Science*, 5(3), 201-209.
- Rothstein, M. (1971). An airline overbooking model. *Transportation Science*, 5(2), 180-192.
- Rust, C.T., Gallups, N.H., Clark, W.S., Jones, D.S., & Wilcox, W.D. (1995). Patient appointment failures in pediatric resident continuity clinics. *Archives of Pediatrics & Adolescent Medicine*, 149(6), 693-695.
- Smith, B.C., Leimkuhler, J.T., & Darrow, R.M. (1992). Yield management at American Airlines. *Interfaces*, 22(1), 8-31.
- Soriano, A. (1966). Comparison of two scheduling systems. *Operations Research*, 14, 388-397.
- Sweeney, D.R. (1996). Your office: A lot of things will have to change. *Medical Economics*, 73(7), 97-102.
- Toh, R. S., & Raven, P. (2003). Perishable asset revenue management: Integrated Internet marketing strategies for the airlines. *Transportation Journal*, 42(4), 30-44.
- Van Ryzin, G. J., & Talluri, K. T. (2003). Revenue management. In R.W. Hall (Ed.), *Handbook of transportation science*. Boston, MA: Kluwer Academic Publishers, 599-659.
- Vissers, J., & Wijngaard, J. (1979). The outpatient appointment system: Design of a simulation study. *European Journal of Operational Research*, 3(6), 459-463.
- Weatherford, L. R., & Bodily, S. E. (1992). A taxonomy and research overview of perishable-asset revenue management. *Operations Research*, 40(5), 831-844.
- Welch, J. D., & Bailey, N. T. (1952). Appointment systems in hospital outpatient departments. *The Lancet*, May 31, 1105-1108.

APPENDICES

Appendix 1 – Derivation of Number Waiting

By assumption, the number of clients a_j that show for slot j is binomially distributed:

$$f(a_j; s_j, \sigma) = \binom{s_j}{a_j} \sigma^{a_j} (1-\sigma)^{s_j-a_j} = \frac{s_j!}{a_j!(s_j-a_j)!} \sigma^{a_j} (1-\sigma)^{s_j-a_j} \quad (1.12)$$

where s_j is the number of clients are scheduled for an appointment slot and σ is their show rate.

At the end of an appointment slot, there are $k \geq 0$ clients that remain un-serviced and who are waiting for service in subsequent slots. Define $\alpha_{jk} = f(k; s_j, \sigma)$ as the probability that k clients arrive for service in slot j given that s_j clients were scheduled for the slot and that the show rate is σ . Define θ_{jk} as the probability of k clients waiting at the start of period j after the arrival of scheduled clients. Then the number of clients waiting for service at the start of slot $j+1$ can be found using the recursive relationship:

$$\theta_{j+1,k} = \theta_{j,0} \alpha_{j+1,k} + \theta_{j,1} \alpha_{j+1,k} + \theta_{j,2} \alpha_{j+1,k-1} + \dots + \theta_{j,k+1} \alpha_{j+1,0} \quad (1.13)$$

Note that if there are one or more clients waiting in the prior slot, then one of them will be serviced during that slot and will leave the system. Each term represents a combination that results in k clients waiting at the start of slot $j+1$. The first two terms represent the joint probabilities that there were no clients waiting at the conclusion of the prior slot (either because there were no clients waiting or there was only one in queue that was serviced) and that k clients arrive for service in the current slot. The third term is the joint probability that that there were two clients waiting in the prior term and $k-1$ clients arrive in the current slot. The series continues to the last term which represents $k+1$ clients waiting at the start of the prior slot and no clients arriving in the current slot. Collecting terms provides the desired recursion:

$$\theta_{j+1,k} = \theta_{j,0} \alpha_{j+1,k} + \sum_{i=0}^k \theta_{j,i+1} \alpha_{j+1,k-i} \quad (1.14)$$

Appendix 2 – Derivation of Expected Costs

Client Waiting Costs

Define $\hat{\Omega}(\mathbf{S})$ as the cost function describing expected client waiting penalties incurred for schedule \mathbf{S} . We derive two forms of the waiting costs function: a linear waiting penalty function $\hat{\Omega}^L(\mathbf{S})$ and a quadratic penalty function $\hat{\Omega}^Q(\mathbf{S})$. Consider k clients waiting for service at the start of appointment slot j after client arrivals.

Linear Waiting Costs. In the linear case, for each client in queue that is not serviced in slot j , the marginal waiting time for possible service in slot $j+1$ will be one time unit. The probability of k clients waiting is θ_{jk} . Summing across all possible realizations of k and across all N appointment slots gives the total expected waiting time for all clients as $\sum_{j=1}^N \sum_k k \theta_{jk}$. At the

conclusion of the last slot j , there remains the possibility that un-served clients remain. The first client in queue will be served immediately and so incurs no further waiting time. The second client in queue will wait 1 time unit for service; the third will wait 2 time units; and so forth. The aggregate expected waiting time for clients waiting for service at the end of the office session is therefore $\sum_k \sum_{i=1}^k i \theta_{N+1,k}$. If \hat{A} is the expected number of clients that arrive in a clinic session and ω

is the per time unit penalty for client waiting, then the expected linear waiting penalty per arriving client is

$$\hat{\Omega}^L(\mathbf{S}) = \frac{\omega}{\hat{A}} \left(\sum_{j=1}^N \sum_k k \theta_{jk} + \sum_k \sum_{i=1}^k (i-1) \theta_{N+1,k} \right) \quad (1.15)$$

Quadratic Waiting Costs. In the quadratic case, the penalty for client waiting increases as the square of the wait; *i.e.* the penalty for waiting is t^2 if t represents the time a client waits. The penalty for waiting 1 time unit is of course 1, the *additional* penalty for waiting a second time unit is 2²-1=3, and in general, the marginal penalty for waiting an additional time unit after waiting $t-1$ units is proportional to $t^2 - (t-1)^2 = 2t-1$. We can make use of this relationship to write an expression for the total expected quadratic waiting penalty for all clients cross all possible realizations of k and across all N appointment slots as $\sum_{j=1}^N \sum_k (2k-1)\theta_{jk}$. Note that this expression calculates waiting penalties in reverse order as a client progresses through a queue. If the client arrives and is number 3 in line for service, the client's quadratic penalty for the first time period will be calculated as 5, for the second time period as 3, and for the third time period as 1. This is of course in the reverse order that the penalties are incurred, but since clients are seen strictly in order of arrival and since addition is commutative, the order of calculation is immaterial. For clients left waiting at the end of a clinic session, the waiting penalty incurred will simply be the square of their expected waiting time, and the aggregate expected waiting penalty at the end of the office session will be $\sum_k \sum_{i=1}^k i^2 \theta_{N+1,k}$. Putting these two terms together provides the expected quadratic waiting penalty per arriving client:

$$\hat{\Omega}^Q(\mathbf{S}) = \frac{\omega}{\hat{A}} \left(\sum_{j=1}^N \sum_k (2k-1)\theta_{jk} + \sum_k \sum_{i=1}^k (i-1)^2 \theta_{N+1,k} \right) \quad (1.16)$$

Clinic Overtime Costs

The development of clinic overtime costs proceeds along lines similar to the calculations for client waiting costs. Denote expected overtime costs for schedule \mathbf{S} as $\hat{T}(\mathbf{S})$. For linear overtime

costs, the expected overtime that the clinic will work is $\sum_k k\theta_{N+1,k}$, the number of clients waiting for service after final appointment slot N . If the marginal cost of additional overtime is τ , then summing across all possible realizations of k provides the expected linear overtime costs of schedule \mathbf{S} is:

$$\hat{T}^L(\mathbf{S}) = \tau \sum_k k\theta_{N+1,k} \quad (1.17)$$

For quadratic overtime costs, expected quadratic overtime that the clinic will work is $\sum_k k^2\theta_{N+1,k}$,

and the expected quadratic overtime cost of schedule \mathbf{S} is:

$$\hat{T}^Q(\mathbf{S}) = \tau \sum_k k^2\theta_{N+1,k} \quad (1.18)$$

TABLES

Table 1: Notation

\hat{A}	Expected number of arriving clients
$D = Nd$	Duration of a session
d	Duration of an appointment (deterministic)
S	Number of appointments scheduled for a session ($S \geq N$)
N	Number of appointment slots in a session
$\rho = 1 - \sigma$	No-show rate of scheduled appointments ($0 \leq \rho \leq 1$)
$\sigma = 1 - \rho$	Show rate of scheduled appointments ($0 \leq \rho \leq 1$)
π	Marginal net benefit of one additional client
τ	Marginal cost or penalty of clinic overtime ($F > C$)
ω	Marginal cost or penalty of client wait time (per client)
s_j	Number of clients scheduled for service in slot j
a_j	Number of scheduled clients that actually arrive in slot j
w_{jk}	Probability that k clients remain waiting for service at the end of appointment slot j
α_{jk}	Probability that k clients arrive in appointment slot j
θ_{jk}	Probability that k clients are ready for service at the start of slot j after new client arrivals
$\Pi(\cdot)$	Net benefit function
$\Omega(\cdot)$	Client waiting cost function
$T(\cdot)$	Office overtime function
$U(\cdot)$	Utility function, where $U(\cdot) = \Pi(\cdot) - \Omega(\cdot) - T(\cdot)$
\mathbf{W}_j	Vector of probabilities of the number of clients waiting for service at the start of slot j before new client arrivals