ABSTRACT

MINOR, MARY CHRISTINA. Stochastic Programming Models for Appointment Scheduling that Compensate for Variation in Server Behavior. (Under the direction of Dr. B. T. Denton.)

A large amount of pressure is placed on the health care industry to minimize costs, and one facet to cutting health care spending is increased efficiency. Appointment scheduling systems can be used to create better and more cost effective ways to match patient demand with health care provider supply. Previous research has focused on queuing methods, the use of simulation models to test various scheduling rules, and optimization models. In this thesis we focus on server behavior and how it may affect appointment scheduling. Three server behaviors were selected based on their basic nature and relevance to the health care setting: learning, fatigue, and congestion response. The three server effects were modeled separately using three different models. The produced learning and fatigue effect models are both two stage stochastic linear programs (2-SLP). The effect of congestion is modeled through a two stage mixed integer program (2-SMIP). Each model is examined for their relevance to producing more effective schedules, and various sensitivities of the models are also examined. Then finally we produced a set of heuristics for scheduling that are based on each of the three server behavior models, and the effectiveness of each heuristic is examined.
Stochastic Programming Models for Appointment Scheduling that
Compensate for Variation in Server Behavior

by
Mary Christina Minor

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Industrial Engineering

Raleigh, North Carolina

2012

APPROVED BY:

Dr. R. E. King
Dr. A. Vila-Parrish

Dr. B. T. Denton
Chair of Advisory Committee
Mary Christina Minor was born on Nov 17, 1987 in Statesville, NC. Mary, who was typically called Christy by her friends and family, grew up under the guidance of her mother Mary and older brother Mason. Her interest in mathematics grew from a young age and was encouraged by her brother’s mutual interest. Christy received her Bachelor of Science in Mathematics from Elon University in 2010. She then decided to continue her education at North Carolina State University in Industrial and Systems Engineering. After the completion of her Masters degree, Christy will begin a career in Information Technology Consulting.
ACKNOWLEDGEMENTS

My achievements up to this point would not have been possible without the support of many people in my life. Thank you to my mom for her unflagging support throughout my life, and for dealing with my numerous stressed out phone calls. Thank you to my brother for always pushing me and never giving me the answers to my homework problems. Thank you to Austin for keeping me motivated and relaxed over these last 3 semesters. Thank you to my friends for their smiles and encouraging words.

This thesis would not have been possible without the help, support, and patience of my advisor Dr. Brian T. Denton. The good advice and friendship of my committee members Dr. Russell King and Dr. Anita Vila-Parrish has been invaluable to me both academically and personally. I would also like to thank Bjorn Berg for his creation of the congestion model used in this research, and the time that he took to explain the model to me.
# Table of Contents

List of Tables .................................................................................. v

List of Figures .................................................................................. vi

Chapter 1 Introduction ..................................................................... 1

Chapter 2 Literature Review ............................................................. 4
  2.1 Introduction .............................................................................. 4
  2.2 Appointment Scheduling ............................................................ 5
  2.3 Server Behavior ....................................................................... 9
    2.3.1 Learning and Fatigue ......................................................... 10
    2.3.2 Congestion ................................................................. 11
  2.4 Contributions of this Thesis ....................................................... 13

Chapter 3 Appointment Scheduling in the Presence of Varying Server Behavior 14
  3.1 Introduction .............................................................................. 14
  3.2 Model Formulations ................................................................. 15
    3.2.1 Learning and Fatigue Formulation .................................... 15
    3.2.2 Congestion Formulation .................................................. 17

Chapter 4 Numerical Results ............................................................ 26
  4.1 Introduction .............................................................................. 26
  4.2 Methodology .......................................................................... 26
  4.3 Model Validation .................................................................... 27
  4.4 Server Effect on Total Cost of the System ................................. 28
  4.5 Sensitivity Analysis ................................................................. 30
  4.6 Allotted Time Sensitivity ......................................................... 35
  4.7 Heuristics .............................................................................. 45
    4.7.1 Learning and Fatigue Heuristic ........................................ 45
    4.7.2 Congestion Heuristic ...................................................... 47
  4.8 Conclusions ........................................................................... 48

Chapter 5 Conclusions .................................................................. 50
  5.1 Introduction .............................................................................. 50
  5.2 Conclusions ........................................................................... 50
  5.3 Limitations and Future Work .................................................... 51

References ...................................................................................... 52
Table 3.1 Linear congestion effect values as a function of the number of patients waiting for $\rho_{min} = 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$. ................................. 19

Table 4.1 Comparison of objective values when a learning effect is present and time allowance values do account for a learning effect and when the learning effect is ignored. * indicates statistically significant difference between means. 29

Table 4.2 Comparison of objective values when a fatigue effect is present and time allowance values do account for a fatigue effect and when the fatigue effect is ignored. ........................ 29

Table 4.3 Comparison of objective values when a congestion effect is present and time allowance values do not account for congestion and when congestion is ignored. * indicates statistically significant difference between means. 29

Table 4.4 Mean and 95% confidence interval on the mean optimal objective value of the server learning behavior model when $\alpha = 1$. ................................. 32

Table 4.5 Mean and 95% confidence interval on the mean optimal objective value of the server learning behavior model when $\alpha = 0.1$. ................................. 32

Table 4.6 Mean and 95% confidence interval on the mean optimal objective value of the server learning behavior system when $\alpha = 10$. ................................. 32

Table 4.7 Mean and 95% confidence interval on the mean optimal objective value of the server fatigue behavior model when $\alpha = 1$. ................................. 33

Table 4.8 Mean and 95% confidence interval on the mean optimal objective value of the server fatigue behavior model when $\alpha = 0.1$. ................................. 33

Table 4.9 Mean and 95% confidence interval on the mean optimal objective value of the server fatigue behavior model when $\alpha = 10$. ................................. 33

Table 4.10 Mean and 95% confidence interval on the mean optimal objective value of the server congestion behavior model when $\alpha = 1$. ................................. 34

Table 4.11 Mean and 95% confidence interval on the mean optimal objective value of the server congestion behavior model when $\alpha = 0.1$. ................................. 34

Table 4.12 Mean and 95% confidence interval on the mean optimal objective value of the server congestion behavior model when $\alpha = 10$. ................................. 35

Table 4.13 Comparison of mean optimal objective values for a learning effect using the 2-SLP learning model, learning heuristic, and ASP model. ................................. 46

Table 4.14 Comparison of mean optimal objective values for a fatigue effect using the 2-SLP fatigue model, fatigue heuristic, and ASP model. ................................. 46

Table 4.15 Comparison of mean optimal objective values for a congestion effect using the 2-SMIP congestion model, congestion heuristic, and ASP model. (+ indicates a statistically significant difference between the mean optimal objective values for the heuristic and ASP, where the difference between the mean optimal objective value of the heuristic and the 2-SMIP is smaller than between the ASP and 2-SMIP.) ................................. 48
LIST OF FIGURES

Figure 3.1  Graph of learning curve effect, $g_{\rho_{\text{min}}}(i)$, for specified $\rho_{\text{min}}$ value.  
Figure 3.2  Graph of fatigue effect, $g_{\rho_{\text{min}}}(i)$, for specified $\rho_{\text{min}}$ value.  
Figure 3.3  Graph of $f_{\rho_{\text{min}}}(j)$ for specified $\rho_{\text{min}}$ value as a function of waiting patients, $j$.  

Figure 4.1  Graph of Patient Allotted Times, in hours, with Learning Effect, $\rho_{\text{min}} = 0.9$  
Figure 4.2  Graph of Patient Allotted Times, in hours, with Learning Effect, $\rho_{\text{min}} = 0.8$  
Figure 4.3  Graph of Patient Allotted Times, in hours, with Learning Effect, $\rho_{\text{min}} = 0.7$  
Figure 4.4  Graph of Patient Allotted Times, in hours, with Learning Effect, $\rho_{\text{min}} = 0.6$  
Figure 4.5  Graph of Patient Allotted Times, in hours, with Learning Effect, $\rho_{\text{min}} = 0.5$  
Figure 4.6  Graph of Patient Allotted Times, in hours, with Learning Effect, $\rho_{\text{min}} = 0.55$  
Figure 4.7  Graph of Patient Allotted Times, in hours, with Fatigue Effect, $\rho_{\text{min}} = 0.9$  
Figure 4.8  Graph of Patient Allotted Times, in hours, with Fatigue Effect, $\rho_{\text{min}} = 0.8$  
Figure 4.9  Graph of Patient Allotted Times, in hours, with Fatigue Effect, $\rho_{\text{min}} = 0.7$  
Figure 4.10 Graph of Patient Allotted Times, in hours, with Fatigue Effect, $\rho_{\text{min}} = 0.6$  
Figure 4.11 Graph of Patient Allotted Times, in hours, with Fatigue Effect, $\rho_{\text{min}} = 0.5$  
Figure 4.12 Graph of Patient Allotted Times, in hours, with Congestion Effect, $\rho_{\text{min}} = 0.9$  
Figure 4.13 Graph of Patient Allotted Times, in hours, with Congestion Effect, $\rho_{\text{min}} = 0.8$  
Figure 4.14 Graph of Patient Allotted Times, in hours, with Congestion Effect, $\rho_{\text{min}} = 0.7$  
Figure 4.15 Graph of Patient Allotted Times, in hours, with Congestion Effect, $\rho_{\text{min}} = 0.6$  
Figure 4.16 Graph of Patient Allotted Times, in hours, with Congestion Effect, $\rho_{\text{min}} = 0.5$
Chapter 1

Introduction

Health care expenditures in 2009 were 17.6% of the United State’s gross domestic product, with expenditures related to physician and clinical services, and hospitals growing 4.0% and 5.1%, respectively that same year[5]. It is predicted that national health expenditures will continue to grow 6.1% each year over the next 7 years. Given that a large proportion of the population is reaching retirement age, and will begin to have greater demand for health services, a large amount of pressure is placed on the health care industry to minimize costs. One facet to cutting health care spending is increased efficiency. Appointment scheduling systems provide a means to better match patient demand with health service provider supply. Improvements may lower the cost of care by increasing the number of patients that can be served and increasing utilization of expensive resources. Improvements to appointment scheduling can also lower waiting times, resulting in a better service experience for patients.

When dealing with appointment systems in a health care setting two main groups are involved: patients and health care providers. From a patient perspective, as waiting times increase there can be several resulting negative effects. Patients may become frustrated and leave before their appointment time, or may choose to find a new provider in the future. In both situations a loss in revenue for the health care provider may occur. Another aspect that may lead to negative effects for health care providers is the occurrence of idle time for the health care providers between appointments. Although the provider may not be working with a patient they are still being compensated for their time. Additionally, poor scheduling may lead to overtime in order to finish all scheduled appointments, thus resulting in additional cost to the provider.

Many factors can create waiting time for patients and idle time for providers. One of the main factors is the schedule created at the beginning of the day dictating when patients will arrive for their appointments. The need to create schedules that effectively assign patient ap-
pointments has prompted a large amount of research. Previous research has focused on queuing methods, the use of simulation models to test various scheduling rules, and optimization models. Standard models consider the arrival of patients to a single stochastic server. The previous use of queuing allows for easier computation and thus more complex models, however the models cannot account for unpredictability. The more recent focus on stochastic models accounts for every day randomness although adding complexity to the models proves to be more difficult.

Numerous extensions of appointment scheduling models have been explored. For example, some models consider the added probability that a patient may not attend their appointment (no show), or there may only be a finite amount of waiting space for patients. The majority of extensions of this type have focused on the patient behavior. On the other hand, little research has been done on the development of appointment scheduling models that account for server behavior.

In this thesis we focus on provider behavior and how it may affect appointment scheduling. We propose three models which examine the effect of three different behaviors: learning, fatigue, and congestion response. The three behaviors were selected because of their relevance to the health care setting. For example, a learning curve may exist if the longer a health care provider is working, the shorter each appointment may become, because the provider becomes more efficient at the required processes. Alternatively, the longer a provider works, the longer each appointment may become because they are affected by fatigue as the day continues. Finally, in the third model we assume the more patients a provider sees waiting for their appointment the faster the provider performs the service. These server behaviors are observed in many service systems. Goodale and Tunc[6] examined server behavior affected by learning curves, while Courtois and Georges[2], Gupta[7], Harris[8], and Posner[12] all created models that determine service time based on the current state of customers in the system.

In the models we propose, the learning and fatigue effects were considered to be similar in nature. In each case the service rate was modeled as changing based on the number of patients that had arrived for their appointment. The learning effect produced shorter mean service time with each successive patient, and the fatigue a longer mean service time. The congestion effect was modeled with a different approach, being determined by how many patients were currently waiting for their appointment. In the model we propose, the more patients that are waiting for their appointment the shorter the current patient’s mean service time.

The three behaviors were considered separately using three different models. The learning and fatigue effect models are both two stage stochastic linear programs (2-SLP). The effect
of congestion is modeled through a two stage mixed integer program (2-SMIP). Each of the models are examined for their relevance to producing more effective schedules, and various sensitivities are also examined. Heuristics for scheduling are also produced based on the three server behavior models. These heuristics may produce accurate schedules without the use of complicated models, therefore creating a fast cost effective method for scheduling.

The remainder of this thesis is organized as follows. Chapter 2 presents a literature review of appointment scheduling. Chapter 3 describes the 2-SLP model formulations for learning and fatigue and the 2-SMIP formulation accounting for congestion. Chapter 4 presents results from numerical experiments performed to evaluate the structure of the optimal schedules under the three proposed types of server behavior. Finally Chapter 5 discusses conclusions reached through the modeling and testing processes, and future research opportunities associated with this work.
Chapter 2

Literature Review

2.1 Introduction

The scheduling of patient appointments in a health care setting, for example a primary care office or hospital operating room, is a difficult task that requires several different considerations. There are a variety of factors that can influence how quickly a patient is seen by a physician or how long a procedure or consultation may last. One factor that affects patient appointment scheduling is that in many environments, such as outpatient settings, patients are only available at or following their scheduled appointment time. If the health care provider (e.g. physician, operating room) is prepared for the patient prior to their scheduled appointment time, then there is an incurred idling time for the health care server until the patient arrives for their appointment. Thus when scheduling patients for the server’s benefit, it is more advantageous to have patients arrive earlier in order to reduce idling time for the health care provider. However, consideration of the potential for increased patient waiting time must be balanced with the cost of health care provider idle time.

Over the past few decades there has been a large amount of research dedicated to effective methods for scheduling. In this literature review we will discuss some of the most common models and methods. The scheduling methods discussed in this literature review are not all directed at a health care setting, but they are applicable to many contexts. Since the work in this thesis is motivated by problems that arise in a health care context the terms customer and patient are used interchangeably. Additionally the terms server and health care provider are used equivalently to indicate resources that provide a service to customers. A server may indicate a particular individual, such as a physician, or a physical location, for example an operating room. Finally, this literature review will also discuss in detail a previously proposed stochastic programming approach that serves as a basis for the extensions discussed in this thesis.
2.2 Appointment Scheduling

There have been several studies of models that seek to more effectively schedule appointments for medical environments. In one study done by Ho and Lau[10], the authors examined the application of various appointment scheduling rules for creating a schedule in an outpatient environment. The authors evaluated each of the rule’s ability to decrease waiting time and server idle time. Some of the rules examined were variations on a previous set of work by Welch and Bailey [14]. These rules focused on the arrival time, \(a_i\), for each patient \(i\) and which patients would have \(a_i = 0\) and how the remaining patient arrival times were determined. For example, one rule sets \(a_1 = a_2 = 0\) and for \(i > 2\), \(a_i = a_{i-1} + \mu\) where \(\mu\) is the mean service rate. Another rule assigns \(a_1 = 0\), \(a_2 = 0.3\), \(a_3 = 0.6\), \(a_4 = 0.9\), and for \(i > 4\), \(a_i = a_{i-1} + \mu\). The scheduling rules, or heuristics, were tested using a Monte Carlo simulation. Through their work Ho and Lau found that an effective scheduling rule could be identified only when three main environmental factors were known. These factors were the probability of patients being a no show, the coefficient of variation of service time, and the total number of patients to be scheduled. Additionally, Ho and Lau suggested that the ratio of cost for patient wait time to provider idle time is an important factor to consider.

Another approach to appointment scheduling that has been taken, is based on optimization models in which the objective is to minimize costs associated with expected patient waiting and server idling. Weiss[13] examined two different aspects of this type of model: assigning patient arrival times based on a given schedule, and designing a sequence of arrivals for a given set of patients. He studies the problem in the context of scheduling procedures in an operating room. Weiss approaches the problem by showing that the problem resembles that of the well-known newsvendor model. Given a set of \(n\) procedures which are already placed in order, each procedure has an associated set of characteristics:

- \(z_i\): random processing time for \(i\)th procedure
- \(l_i\): completion time of \(i\)th procedure
- \(a_i\): estimated start (arrival) time of \(i\)th procedure
- \(p_i\): actual start time of \(i\)th procedure

Weiss first makes the assumption that no procedure can begin before its estimated, or assigned, start time. Therefore if procedure \(i - 1\) ends early or on time then:
\[ l_{i-1} \leq a_i \Rightarrow p_i = a_i \] (2.1)

where if \( l_{i-1} \) is strictly less than \( a_i \) then this would result in idle time for the health care staff. Conversely if \( i - 1 \) ends on time or later then:

\[ l_{i-1} \geq a_i \Rightarrow p_i = l_{i-1} \] (2.2)

and if \( l_{i-1} \) is strictly greater than \( a_i \) then this results in a waiting time for both the \( i \)th patient and their associated staff. In Weiss’ model the total wait (\( w_i \)) and idle (\( s_i \)) time, respectively, can be represented as:

\[ w_i = \max(0, p_i - a_i) = \max(0, l_{i-1} - a_i) \] (2.3)

\[ s_i = \max(0, a_i - l_{i-1}). \] (2.4)

The objective is to minimize the total expected cost, which can be represented as:

\[ E(\text{cost}) = E \left( \sum_i c_w w_i + \sum_i c_s s_i \right). \] (2.5)

Here \( c_w \) and \( c_s \) represent the cost of wait and idle time, respectively.

In Equation (2.5) the cost associated with idle time is analogous to the surplus cost in the newsvendor model, since idling indicates that the actual demand for server time was less than the server’s actual time supply. The cost of waiting corresponds to the newsvendor’s cost of a shortage, because waiting occurs when the actual demand for server time is greater than the actual supply of time.

Weiss additionally studied a model in which the sequence of patients may vary. This model first determined the patient sequence through an equation examining the cost differences of varying sequences. Then each patient’s arrival time was determined using the properties of the previously mentioned newsvendor model. The combination of these procedures provided a method for finding an effective schedule with assigned patient waiting times that minimizes cost and balances patient wait and server idle time.
Another study on appointment scheduling by Bosch and Dietz[1] views the appointment scheduling problem as a first-come first-serve queuing system. In this model they follow the same premise as previously mentioned models, where they seek to minimize the cost of customer waiting time and server overtime by assigning customer arrivals. However, there are two additional considerations, the probability of customer no shows and the requirement to evenly space customer arrivals. As with previous models Bosch and Dietz[1] determine a customer’s service time by considering it to be a random variable from a chosen positive distribution. However, another probability is also associated with customers, their potential to fail to show for an appointment. The central goal of this model is to find a set of customer arrival times that must occur at evenly spaced intervals.

A different approach to appointment scheduling, taken by Murray and Berwick[11] introduces the concept of advanced access appointment scheduling. They examine a scheduling strategy which allows patients to request an appointment same day, rather than call and make an appointment for another day in the future. Murray and Berwick argue that this form of appointment scheduling would allow medical offices to fulfill daily demand as it arrives rather than postpone the demand until a future date. Theoretically this process would decrease the factors that increase patient and staff idle time. These factors include: time spent by the medical staff booking and reminding patients of their appointments and a patient’s probability of skipping the appointment. Murray and Berwick assert that the predicted advantage of advanced access care depends on three general designs. First, advanced access aims to reduce the gap between when demand first appears and when it is satisfied. Then second, the reduction between demand and supply would additionally lead to a reduced patient demand since the physician has become more accessible. Third the authors suggest, these changes would lead to an increase in the supply of physician appointments on a daily basis, and would thus more effectively handle a sudden increase in demand. However, such a scheduling policy has the limitation of increasing physician idle time, i.e., unused appointment slots.

A more recent generalization of Weiss’ optimization model was proposed by Denton and Gupta[3]. They examined the scheduling of $n$ patients when a patient’s appointment duration is uncertain. Denton and Gupta proposed a two stage stochastic linear program (2-SLP) that seeks to minimize the expected cost of patient waiting time, server idle time, and system overtime. They refer to this model as the appointment scheduling problem (ASP). Their model serves as a basis for the models studied in this thesis. Thus, we describe it in detail below.

The model contains decision variables divided into a first and second stage process which in-
volves a set of $n$ jobs, where patients are assumed to always arrive on time to their appointment. The first stage decision variables are denoted by:

\[ x: \text{vector of job allowances for first } n-1 \text{ jobs} \]

The second stage decision variables are:

\[ w: \text{vector of wait times associated with } x \text{ and } Z \]
\[ s: \text{vector of server idle times associated with } x \text{ and } Z \]
\[ o: \text{system overtime associated with given set of jobs} \]
\[ g: \text{earliness for a given sequence of jobs} \]

The model parameters include:

\[ Z: \text{vector of random job durations} \]
\[ d: \text{day length, total allotted time for job sequence} \]
\[ c^w: \text{vector of cost coefficients for customer waiting time} \]
\[ c^s: \text{vector of cost coefficients for server idle time} \]
\[ c^o: \text{cost coefficient for server overtime} \]

Note that we use bold font to denote vectors and upper case to denote random variables.

The ASP has the following objective function:

\[
\min E_Z \left\{ \sum_{i=2}^{n} c^w_i w_i + \sum_{i=2}^{n} c^s_i s_i + c^o o \right\}
\]

(2.6)

The expectation in (2.6) is with respect to a probability distribution defined by a finite set of scenarios. A scenario, $\omega_k$, is a single set of realizations of job durations where $k$ is the index of a set of $K$ scenarios. Each scenario, $k$, has probability $p^k$ of occurring. The objective function can be expressed as:

\[
\min \sum_{k=1}^{K} p^k \left( \sum_{i=2}^{n} c^w_i u^k_i + \sum_{i=2}^{n} c^s_i s^k_i + c^o o^k \right)
\]

(2.7)

The subscript, $k$, denotes that each variable corresponds to a specific scenario $k$. 


The constraints in the ASP follow the logical argument that the difference between the actual duration of a patient’s appointment \((Z_i)\) and their estimated time allowance \((x_i)\) must be equal to the difference between the following patient’s waiting time \((w_{i+1})\) and the current patient’s waiting time \((w_i)\), minus the server idle time that occurs between patient \(i\) and \(i + 1\), \((s_{i+1})\). This must be true for each patient \(i\), given that \(w_1\) and \(s_1\) are zero since the first patient arrives on time at time zero, creating \(n - 1\) decision variables. Therefore the given objective function, (2.7), is subject to the following constraints:

\[
\begin{align*}
  w_2^k - s_2^k &= Z_1^k - x_1 \\
  -w_2^k + &w_3^k - s_3^k = Z_2^k - x_2 \\
  \quad \ldots \quad \ldots \\
  -w_{n-1}^k + &w_n^k - s_n^k = Z_{n-1}^k - x_{n-1}
\end{align*}
\]

\[x_i \geq 0 \forall i, w_i^k, s_i^k \geq 0 \forall (i, k) \quad (2.8)\]

The final constraint placed on the model restricts the calculation of overtime and earliness:

\[-w_n^k + o_k^k - g_k^k = Z_n^k - d + \sum_{j=1}^{n-1} x_j \quad (2.9)\]

Denton and Gupta studied results based on this model to draw insights about optimal appointment schedules. This model forms the basis for the extensions discussed in Chapters 3 and 4.

### 2.3 Server Behavior

The previous section discussed several different types of appointment scheduling models, each with the goal of determining the amount of time that each patient should be allotted for their appointment. However, each of the discussed methods fails to consider how server behavior may influence appointment scheduling. As discussed in Chapter 1, over the course of a day a server may begin to speed up or slow down their service process due to factors such as learning, fatigue, or a response to congestion. Failing to account for an existing change in service time over the course of a day when modeling appointment schedules may lead to incorrect schedules, and therefore potential increases in cost that could be avoided. In this section we will discuss three
types of server behaviors that could affect appointment scheduling, and some of the literature supporting the selection of these behaviors.

2.3.1 Learning and Fatigue

One environmental factor that could have an effect on servers is related to changes during progression of the day. Two potential effects that may result in this are: learning and fatigue. A learning effect may be present when the server decreases the mean service time as the day continues. Conversely, a server may also be affected by fatigue and increase their mean service time as the day progresses.

Goodale and Tunc[6] produced a mixed integer program that includes service rates that are affected by a server’s learning curve. When modeling the effect of a learning curve, their model used two sets of servers. One set of servers known as the core employees maintained a constant service rate, and the other set, the contingent employees, had varying service rates that were affected by learning or fatigue. Once an optimal schedule was produced by the model the costs associated with each of the employee sets were compared to see if there was an advantage to considering service time variations.

The learning curve employed by Goodale and Tunc was modeled through the following exponential equation:

\[ \frac{f \cdot n^{-b}}{\mu_{ave}} \]  

(2.10)

where \( f \) is the initial service time for contingent employees, \( n \) is the number of customers served by contingent employees, \( b \) is the learning parameter, and \( \mu_{ave} \) is the average service rate of the employee population. Examining the equation reveals that as the number of customers served by a contingent employee increases the fraction, (2.10), also becomes smaller meaning a decrease in service time.

In their research, Goodale and Tunc determined that the schedule produced using a learning curve was more accurate and lowered costs. They also determined that the schedule was more realistic and indicated that server behavior should be considered more in the future. While their research focused on the effect of a learning curve, fatigue was also mentioned on several occasions because of its similarity to learning, and the author’s belief in its relevance to scheduling.
The Goodale and Tunc model uses values that are constant or part of a distribution, and fails to consider how different a set of appointments may be from day to day. Two of the appointment scheduling models produced in this thesis focus on the server behaviors of learning and fatigue and utilize a stochastic model. Our learning and fatigue server behavior models have random service times that more accurately represent appointment schedules in a physician's office.

2.3.2 Congestion

In this section we discuss how waiting room congestion may affect the amount of time that each patient should be allotted for their appointments. A common observation for many service systems is that as a queue for a service grows, the server increases its pace. Some studies suggest this theory can be incorporated into models.

Some studies have attempted to measure the effect of the queue on the server’s pace in practice. In a study by Deveugele, et al.[4] a group of physicians and their patients were observed to determine what factors determine the length of a consultation. They found that a negative correlation existed between the physician’s workload and the patient’s consultation lengths. Another study on a collection of physicians and their patients done by Heaney, et al.[9] had similar results. They also found a negative relationship between queue size and consultation length, and additionally discovered a similar negative relationship between a patient’s placement in the queue and their consultation length.

Some authors have studied models that account for the effects the environment may have on the system. One example is in a paper by Courtois and Georges[2] which extends the standard M/G/1 queuing model. Their queuing model assigns customer arrivals and service times as being dependent on the current state of the system. The state of the system is considered to be the current number of customers in the queue. Courtois and Georges assume that the probability of an arrival at a particular point in time depends on how many customers are waiting at a given point in time. Similarly, the service time for a customer is also determined by the number of customers that are in the queue at the time the customer begins service. The total service time in Courtois and Georges’ model is determined by two different factors: the initial set length of service time (which does not vary), and the amount of extra time for secondary services that the customer requests. The length of secondary services is the value which will vary according the number of customers waiting in line when the customer begins their service.
An article by Gupta[7] considers a queuing system that uses the current state of the system to determine particular distributions. The queuing model produced by Gupta is defined by an input which is hyper-Poisson with an assigned mean arrival rate, a first-come first-served service policy, and an exponential service time distribution. Gupta’s model assigns the arrival and service time mean rates to be state dependent. Therefore, the mean arrival and service time rates change based on arbitrary functions relating the rates to the current state of the system. Gupta also assumes that there is a finite number of patients that can be in the system at any time. The finite consideration not only adds a more realistic aspect to the system but also allows Gupta to obtain solutions for their model.

Another model considering environmental effects was proposed by Harris[8]. Harris’ takes an M/G/1 queuing system and treats each customer’s service time as a stochastic process that is classified by the total number of customers that are in the queue at the time a customer begins service. Each customer’s service time is created by determining the number of customers in the queue at the time each customer begins their service. The number of customers, $i$, at the beginning of service is then used to find the variable, $M_i$, which is part of a predetermined group. The determined $M_i$ is then used to create the service time, $T_i$, for each customer. Harris examines three different queuing models: one that has state independent service times, the second that considers an independent service time distribution for customer 1 and a state dependent distribution for customers 2 through $n$, and a third that has a state dependent service time distribution for all $n$ customers. Harris determined that the use of state dependent processes requires different solving methods, and that most future research would be centered around finding these methods.

Finally, Posner[12] looks at a slightly different approach for an M/M/1 queuing model with state dependent service times for customers. In previous examples the service time has been dependent on the current number of customers in the queue at the start of a customer’s service. Posner considers that a customer’s service time is dependent on the amount of time that the customer has spent in the queue waiting to be served. The approach used by Posner seeks to account for the customer’s needs, based on the amount of time that they have waited. A customer may require a longer service time if they have been waiting longer, or the server may spend a longer time with the customer to counter the negative effect incurred by making the customer wait longer. Posner considers an approach in which the service time for a customer is exponentially distributed with a parameter dependent on the customer’s wait time. The course used here is determined to have a more accessible solution.
2.4 Contributions of this Thesis

This thesis proposes a stochastic model that accounts for the variability in behavior of a server as opposed to using a fixed workload-invariant service time distribution. Although there are some (descriptive) queuing models that address this issue, our model appears to be the first that addresses it from an (prescriptive) optimization perspective. This thesis provides, to our knowledge, the first stochastic programming model formulation of a problem of this nature. Using our model a number of novel results are provided to evaluate the potential influence of server behavior on the optimal design of appointment schedules, and to measure the relative importance of considering server behavior compared to standard models previously proposed in the literature.
Chapter 3

Appointment Scheduling in the Presence of Varying Server Behavior

3.1 Introduction

Adjusting an appointment scheduling model to appropriately account for changes in the server’s behavior over time may improve appointment schedules by reducing expected waiting and/or idle time. Failing to account for changes, on the other hand, could lead to a schedule that over or underestimates the duration of each patient’s appointment. The server behaviors we focus on in this Chapter are: learning, fatigue, and congestion response. In this section we intend to produce our own models that will account for these selected server behaviors.

We propose two 2-SLP models that include the server behaviors of learning and fatigue and one 2-SMIP that includes server behavior affected by congestion. The models are extensions to the ASP model proposed by Denton and Gupta [3]. We explain the assumptions associated with each of the models, and we provide a detailed description of the stochastic programming formulation.

This chapter is organized as follows. The formulation of the 2-SLPs that include learning and fatigue are presented first in Section 3.2.1, followed by the 2-SMIP that accounts for congestion in Section 3.2.2. The methodology used to obtain solutions for these formulations is explained in Section 4.2.
3.2 Model Formulations

We extend the model by Denton and Gupta[3] to include the influence of the three types of server behavior discussed in Chapter 2: learning, fatigue, and congestion. The formulation of the first two server behaviors, learning and fatigue, will be provided first followed by the formulation of the congestion model.

3.2.1 Learning and Fatigue Formulation

In our model we represent the effects of learning and fatigue as a change in the proportion of each customer’s service time. We use the following equations to represent the effects of learning and fatigue, respectively:

\[ g^{L}_{\rho_{\text{min}}}(i) = (1 - \rho_{\text{min}})e^{(1-i)} + \rho_{\text{min}} \] (3.1)

\[ g^{F}_{\rho_{\text{min}}}(i) = (\rho_{\text{min}} - 1)e^{(1-i)} + 1 \] (3.2)

In each equation \( \rho_{\text{min}} \) represents the selected server behavior minimum effect value for learning (L) and fatigue (F), and \( i \) is an index denoting the position of the patient in the appointment sequence. The functions in equations (3.1) and (3.2) are multiplied with service time in each of their respective models. Thus, each behavior effect equation produces the effect of scaling each customer’s random service time. A better understanding of the two equations can be gained by looking at Figures 3.1 and 3.2 which show a graph of the value of each equation by patient number for a set of \( n = 10 \) patients, and several different \( \rho_{\text{min}} \) values. In Figure 3.1 the server learning effect is decreasing, resulting in a decrease in the mean service time, with respect to the patient index. Figure 3.2 shows the server fatigue effect increasing with each additional patient which will result in an increase in the mean service time with respect to patient index.

The extended models for learning and fatigue are created by simply multiplying the respective server behavior equations with the actual job duration time, \( Z_i \). In the case of the learning formulation, the objective remains the same as Equation 2.6 in the ASP model,

\[ \min E \left\{ \sum_{i=2}^{n} c^w_i w_i + \sum_{i=2}^{n} c^s_i s_i + c^o o \right\} \] (3.3)
Figure 3.1: Graph of learning curve effect, $g_{\rho_{\text{min}}}^{L}(i)$, for specified $\rho_{\text{min}}$ value.

Figure 3.2: Graph of fatigue effect, $g_{\rho_{\text{min}}}^{F}(i)$, for specified $\rho_{\text{min}}$ value.
and is now subject to a new set of constraints:

\[
\begin{align*}
  w_2 - s_2 &= Z_1 \cdot g_{\rho_{\text{min}}}^L(1) - x_1 \\
  -w_2 + w_3 - s_3 &= Z_2 \cdot g_{\rho_{\text{min}}}^L(2) - x_2 \\
  \quad \ldots \quad \ldots \\
  -w_{n-1} + w_n - s_n &= Z_{n-1} \cdot g_{\rho_{\text{min}}}^L(n-1) - x_{n-1}
\end{align*}
\]

\[x_i \geq 0, w_i \geq 0, s_i \geq 0 \quad \forall \quad i = 1, \ldots, n\]  
\[\text{(3.4)}\]

The fatigue formulation is very similar, differing only in the server behavior parameter that is used.

\[
\min E \left\{ \sum_{i=2}^{n} c_i^w w_i + \sum_{i=2}^{n} c_i^s s_i + c^o \right\}
\]

\[\text{(3.5)}\]

\[
\begin{align*}
  w_2 - s_2 &= Z_1 \cdot g_{\rho_{\text{min}}}^F(1) - x_1 \\
  -w_2 + w_3 - s_3 &= Z_2 \cdot g_{\rho_{\text{min}}}^F(2) - x_2 \\
  \quad \ldots \quad \ldots \\
  -w_{n-1} + w_n - s_n &= Z_{n-1} \cdot g_{\rho_{\text{min}}}^F(n-1) - x_{n-1}
\end{align*}
\]

\[x_i \geq 0, w_i \geq 0, s_i \geq 0 \quad \forall \quad i = 1, \ldots, n\]  
\[\text{(3.6)}\]

### 3.2.2 Congestion Formulation

The concept of adding learning and fatigue is a simple model extension, but the addition of server behavior in the presence of congestion is more complicated. In our model we assumed that as congestion builds up in the system, the server would begin to work faster, in order to compensate for the growing queue. In our model congestion is defined by the number of patients that are waiting for service. Thus, the model needs to keep track of the current number of patients waiting for their appointment and make the congestion effect a function of this number. The resulting congestion model considers a single server system where patients consistently arrive at their scheduled appointment time, and would then be attended to in the order in which
they arrived. Following is a detailed mathematical formulation of this model.

The following indices were used in the model:

- \( i \): index for patients
- \( j \): index for server queue size (number of patients waiting)
- \( \omega \): index for scenarios

In addition, the following notation, coupled with the indices listed above, defines the model parameters.

- \( n \): number of routine patients
- \( c^w \): cost vector associated with patient waiting time
- \( c^o \): overtime cost coefficient with respect to \( d \)
- \( d \): planned length of clinic day
- \( f_{\rho_{\min}}(j) \): function determining congestion effect produced by \( j \) patients waiting and selected \( \rho_{\min} \) parameter
- \( \rho_{\min} \): minimum congestion parameter
- \( z(\omega) \): vector of random job durations for patients
- \( \xi(\omega) \): random vector containing second stage scenario dependent parameters,
  \[ \xi(\omega) = \{z_1(\omega), ..., z_n(\omega)\} \] where \( n \) is the number of patients, \( z \in \mathbb{R}^n \).

The cost vectors \( c^w \) and \( c^o \) can be adjusted accordingly for each job, depending on how cost is affected by a particular customer or server. The congestion effect \( f_{\rho_{\min}}(j) \) is a function of the number of patients that are currently waiting in the queue and a previously specified value of \( \rho_{\min} \). The function produces a scalar value that is used to find the fraction of the original time the server will take to service a particular patient. In this thesis we use a linear model for congestion, the applied numerical values for five different \( \rho_{\min} \) values are displayed in Table 3.1. Additionally a graphical display of each of the five functions is shown in Figure 3.3.

The vector \( z(\omega) \), which varies according to scenario, has probability distribution \( P \in \mathbb{R}^n \) and support \( \xi \). There are several decision variables used in this model but they are divided as first and second stage. The first stage decision variables are:
Table 3.1: Linear congestion effect values as a function of the number of patients waiting for $\rho_{\text{min}} = 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$.

<table>
<thead>
<tr>
<th>$j$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_1(j)$</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$f_{0.9}(j)$</td>
<td>1.0</td>
<td>0.989</td>
<td>0.978</td>
<td>0.967</td>
<td>0.956</td>
<td>0.944</td>
<td>0.933</td>
<td>0.922</td>
<td>0.911</td>
<td>0.9</td>
</tr>
<tr>
<td>$f_{0.8}(j)$</td>
<td>1.0</td>
<td>0.978</td>
<td>0.956</td>
<td>0.933</td>
<td>0.911</td>
<td>0.889</td>
<td>0.867</td>
<td>0.844</td>
<td>0.822</td>
<td>0.8</td>
</tr>
<tr>
<td>$f_{0.7}(j)$</td>
<td>1.0</td>
<td>0.967</td>
<td>0.933</td>
<td>0.9</td>
<td>0.867</td>
<td>0.833</td>
<td>0.8</td>
<td>0.767</td>
<td>0.733</td>
<td>0.7</td>
</tr>
<tr>
<td>$f_{0.6}(j)$</td>
<td>1.0</td>
<td>0.956</td>
<td>0.911</td>
<td>0.867</td>
<td>0.822</td>
<td>0.778</td>
<td>0.733</td>
<td>0.689</td>
<td>0.644</td>
<td>0.6</td>
</tr>
<tr>
<td>$f_{0.5}(j)$</td>
<td>1.0</td>
<td>0.944</td>
<td>0.889</td>
<td>0.833</td>
<td>0.778</td>
<td>0.722</td>
<td>0.667</td>
<td>0.611</td>
<td>0.556</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Figure 3.3: Graph of $f_{\rho_{\text{min}}}(j)$ for specified $\rho_{\text{min}}$ value as a function of waiting patients, $j$. 
$x_i$: time allotted for patient $i$

$a_i$: decision variable defining the time when patient $i$ arrives

$u_{ii'}$: binary decision variable defining the arrival sequence of patients (1 if $i'$ arrives after $i$ arrives, 0 otherwise)

The second stage decision variables are:

$l_i(\omega)$: decision variable defining the time when patient $i$ leaves

$p_i(\omega)$: decision variable defining the time when patient $i$ begins their procedure

$v_{ii'}(\omega)$: binary decision variable defining which patients arrive before patient $i'$ begins their procedure (1 if $i'$ is arrives before $i$ begins their procedure, 0 otherwise)

$q_{ii'}(\omega)$: binary decision variable defining which patients are waiting when patient $i'$ begins their procedure (1 if $i'$ is waiting when $i$ begins their procedure, 0 otherwise)

$r_i(\omega)$: integer decision variable defining the number of patients waiting when $i$ begins their procedure

$t_{ij}(\omega)$: binary decision variable defining the number of patients waiting when $i$ begins their procedure (1 if $j$ patients are waiting, 0 otherwise)

$\hat{z}_i(\omega)$: decision variable defining the procedure duration for patient $i$ where congestion is included

$w_i(\omega)$: waiting time for patient $i$

$s_i(\omega)$: idling time associated with patient $i$

$o(\omega)$: overtime with respect to the length of clinic day $d$

The objective function that is to be minimized in the second stage, given scenario $\omega$, is:

$$Q(x, \xi(\omega)) = \min \sum_{i=1}^{n} c^w w_i(\omega) + c^o o(\omega).$$ (3.7)

Thus, the first stage problem of the model is:
\[
\begin{align*}
\min & \quad Q(x) \tag{3.8a} \\
\text{s.t.} & \\
& a_i = \sum_{k=1}^{i-1} x_k \quad \forall (i,\omega) \tag{3.8b} \\
& a_{i'} - a_i - Mu_{ii'} \leq 0 \quad \forall (i,i') \tag{3.8c} \\
& x_i, a_i \geq 0, \quad u_{ii'} \in \{0,1\} \tag{3.8d}
\end{align*}
\]

where

\[Q(x) = E_{\omega}[Q(x,\xi(\omega))].\tag{3.9}\]

The first stage problem is then combined with the second stage problem to obtain the complete formulation, which is:

\[
\begin{align*}
\min & \quad E_{\omega} \left[ \sum_{i=1}^{n} c^w w_i(\omega) + c^o o(\omega) \right] \tag{3.10a}
\end{align*}
\]
s.t.

\[ l_i(\omega) - w_i(\omega) - \hat{z}_i(\omega) = \sum_{k=1}^{i-1} x_k \quad \forall (i > 1, \omega) \quad (3.10b) \]

\[ l_1(\omega) - \hat{z}_1(\omega) = 0 \quad \forall (\omega) \quad (3.10c) \]

\[ p_i(\omega) - a_i - w_i(\omega) = 0 \quad \forall (i, \omega) \quad (3.10d) \]

\[ p_i(\omega) - a_{i'} - M v_{ii'}(\omega) \leq 0 \quad \forall (i, i', \omega) \quad (3.10e) \]

\[ 2 q_{ii'}(\omega) - u_{ii'} - v_{ii'}(\omega) \leq 0 \quad \forall (i, i', \omega) \quad (3.10f) \]

\[ r_i(\omega) - \sum_{i'=1}^{i} q_{ii'}(\omega) = 0 \quad \forall (i, \omega) \quad (3.10g) \]

\[ r_i(\omega) - \sum_{j=0}^{n-1} j t_{ij}(\omega) = 0 \quad \forall (i, \omega) \quad (3.10h) \]

\[ \sum_{j=0}^{n-1} t_{ij}(\omega) = 1 \quad \forall (i, \omega) \quad (3.10i) \]

\[ \hat{z}_i - \sum_{j=0}^{n-1} f_{\rho_{\min}}(j) t_{ij}(\omega) z_i(\omega) = 0 \quad \forall (i, \omega) \quad (3.10j) \]

\[ w_i(\omega) - s_i(\omega) - w_{i-1}(\omega) - \hat{z}_{i-1}(\omega) = -x_{i-1} \quad \forall (i > 1, \omega) \quad (3.10k) \]

\[ w_1(\omega) = 0 \quad \forall (\omega) \quad (3.10l) \]

\[ \sum_{i=1}^{n} (\hat{z}_i(\omega) + s_i(\omega)) - o(\omega) \leq d \quad \forall (\omega) \quad (3.10m) \]

\[ l_i(\omega), p_i(\omega), \hat{z}_i(\omega), w_i(\omega), s_i(\omega), o(\omega) \geq 0 \quad \forall (i, \omega) \quad (3.10n) \]

\[ q_{ii'}(\omega), v_{ii'}(\omega), t_{ij}(\omega) \in \{0, 1\} \quad \forall (i, i', j, \omega) \quad (3.10o) \]

\[ r_i(\omega) \in \mathbb{Z}_+ \quad \forall (i, \omega). \quad (3.10p) \]

The equations (3.8b) and (3.10b) through (3.10d) assign arrival time, leaving time and procedure start time for each patient. Constraints (3.8c),(3.10e), and (3.10f) are used to determine which patients arrive in relation to some patient's arrival time, procedure start time, and which patients are waiting when patient begins their procedure. The next three constraints (3.10g) through (3.10i) assign the number of patients that are waiting when some patient begins their procedure. The congestion effect is assigned by constraint (3.10j). Finally constraints (3.10k) through (3.10m) are from the previously discussed Denton and Gupta[3] model, and determine wait, idle, and overtime for each patient and scenario, \( \omega \).
The constraints (3.8b), (3.8c), and (3.10e)-(3.10j) are what define the congestion response of the server. The constraints are broken into two groups for description, (3.8c),(3.10e), and (3.10f), and (3.10g)-(3.10j). The constraints (3.8c), (3.10e), and (3.10f) determine values that are necessary to be known to apply a congestion effect to this model. Constraint (3.8c),

\[ a_{i'} - a_i - Mu_{ii'} \leq 0 \quad \forall (i, i', \omega), \]

is a first stage constraint defined by first stage decision variables \( a_i \) and \( u_{ii'} \). This constraint determines which patients arrive after patient \( i \). When \( u_{ii'} = 0 \) the remaining portion of the equation must be less than or equal to zero, and since the arrival times for patient \( i \) and \( i' \), where \( i \neq i' \), cannot be the same this means it must be less than zero. Therefore, the arrival time for patient \( i, a_i \), must be a larger value or later time than the arrival time for patient \( i' \), \( a_{i'} \). Conversely, when \( u_{ii'} = 1 \), since it is multiplied by some sufficiently large number \( M \), this causes the left side of the equation to be a large negative number, allowing \( a_{i'} \) to be later than \( a_i \). Constraint (3.10e),

\[ p_i(\omega) - a_{i'} - Mv_{ii'}(\omega) \leq 0 \quad \forall (i, i', \omega), \]

determines which patients arrive before patient \( i \) begins their procedure. When \( v_{ii'}(\omega) = 0 \) the remaining portion of the equation must be less than or equal to zero. Therefore, the arrival time for patient \( i' \) must be equal to, or later than, the procedure start time for patient \( i, p_i \), and this would be consistent with \( v_{ii'}(\omega) = 0 \). When \( v_{ii'}(\omega) = 1 \), since it is multiplied by some sufficiently large number, \( M \), this removes the restrictions on the rest of the equation allowing \( a_{i'} \) to occur before \( p_i \). Constraint (3.10f),

\[ 2q_{ii'}(\omega) - u_{ii'} - v_{ii'}(\omega) \leq 0 \quad \forall (i, i', \omega), \]

is a combination of constraints (3.8c) and (3.10e). When \( q_{ii'}(\omega) = 1 \), then because \( q_{ii'}(\omega) \) is multiplied by 2 this forces the two binary variables \( u_{ii'} \) and \( v_{ii'}(\omega) \) to both be equal to 1 to satisfy the inequality. In this case, where \( u_{ii'} \) and \( v_{ii'}(\omega) \) are both equal to 1, this would be
consistent with patient $i'$ waiting when patient $i$ begins their procedure, or $q_{ii'}(\omega) = 1$. Additionally, when $q_{ii'}(\omega) = 0$ the restrictions are released from the remaining variables, allowing patient $i'$ to not be waiting when patient $i$ begins their procedure.

The next group of constraints discussed are (3.10g)-(3.10i). The first two constraints (3.10g) and (3.10h):

\[
\begin{align*}
r_i(\omega) - \sum_{i'=1}^{i} q_{ii'}(\omega) &= 0 \quad \forall (i, \omega) \\
r_i(\omega) - \sum_{j=0}^{n-1} j t_{ij}(\omega) &= 0 \quad \forall (i, \omega)
\end{align*}
\]

both assign the value for the number of patients waiting when patient $i$ begins their procedure. Constraint (3.10g) requires that $r_i(\omega)$ is equal to the sum over the binary variable $q_{ii'}(\omega)$. Constraint (3.10h) requires that $r_i(\omega)$ is equal to the binary variable $t_{ij}(\omega)$, determining whether $j$ patients are waiting when patient $i$ begins their procedure times those $j$ patients. The constraint (3.10i),

\[
\sum_{j=0}^{n-1} t_{ij}(\omega) = 1 \quad \forall (i, \omega)
\]

simply assures that there can only be one value, $j$, of patients waiting when patient $i$ begins their procedure.

These two groups of constraints, (3.8c), (3.10e), and (3.10f), and (3.10g)-(3.10i), determine the congestion effect to be applied, which is defined by constraint (3.10j),

\[
\hat{z}_i - \sum_{j=0}^{n-1} f_{\rho \text{min}}(j) t_{ij}(\omega) z_i(\omega) = 0 \quad \forall (i, \omega).
\]

The decision variable defining procedure duration is determined in this equation. The variable,
\( z_i \), must be equal to the summation over the potential number of patients \( j \), however since the binary variable \( t_{ij} \) is a product in this summation only one value for patient \( i \) with \( j \) patients waiting not equal to zero may be found. Therefore, the congestion effect for patient \( i \) will be equal to the product of the congestion effect for \( j \) patients waiting and the job duration for patient \( i \), \( z_i(\omega) \). The congestion effect function, \( f_{\rho_{\text{min}}}(j) \), will be defined by several different functions in the next chapter. The values are fractional to represent the portion of the previous procedure duration that will be used.

Solving the congestion formulation is more difficult than solving the learning and fatigue formulations. The increased difficulty is a result of the formulation’s additional decision variables and constraints. The congestion formulation uses big M constraints, which could potentially make solving the formulation significantly more difficult. Additionally the model is a MIP, compared to an LP in the learning and fatigue cases.
Chapter 4

Numerical Results

4.1 Introduction

A series of numerical experiments were performed to evaluate the three models described in Chapter 3. The results presented in this Chapter include an evaluation of the structure of the optimal schedule, a sensitivity analysis of the models with respect to input parameters, and a set of three heuristics based on each of the server behavior models. Through this analysis we seek to determine if the models are able to produce significantly more cost effective schedules than standard models that ignore server behavior. We further hope to understand which variables have the most effect on cost, and should therefore be considered in the future.

The models were solved using OPL-CPLEX 12.2. The solutions, on average took from thirty to fifty seconds to be returned on a computer with a 2.83 GHz IntelCore\textsuperscript{TM}2 Quad Processor and 8 GB RAM.

4.2 Methodology

The models were programmed using OPL-CPLEX. The following settings were used throughout Chapter 4 unless noted otherwise: $n = 10$, $c^w = 1$, $c^o = 1$, $c^s = 0$, $d = 4$, $z(\omega) \sim U(0, 1)$. The determination of the learning ($g_{\rho_{\min}}^L$), fatigue ($g_{\rho_{\min}}^F$), and congestion ($f_{\rho_{\min}}$) effect functions, are discussed in Section 4.4. The probability of each scenario occurring was sampled from a uniform distribution, $U(0, 1)$, and there were 1000 scenarios for each model evaluation unless noted otherwise. We used 1000 scenarios because this number is sufficient to generate 95% confidence intervals on the objective function and allotted time that are tight relative to the mean. In Sections 4.4, 4.5, and 4.7 statistical analysis is performed on the optimal objective value based
on the mean and 95% confidence intervals which were computed using 30 independent model runs. Similarly in Section 4.6 an analysis is performed on the allotted time for each patient and again thirty model runs were used to compute the mean and 95% confidence interval for the analysis.

4.3 Model Validation

The validity of each model was determined through comparison to the ASP model proposed by Denton and Gupta[3] (which we refer to as the ASP model in the remainder of this Chapter). We assumed that because the ASP model was previously validated we could verify our models by showing that each model is equivalent to the ASP model. Each of the three behavior models were validated by reducing each model back to the ASP model. Then the same set of values were run through each reduced server behavior model and the ASP model, and then each model was compared back to the base ASP model to conclude that identical solutions were being produced.

In the case of the learning and fatigue server behavior models their similarity with the ASP model allowed for easy verification. Each of the two behavior models were reduced by restricting their respective effect functions to be equal to 1, thus resulting in no effect from learning and fatigue. Then equivalent values for job duration were run through each of the behavior models and the base ASP model while the same cost for overtime, waiting, and idle time were used along with 10 patients and scenarios, and equal clinic day length. Identical solutions were produced for optimal objective value and allotted time between each behavior model and the ASP model, indicating that the learning and fatigue models are only a variation on the previously validated ASP model.

The verification of the congestion model was done by first reducing the model by keeping the congestion effect function equal to 1 for our model, meaning that congestion would have no effect. Identical values for scenario probability and procedure duration were used between each model, and the cost of overtime, waiting, and idle time were held to 1, 1, and 0 respectively. There were 10 scenarios and patients for each model and equal clinic day lengths. The application of these restrictions to each model returned the same objective solutions and allotted patient times, verifying the validity of our congestion model. Therefore the server behavior model affected by congestion is valid based on the Denton and Gupta ASP model.
4.4 Server Effect on Total Cost of the System

The first set of experiments that was run on each of the models sought to determine if failing to account for server behavior has a significant effect on the total cost of the system. In order to conduct this experiment the results from the three server behavior models were compared to the ASP model that does not account for server behavior. The effect of ignoring server behavior when it is present was modeled by manipulating each of the server behavior models such that their solutions were no longer accounting for their respective server behavior, but the service times were still being altered by the server behavior. The resulting server behavior models were then used to determine the allotted time for each patient (appointment schedule) and the mean optimal objective value (total cost) associated with each of these sets of allotted times, then this data was examined to determine if there was an associated loss in profit or time.

The first step of the experiment was conducted by finding and recording the mean optimal objective values corresponding to applying each of the three server behavior models, and then comparing these results to the standard model. Each model uses a server effect function \((g_{\rho_{\text{min}}}^L(i), g_{\rho_{\text{min}}}^F(i), f_{\rho_{\text{min}}}(j))\) that is determined by an initially selected value of \(\rho_{\text{min}}\) and either the current patient number (learning and fatigue models) or the current number of waiting patients (congestion model). The experiment was run on each model with \(n = 10\) patients and \(\rho_{\text{min}} = 0.5, 0.6, 0.7, 0.8, 0.9\). The mean optimal objective function values were recorded for each server behavior model and compared to the standard model.

The second step requires creating a model for each server behavior that does not account for the respective server behavior when producing a schedule. Setting \(\rho_{\text{min}} = 1.0\) for each model results in no server behavior effect. Therefore each model, the 2-SLP learning and fatigue, and the 2-SMIP congestion model were run setting \(\rho_{\text{min}} = 1.0\) and the produced set of allotted times, \(x_i\), for each patient was recorded. The program was then run again for each model and their corresponding effect function, with \(\rho_{\text{min}} = 0.9, 0.8, 0.7, 0.6, 0.5\); however, the allotted times were held to be equal to those previously recorded for the \(\rho_{\text{min}} = 1.0\). Restriction of the time allowance values in this way simulates the effect of using a model that does not account for server behavior to produce a schedule in a situation where server behavior alters patient service times.

The results for the learning effect and fatigue effect are shown in Table 4.1 and Table 4.2, respectively. When \(\rho_{\text{min}} \leq 0.6\) Table 4.1 shows a significant increase between mean objective values when a learning effect is accounted for and when it is ignored, as shown by the *. Since significant differences in the means are only seen for low values of \(\rho_{\text{min}}\), it seems that increases
Table 4.1: Comparison of objective values when a learning effect is present and time allowance values do account for a learning effect and when the learning effect is ignored. * indicates statistically significant difference between means.

<table>
<thead>
<tr>
<th></th>
<th>Optimal Objective Value</th>
<th>Optimal Objective Value Ignoring Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>95% CI</td>
</tr>
<tr>
<td>$g_{L.1}(i)$</td>
<td>8781.44 (8662.21, 8900.68)</td>
<td>-</td>
</tr>
<tr>
<td>$g_{L.0.9}(i)$</td>
<td>6521.35 (6443.09, 6599.61)</td>
<td>6031.39 (5281.96, 6780.81)</td>
</tr>
<tr>
<td>$g_{L.0.8}(i)$</td>
<td>4288.64 (4235.86, 4341.41)</td>
<td>4208.216 (3684.53, 4731.90)</td>
</tr>
<tr>
<td>$g_{L.0.7}(i)$</td>
<td>2469.28 (2441.26, 2497.30)</td>
<td>2791.14 (2443.785, 3138.49)</td>
</tr>
<tr>
<td>$g_{L.0.6}(i)$</td>
<td>1165.82 (1143.59, 1188.04)</td>
<td>1893.42* (1658.62, 2128.21)</td>
</tr>
<tr>
<td>$g_{L.0.5}(i)$</td>
<td>456.45* (448.28, 464.61)</td>
<td>1304.08* (1142.23, 1465.93)</td>
</tr>
</tbody>
</table>

Table 4.2: Comparison of objective values when a fatigue effect is present and time allowance values do account for a fatigue effect and when the fatigue effect is ignored.

<table>
<thead>
<tr>
<th></th>
<th>Optimal Objective Value</th>
<th>Optimal Objective Value Ignoring Fatigue</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>95% CI</td>
</tr>
<tr>
<td>$g_{F.1}(i)$</td>
<td>8781.44 (8662.21, 8900.68)</td>
<td>-</td>
</tr>
<tr>
<td>$g_{F.0.9}(i)$</td>
<td>8300.18 (8216.23, 8384.12)</td>
<td>7602.50 (6657.22, 8547.78)</td>
</tr>
<tr>
<td>$g_{F.0.8}(i)$</td>
<td>7828.05 (7750.73, 7905.38)</td>
<td>7306.17 (6397.70, 8214.65)</td>
</tr>
<tr>
<td>$g_{F.0.7}(i)$</td>
<td>7395.68 (7317.86, 7473.50)</td>
<td>6842.60 (5993.86, 7691.33)</td>
</tr>
<tr>
<td>$g_{F.0.6}(i)$</td>
<td>6960.64 (6892.39, 7028.90)</td>
<td>6518.18 (5704.97, 7331.39)</td>
</tr>
<tr>
<td>$g_{F.0.5}(i)$</td>
<td>6475.96 (6405.86, 6546.06)</td>
<td>6153.33 (5389.32, 6917.34)</td>
</tr>
</tbody>
</table>

Table 4.3: Comparison of objective values when a congestion effect is present and time allowance values do not account for congestion and when congestion is ignored. * indicates statistically significant difference between means.

<table>
<thead>
<tr>
<th></th>
<th>Optimal Objective Value</th>
<th>Optimal Objective Value Ignoring Congestion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>95% CI</td>
</tr>
<tr>
<td>$f_{1}(j)$</td>
<td>9011.72</td>
<td>9011.72</td>
</tr>
<tr>
<td>$f_{0.9}(j)$</td>
<td>7193.90 (7118.42, 7269.37)</td>
<td>7220.61 (7144.46, 7296.77)</td>
</tr>
<tr>
<td>$f_{0.8}(j)$</td>
<td>5635.54 (5574.27, 5696.81)</td>
<td>5721.61 (5664.73, 5778.49)</td>
</tr>
<tr>
<td>$f_{0.7}(j)$</td>
<td>4067.90* (4014.83, 4120.98)</td>
<td>4320.13* (4262.99, 4377.28)</td>
</tr>
<tr>
<td>$f_{0.6}(j)$</td>
<td>2718.07* (2674.98, 2761.17)</td>
<td>3167.03* (3134.69, 3199.19)</td>
</tr>
<tr>
<td>$f_{0.5}(j)$</td>
<td>1610.34* (1584.92, 1635.76)</td>
<td>2263.25* (2228.30, 2298.19)</td>
</tr>
</tbody>
</table>
The results for the fatigue model, in Table 4.2, do not show the same effect as the learning model. There is a decrease between mean costs when fatigue is accounted for and when it is not. Comparing the 95% confidence intervals, however, does not show a significant decrease, indicating that failing to account for fatigue in appointment scheduling may not lead to a significant difference for the specific test cases we explored. We believe that one main reason why no significant difference was found for the fatigue model is due to the fatigue effect function, Equation 3.2. The fatigue effect function, the graph of which can be seen in Figure 3.2, is an increasing exponential function with an upper bound of 1. The speed at which the function increases to 1 causes, for the majority of patients, the fatigue effect to be fairly minimal (i.e. the fatigue effect is generally $\geq 0.9$). Therefore, the majority of service times are only minimally altered by fatigue and it is not surprising that there is no significant difference in the mean optimal objective value.

The results for the congestion model are shown in Table 4.3. Table 4.3 shows an increase in mean optimal objective value between when congestion is considered in the appointment scheduling model and when congestion is ignored. The comparison of the confidence intervals also reveals statistically significant differences, as indicated by the *. In the case of the congestion coefficient function with the $\rho^\text{min}$ values 0.8 and 0.9 the confidence intervals overlap, which would indicate that there is no significant difference in cost between when congestion is considered and when it is not considered. However, for $\rho^\text{min} \leq 0.7$ the differences are statistically significant. This difference indicates that failing to consider congestion when it has an effect on the system may lead to increased costs that could be avoided.

### 4.5 Sensitivity Analysis

The second set of experiments performed on each of the server behavior models investigates the sensitivity of each model to $\alpha$, where $\alpha = \frac{c_w}{c_o}$. This measures the effect of varying ratios of waiting and overtime cost on the optimal objective value. The experiments were carried out by testing three different values $\alpha = 1, 0.1, 10$. These three values were selected to cover all possible scenarios where waiting and overtime cost were either: equal, overtime cost was higher, or wait cost was higher. Each of these three $\alpha$ values were separately applied to each of the three models. As with the first set of experiments the mean and 95% confidence interval on the mean for the optimal objective values were calculated from the results of the 30 runs. The returned
mean optimal objective value, for each associated $\alpha$ and $\rho_{\text{min}}$ value were recorded. The results for the learning server behavior model are shown in Tables 4.4, 4.5, and 4.6, then the fatigue model results are displayed in Tables 4.7, 4.8, and 4.9. Finally the sensitivity of changes in cost of waiting and overtime for the congestion behavior model are shown in Tables 4.10, 4.11, and 4.12. There is a table for all three server behavior models when $\alpha = 1$ (4.4, 4.7, 4.10), meaning $c^w$ and $c^o$ are both equal, this table serves as a comparison to the tables for each server behavior model when $c^o$ and $c^w$ are separately increased. The tables when $\alpha = 1$ help us to determine the effect on the mean optimal objective value when $c^w$ and $c^o$ are increased.

Each Table of different $\alpha$ values for the learning server behavior effect show a significant increase in the mean optimal objective value as $\rho_{\text{min}}$ increases. Table 4.4 when compared to Tables 4.5 and 4.6 shows a significant increase in the mean optimal objective value with respect to $\rho_{\text{min}}$ when $c^w$ and $c^o$ are increased, except for $\rho_{\text{min}} = 0.5$ in Table 4.5. The significantly lower mean optimal objective value for $\alpha = 1$ compared to $\alpha = 0.1$ and $\alpha = 10$ follows intuitively because of an increase in either $c^o$ or $c^w$. The significant decrease in mean optimal objective value for $\rho_{\text{min}} = 0.5$ when $\alpha = 1$ compared to when $\alpha = 0.1$ we estimate stems from the large decrease in allotted time for patients when $\rho_{\text{min}} = 0.5$. The large decrease in allotted time indicates that any increase in overtime cost has minimal consequence because overtime is very seldom required.

Looking at the effect of increasing either $c^o$ or $c^w$ in the learning behavior model. Table 4.5 reveals higher mean optimal objective values compared to Table 4.6 with respect to $\rho_{\text{min}}$. Table 4.5 represents when $c^o$ is increased. The higher optimal objective values in Table 4.5 imply that increasing only $c^o$ will lead to higher optimal objective values than when increasing $c^w$ only. These results indicate that the learning server behavior model may be more sensitive to changes in cost of overtime than the cost of waiting.

The Tables for each $\alpha$ value for the fatigue server behavior effect again show a significant increase in the mean optimal objective value as $\rho_{\text{min}}$ increases. Table 4.7 when compared to Tables 4.8 and 4.9 shows a significant increase in mean optimal objective value with respect to $\rho_{\text{min}}$ when the waiting or overtime costs are increased. Again, the significant lower mean optimal objective value for $\alpha = 1$ compared to $\alpha = 0.1$ and $\alpha = 10$ results from the increase in $c^o$ or $c^w$. The comparison of Tables 4.8 and 4.9 similar to the learning behavior model, show a significant increase in mean optimal objective value with respect to $\rho_{\text{min}}$ for Table 4.8. The similar results for the fatigue server behavior model indicate that this model may also be more sensitive to changes in overtime cost.

In the case of the congestion model, similar results to the learning and fatigue models are
Table 4.4: Mean and 95% confidence interval on the mean optimal objective value of the server learning behavior model when $\alpha = 1$

<table>
<thead>
<tr>
<th>$\rho_{min}$</th>
<th>Objective Value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>Mean: 8935.01</td>
<td>(8872.45, 8997.45)</td>
</tr>
<tr>
<td>0.9</td>
<td>7193.895</td>
<td>(7118.42, 7269.37)</td>
</tr>
<tr>
<td>0.8</td>
<td>5635.54</td>
<td>(5574.27, 5696.81)</td>
</tr>
<tr>
<td>0.7</td>
<td>4004.92</td>
<td>(3920.94, 4088.90)</td>
</tr>
<tr>
<td>0.6</td>
<td>2747.98</td>
<td>(2654.03, 2841.93)</td>
</tr>
<tr>
<td>0.5</td>
<td>1638.77</td>
<td>(1595.46, 1682.08)</td>
</tr>
</tbody>
</table>

Table 4.5: Mean and 95% confidence interval on the mean optimal objective value of the server learning behavior model when $\alpha = 0.1$

<table>
<thead>
<tr>
<th>$\rho_{min}$</th>
<th>Objective Value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>Mean: 59096.31</td>
<td>(59094.26, 60521.55)</td>
</tr>
<tr>
<td>0.9</td>
<td>39374.43</td>
<td>(38907.57, 39841.30)</td>
</tr>
<tr>
<td>0.8</td>
<td>22915.45</td>
<td>(22498.28, 23332.62)</td>
</tr>
<tr>
<td>0.7</td>
<td>10475.22</td>
<td>(10240.22, 10710.22)</td>
</tr>
<tr>
<td>0.6</td>
<td>3461.42</td>
<td>(3371.19, 3551.65)</td>
</tr>
<tr>
<td>0.5</td>
<td>752.01</td>
<td>(726.20, 777.81)</td>
</tr>
</tbody>
</table>

Table 4.6: Mean and 95% confidence interval on the mean optimal objective value of the server learning behavior system when $\alpha = 10$

<table>
<thead>
<tr>
<th>$\rho_{min}$</th>
<th>Objective Value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>Mean: 17498.52</td>
<td>(17496.47, 17654.97)</td>
</tr>
<tr>
<td>0.9</td>
<td>14296.34</td>
<td>(14170.78, 14421.90)</td>
</tr>
<tr>
<td>0.8</td>
<td>11255.31</td>
<td>(11156.12, 11354.51)</td>
</tr>
<tr>
<td>0.7</td>
<td>8288.40</td>
<td>(8212.21, 8364.59)</td>
</tr>
<tr>
<td>0.6</td>
<td>5105.60</td>
<td>(5056.08, 5155.11)</td>
</tr>
<tr>
<td>0.5</td>
<td>2428.57</td>
<td>(2397.03, 2460.10)</td>
</tr>
</tbody>
</table>
Table 4.7: Mean and 95% confidence interval on the mean optimal objective value of the server fatigue behavior model when $\alpha = 1$

<table>
<thead>
<tr>
<th>$\rho_{\text{min}}$</th>
<th>Objective Value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{\text{min}} = 1.0$</td>
<td>8935.01</td>
<td>(8872.45, 8997.45)</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.9$</td>
<td>6215.23</td>
<td>(6110.68, 6325.94)</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.8$</td>
<td>5635.54</td>
<td>(5574.27, 5696.81)</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.7$</td>
<td>4004.92</td>
<td>(3920.94, 4088.90)</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.6$</td>
<td>2747.98</td>
<td>(2654.03, 2841.93)</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.5$</td>
<td>1638.77</td>
<td>(1595.46, 1682.08)</td>
</tr>
</tbody>
</table>

Table 4.8: Mean and 95% confidence interval on the mean optimal objective value of the server fatigue behavior model when $\alpha = 0.1$

<table>
<thead>
<tr>
<th>$\rho_{\text{min}}$</th>
<th>Objective Value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{\text{min}} = 1.0$</td>
<td>59096.31</td>
<td>(59094.26, 60521.55)</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.9$</td>
<td>54353.55</td>
<td>(53771.96, 54935.14)</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.8$</td>
<td>50832.45</td>
<td>(50182.11, 51482.78)</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.7$</td>
<td>47869.03</td>
<td>(47129.04, 48609.02)</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.6$</td>
<td>44330.79</td>
<td>(43792.64, 44868.94)</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.5$</td>
<td>39791.50</td>
<td>(39205.66, 40377.33)</td>
</tr>
</tbody>
</table>

Table 4.9: Mean and 95% confidence interval on the mean optimal objective value of the server fatigue behavior model when $\alpha = 10$

<table>
<thead>
<tr>
<th>$\rho_{\text{min}}$</th>
<th>Objective Value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{\text{min}} = 1.0$</td>
<td>17498.52</td>
<td>(17496.47, 17654.97)</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.9$</td>
<td>17011.29</td>
<td>(16845.58, 17177.00)</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.8$</td>
<td>16280.14</td>
<td>(16140.99, 16419.29)</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.7$</td>
<td>15597.57</td>
<td>(15452.08, 15743.07)</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.6$</td>
<td>15046.39</td>
<td>(14920.50, 15172.28)</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.5$</td>
<td>14372.84</td>
<td>(14271.09, 14474.60)</td>
</tr>
</tbody>
</table>
found. Tables 4.10, 4.11, and 4.12 show that as the $\rho_{\text{min}}$ value increases the mean optimal objective value also increases. Tables 4.11 and 4.12 display the same increased mean optimal objective values for Table 4.11 when $c^{\alpha}$ is increased, however these results are only found when $\rho_{\text{min}} \geq 0.6$. Conversely, for $\rho_{\text{min}} = 0.5$ Table 4.12 and Table 4.11 return lower and higher mean optimal objective values, respectively. It appears that the congestion model may be more sensitive to increases in overtime cost, but sensitivity may also be dependent upon the $\rho_{\text{min}}$ value.

Table 4.10: Mean and 95% confidence interval on the mean optimal objective value of the server congestion behavior model when $\alpha = 1$

<table>
<thead>
<tr>
<th>Objective Value</th>
<th>$\alpha = (1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{\text{min}} = 1.0$</td>
<td>$8935.01$</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.9$</td>
<td>$7193.895$</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.8$</td>
<td>$5635.54$</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.7$</td>
<td>$4004.92$</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.6$</td>
<td>$2747.98$</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.5$</td>
<td>$1638.77$</td>
</tr>
</tbody>
</table>

Table 4.11: Mean and 95% confidence interval on the mean optimal objective value of the server congestion behavior model when $\alpha = 0.1$

<table>
<thead>
<tr>
<th>Objective Value</th>
<th>$\alpha = (0.1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{\text{min}} = 1.0$</td>
<td>$59096.31$</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.9$</td>
<td>$44668.58$</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.8$</td>
<td>$32018.78$</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.7$</td>
<td>$21185.85$</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.6$</td>
<td>$11955.25$</td>
</tr>
<tr>
<td>$\rho_{\text{min}} = 0.5$</td>
<td>$5860.72$</td>
</tr>
</tbody>
</table>

The sensitivity analysis in this section allows us to discuss the practical implications of how the system reacts to an increases in the costs of overtime or waiting. The first trend that was
noted was the positive correlation between optimal objective value and the $\rho^{min}$ value. This positive correlation indicates that as the effect of congestion increases ($\rho^{min}$ decreases), there tend to be lower expected costs of waiting and overtime. The second observation made when comparing increased overtime cost ($\alpha = 0.1$) and waiting cost ($\alpha = 10$) shows that the learning, fatigue, and congestion (when $\rho^{min} \geq 0.7$) behavior models are more sensitive to changes in cost of overtime. Then finally it was observed that for the congestion effect function when $\rho^{min} = 0.5$ there is a higher sensitivity to increases in the cost of waiting time, relative to overtime.

### 4.6 Allotted Time Sensitivity

The effect of server behavior on the optimal patient allotted times (i.e. appointment schedule) was evaluated for each of the three server behavior models. In order to understand this potential sensitivity, each of the three server behavior effect models were solved with multiple $\rho^{min}$ values, and the allotted times for each of the 10 patients were graphed. Each graph plots the mean allotted time for each patient for 30 runs of each server behavior model, and a 95% confidence intervals are provided on the graphs. The results for the learning model are shown in Figures 4.1–4.5, the fatigue model results are shown in Figures 4.7 through 4.11, and the congestion model results are shown in Figures 4.12 through 4.16.

The graphs of allotted time with a learning effect show a progressively decreasing amount of allotted time for patients 3 through 10 as the $\rho^{min}$ value decreases. The allotted time for patients 3 through 10 begin around 0.45 hours for $\rho^{min} = 0.9$ and end at around 0.35 hours for $\rho^{min} = 0.5$. The slow decrease for patients 3 through 10 is intuitive and follows from the

<table>
<thead>
<tr>
<th>$\rho^{min}$</th>
<th>Objective Value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>17498.52</td>
<td>(17496.47, 17654.97)</td>
</tr>
<tr>
<td>0.9</td>
<td>15309.83</td>
<td>(15188.14, 15431.52)</td>
</tr>
<tr>
<td>0.8</td>
<td>13197.15</td>
<td>(13084.72, 13309.58)</td>
</tr>
<tr>
<td>0.7</td>
<td>11039.84</td>
<td>(10948.04, 11131.63)</td>
</tr>
<tr>
<td>0.6</td>
<td>8858.30</td>
<td>(8763.87, 8952.73)</td>
</tr>
<tr>
<td>0.5</td>
<td>6707.66</td>
<td>(6653.36, 6761.95)</td>
</tr>
</tbody>
</table>
Figure 4.1: Graph of Patient Allotted Times, in hours, with Learning Effect, $\rho^{\text{min}} = 0.9$.

Figure 4.2: Graph of Patient Allotted Times, in hours, with Learning Effect, $\rho^{\text{min}} = 0.8$. 
Figure 4.3: Graph of Patient Allotted Times, in hours, with Learning Effect, $\rho^{\text{min}} = 0.7$

Figure 4.4: Graph of Patient Allotted Times, in hours, with Learning Effect, $\rho^{\text{min}} = 0.6$
decreased amount of time needed to complete service because the health care server is learning, and thus working faster. Also as $\rho_{min}$ decreases the range of allotted time for all patients begins to decrease, until $\rho_{min} = 0.5$ where the range again increases which is most likely because the learning effect function has such a large impact on the procedure duration for each patient. From the five figures we also observe that the first patient is given the shortest amount of allotted of all 10 patients time when $\rho_{min} \geq 0.7$.

A large difference in shape can be seen between the graph of allotted time for $\rho_{min} \geq 0.6$ and $\rho_{min} = 0.5$. When $\rho_{min} \geq 0.6$ the first patient requires a shorter amount of allotted time than the second patient however when $\rho_{min} = 0.5$ the second patient has a shorter allotted time than the first. In an attempt to further understand the pattern as $\rho_{min}$ decreases a graph of allotted time when $\rho_{min} = 0.55$ is shown in Figure 4.6. The graph shows a decrease in allotted time between the first and second patient, but a smaller range of decrease than in Figure 4.5. The pattern shown for the first patient indicates that as the learning effect becomes more prominent (i.e. $\rho_{min}$ decreases) the first patient’s allotted time increases.

In the case of the graphs of allotted time with a fatigue effect (Figures 4.7 through 4.11), the graphs show an increasing allotted time for patients as the $\rho_{min}$ value decreases. The in-
Figure 4.6: Graph of Patient Allotted Times, in hours, with Learning Effect, $\rho^{\text{min}} = 0.55$

Figure 4.7: Graph of Patient Allotted Times, in hours, with Fatigue Effect, $\rho^{\text{min}} = 0.9$
Figure 4.8: Graph of Patient Allotted Times, in hours, with Fatigue Effect, $\rho^{\text{min}} = 0.8$

Figure 4.9: Graph of Patient Allotted Times, in hours, with Fatigue Effect, $\rho^{\text{min}} = 0.7$
Figure 4.10: Graph of Patient Allotted Times, in hours, with Fatigue Effect, $\rho^{\min} = 0.6$

Figure 4.11: Graph of Patient Allotted Times, in hours, with Fatigue Effect, $\rho^{\min} = 0.5$
increasing allotted time follows with the pattern of the fatigue model. The fatigue model has a dome shape for the allotted time with each successive patient and thus the optimal schedule has appointment times that are increasing and then begin to gradually decrease with respect to the patient position in the schedule. Further, as the $\rho^{\text{min}}$ value decreases there is a larger range of allotted times beginning with a range of 0.3 hours for $\rho^{\text{min}} = 0.9$ and ending with a range of 0.4 hours for $\rho^{\text{min}} = 0.5$. The increase in allotted time as $\rho^{\text{min}}$ decreases mirrors the increase in range of the fatigue effect for each patient as $\rho^{\text{min}}$ decreases.

A quick examination of the congestion graphs reveals that for each $\rho^{\text{min}}$ value the allotted time for each patient, beginning with patient 2, shows a decreasing trend as they are scheduled later in the day. The decrease in allotted time is most likely because congestion in the office will begin to build up throughout the day as more patients arrive for their appointment. Then, since the model is set up to have patient’s procedure durations decrease by a set factor according to the number of patient’s currently waiting, the assigned allotted time would be less for each patient. Additionally a closer look at the provided graphs (Figures 4.12 - 4.16) shows that there is a correlation between a smaller $\rho^{\text{min}}$ value and a larger range of allotted time for patients 2 through 10. When $\rho^{\text{min}} = 0.9$ patient allotted time has a range of 0.11 hours compared to when $\rho^{\text{min}} = 0.5$ patient allotted has a range of 0.18 hours, for patients 2 through 10. An increase in
Figure 4.13: Graph of Patient Allotted Times, in hours, with Congestion Effect, $\rho_{\text{min}}^{\text{in}} = 0.8$

Figure 4.14: Graph of Patient Allotted Times, in hours, with Congestion Effect, $\rho_{\text{min}}^{\text{in}} = 0.7$
Figure 4.15: Graph of Patient Allotted Times, in hours, with Congestion Effect, $\rho_{\text{min}} = 0.6$

Figure 4.16: Graph of Patient Allotted Times, in hours, with Congestion Effect, $\rho_{\text{min}} = 0.5$
range of .07 hours is shown between when $\rho_{\text{min}} = 0.9$ and when $\rho_{\text{min}} = 0.5$. Finally a noticeable characteristic of the congestion model allotted time graphs is that the initial scheduled patient is always given the shortest allotted time of all 10 patients, which is reminiscent of the fatigue model.

### 4.7 Heuristics

In this section we propose some easy to implement heuristics for each of the three server behavior models. The heuristics for the learning and fatigue models were determined based on their respective server behavior effect equations. The congestion behavior heuristic was derived from the analysis of the results in Section 4.6.

Below we describe the motivation behind each model heuristic, and support why each approach was selected. Then to understand how effective each heuristic is we compare the mean optimal objective value and 95% confidence interval to that for each of the heuristics. The mean optimal objective values are then compared to those found in Section 4.4 to determine the effectiveness of each heuristic.

#### 4.7.1 Learning and Fatigue Heuristic

The heuristic determined based on the learning and fatigue models is based on the effect equations, 3.1 and 3.2, mentioned in Section 3.2.1. In a situation where a learning behavior is believed to affect appointment times we determined that patient’s appointments could be scheduled based on the following heuristic determining allotted time for each patient:

$$x_i = g^{L}_{\rho_{\text{min}}}(i) \cdot \mu_i, \ \forall i.$$  \hspace{1cm} (4.1)

Similarly, for the fatigue behavior model the following heuristic is suggested:

$$x_i = \rho_{\text{min}}^{F}(i) \cdot \mu_i, \ \forall i.$$  \hspace{1cm} (4.2)

Each heuristic calculates each patient’s allotted time by determining the current server behavior effect for that patient, $i$, and multiplying this value by the mean service time for patient $i$, $\mu_i$. 

45
Table 4.13 shows the mean optimal objective value when a learning effect is present using the 2-SLP learning model, learning heuristic, and ASP model for \( n = 10 \) patients and our previously defined service time distribution, \( U(0, 1) \). Table 4.13 reveals significantly higher values for the heuristic and ASP for all \( \rho^{\text{min}} \) values when compared to the mean optimal objective value using the learning effect 2-SLP. The difference in means between the 2-SLP and the heuristic are much smaller for \( \rho^{\text{min}} \geq 0.6 \) with a differences ranging from around 125 to 515, as opposed to \( \rho^{\text{min}} = 0.5 \) that has a difference of over 1000. It should however be noted that the mean optimal objective values using the ASP are smaller than the mean optimal objective values using the heuristic for \( \rho^{\text{min}} \geq 0.8 \). The lower mean optimal objective values between the ASP and the heuristic suggests that the current heuristic may be ineffective. In some cases however, when the heuristic mean optimal objective values fall below those of the ASP, this may indicate the heuristic may be on the correct path.

Table 4.14 shows the mean optimal objective value when a fatigue effect is present using the 2-SLP fatigue model, fatigue heuristic, and ASP model. Comparing each column reveals

Table 4.13: Comparison of mean optimal objective values for a learning effect using the 2-SLP learning model, learning heuristic, and ASP model.

<table>
<thead>
<tr>
<th>Learning Effect Mean Optimal Objective Values</th>
<th>2-SLP</th>
<th>Heuristic</th>
<th>ASP</th>
</tr>
</thead>
<tbody>
<tr>
<td>( g_{0.9}^L(i) )</td>
<td>6521.35</td>
<td>6965.41</td>
<td>6031.39</td>
</tr>
<tr>
<td>( g_{0.8}^L(i) )</td>
<td>4288.64</td>
<td>4606.03</td>
<td>4208.216</td>
</tr>
<tr>
<td>( g_{0.7}^L(i) )</td>
<td>2469.28</td>
<td>2595.54</td>
<td>2791.14</td>
</tr>
<tr>
<td>( g_{0.6}^L(i) )</td>
<td>1165.82</td>
<td>1678.74</td>
<td>1893.42</td>
</tr>
<tr>
<td>( g_{0.5}^L(i) )</td>
<td>456.45</td>
<td>1533.83</td>
<td>1304.08</td>
</tr>
</tbody>
</table>

Table 4.14: Comparison of mean optimal objective values for a fatigue effect using the 2-SLP fatigue model, fatigue heuristic, and ASP model.

<table>
<thead>
<tr>
<th>Fatigue Effect Mean Optimal Objective Values</th>
<th>2-SLP</th>
<th>Heuristic</th>
<th>ASP</th>
</tr>
</thead>
<tbody>
<tr>
<td>( g_{0.9}^F(i) )</td>
<td>8300.18</td>
<td>8988.09</td>
<td>7602.50</td>
</tr>
<tr>
<td>( g_{0.8}^F(i) )</td>
<td>7828.05</td>
<td>8429.37</td>
<td>7306.17</td>
</tr>
<tr>
<td>( g_{0.7}^F(i) )</td>
<td>7395.68</td>
<td>7978.17</td>
<td>6842.60</td>
</tr>
<tr>
<td>( g_{0.6}^F(i) )</td>
<td>6960.64</td>
<td>7493.10</td>
<td>6518.18</td>
</tr>
<tr>
<td>( g_{0.5}^F(i) )</td>
<td>6475.96</td>
<td>6995.14</td>
<td>6153.33</td>
</tr>
</tbody>
</table>
that the fatigue heuristic produces a similar change in optimal objective value as the learning heuristic. The optimal objective value using the fatigue heuristic again returns significantly higher values than the 2-SLP model for fatigue. However, for the fatigue heuristic the difference in means remains smaller for all \( \rho_{\text{min}} \) values with differences ranging from around 130 to 580. Comparing the mean optimal objective values for the heuristic to those of the ASP are not a viable comparison in this case because of a decrease in mean, see Table 4.2. The results for the fatigue heuristic indicate that it may be a more accurate portrayal of the behavior without use of a complicated model, but some adjustments are still needed for accuracy.

### 4.7.2 Congestion Heuristic

Determining a heuristic for the congestion model was a more difficult task than for the learning and fatigue models, because the congestion model is more complex and therefore has a less intuitive set of results. In order to better understand the results produced by the congestion model a deterministic version of the system was created. The deterministic version of the congestion model removes the randomness associated with the stochastic model. Therefore, a more in-depth understanding of the model and the relationship between its variables and solutions may be gained.

The deterministic mean value model was created by replacing the stochastic variables, \( z(\omega) \), with their mean. Thus, since the probability distribution for each variable was \( U(0,1) \) in our experiments, their mean value was 0.5. Using these values the model was run using five different \( \rho_{\text{min}} \) values, \( \rho_{\text{min}} = 1.0, 0.9, 0.8, 0.7, 0.5 \), in the linear congestion effect function, \( f_{\rho_{\text{min}}}(j) \), and a varying number of patients, \( n \). The resulting allotted times for each \( \rho_{\text{min}} \) value and selected \( n \) were recorded and examined for patterns. The resulting pattern found that for \( n \geq 4 \) the allotted time for patients 1 through \( n - 1 \) could be calculated by the equation:

\[
x_i = \frac{1}{2} - \left( \frac{1 - \rho_{\text{min}}}{2(n - 1)} \right) i, \forall i.
\]

Table 4.15 displays the mean optimal objective value when congestion effect is present using the 2-SMIP congestion model, congestion heuristic, and ASP model. Comparing the columns of Table 4.15 reveals lower mean objective values for the heuristic compared to the ASP model, when \( \rho_{\text{min}} \) is small. Therefore we conclude that the heuristic provides some benefit, particularly when the congestion effect is significant.
4.8 Conclusions

The experiments for the three server behavior models help to determine several important factors about each model. The learning and fatigue models while similar in formulation have different optimal schedules. It also seems to be that more significant effects are observed with the learning model. Suggesting that learning may have a greater influence on optimal schedule than fatigue. The congestion model has a more significant effect on the optimal appointment schedule compared to the learning and fatigue models, which result from the increased constraint considerations needed for the congestion model.

The analysis of the sensitivity of the models shows each model’s response when a particular server behavior is present and accounted for, and when the behavior is ignored. The learning and fatigue models have opposing reactions in their resulting optimal objective values in this case. The learning model returns significant increases in the objective function when a learning behavior is present and not accounted for in the scheduling model. The fatigue model suggests a decrease in cost when fatigue is present and not accounted for, however there are no statistically significant decreases that can be observed for the specific test cases we considered. The congestion model also returns significant increases in cost when server’s react to the presence of congestion but the model does not consider an effect from congestion. Each of the models also tend to show more significant decreases when the server behavior effect is greater, meaning a smaller $\rho_{min}$ value.

Table 4.15: Comparison of mean optimal objective values for a congestion effect using the 2-SMIP congestion model, congestion heuristic, and ASP model. ($^+$ indicates a statistically significant difference between the mean optimal objective values for the heuristic and ASP, where the difference between the mean optimal objective value of the heuristic and the 2-SMIP is smaller than between the ASP and 2-SMIP.)

<table>
<thead>
<tr>
<th>Congestion Effect</th>
<th>Mean Optimal Objective Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{0.9}(j)$</td>
<td>7193.90</td>
</tr>
<tr>
<td></td>
<td>7738.15</td>
</tr>
<tr>
<td></td>
<td>7220.61</td>
</tr>
<tr>
<td>$f_{0.8}(j)$</td>
<td>5635.54</td>
</tr>
<tr>
<td></td>
<td>5922.68</td>
</tr>
<tr>
<td></td>
<td>5721.61</td>
</tr>
<tr>
<td>$f_{0.7}(j)$</td>
<td>4067.90</td>
</tr>
<tr>
<td></td>
<td>4316.86</td>
</tr>
<tr>
<td></td>
<td>4320.13</td>
</tr>
<tr>
<td>$f_{0.6}(j)$</td>
<td>2718.07</td>
</tr>
<tr>
<td></td>
<td>2896.07$^+$</td>
</tr>
<tr>
<td></td>
<td>3167.03</td>
</tr>
<tr>
<td>$f_{0.5}(j)$</td>
<td>1610.34</td>
</tr>
<tr>
<td></td>
<td>1946.01$^+$</td>
</tr>
<tr>
<td></td>
<td>2263.25</td>
</tr>
</tbody>
</table>
Analysis on cost sensitivities for the three server behavior models show higher sensitivity for all three models to an increase in overtime cost. In the majority of cases for the behavior models, increasing the cost of wait time leads to a significant optimal objective value increase, however a more significant increase in optimal objective value can be observed for an increase in the cost of overtime. These results suggest that when server behavior has an effect on appointment duration, scheduling should be focused in a manner such that employees are prevented from working longer than their scheduled hours. However, further analysis should be performed on our models’ sensitivity to overtime cost before a final conclusion can be drawn.

The main focus of each model is to determine the days schedule for a sequence of patients. The schedule is determined by finding the amount of allotted time for each patient and determining the arrival time for each patient based off of these values. Evaluation of allotted time reveals that as $\rho^{\text{min}}$ decreases the server behavior effect becomes more significant. As $\rho^{\text{min}}$ decreases the learning model begins to return shorter allotted times for later scheduled patients and a longer allotted time for the initial patient. The fatigue model responds oppositely to decreasing $\rho^{\text{min}}$ values, returning longer allotted times. Both of these characteristic of the optimal schedule are consistent with intuition. In the case of the congestion model it, similar to the learning model, returns shorter allotted time for patients as $\rho^{\text{min}}$ decreases, and further shows a larger decrease between successive patients. The decrease in allotted time for the congestion model follows the pattern that the server will increase their speed during the day because there is more probability that there will be patients waiting for their appointment.

The heuristics created for each of the three models help to gain an understanding of how each model works. Our heuristics we were able to produce some reasonable schedule estimates, but comparison to each of the server behavior models indicated that the heuristics may not be as effective as would be desired for some choices of $\rho^{\text{min}}$, suggesting opportunities for future research. However, some aspects of these models, such as the deterministic model for the congestion effect or similar effect equations for learning and fatigue, could be used to produce more effective heuristics in the future.
Chapter 5

Conclusions

5.1 Introduction

We examined and produced several models that represented different server behaviors that may affect health care providers. Our models focused on the sever behavior effects of a learning curve, fatigue, and a reaction to congestion. The focus of each model was to minimize the total expected cost associated with patient waiting and server overtime. Our models sought to determine better methods for appointment scheduling, and ways to find easier approaches to appointment scheduling in the future. The production of each model required an understanding of environmental factors that could influence servers. Producing and validating each of the models helped to determine more effective and accurate ways of modeling.

5.2 Conclusions

Model analysis revealed several similar factors that should be focused on more in the future when attempting to decrease costs. In the case of the learning and congestion models we found that when each of the server behaviors were ignored increased costs were incurred. These results suggest that more models in the future should account for these server behaviors in order to correctly schedule patients. Another important conclusion drawn from our work involves each server behavior model’s sensitivity to changes in overtime cost. As overtime costs increased each model returned higher optimal objective values. The results indicate that future models should focus on maintaining low amounts of overtime for servers. Finally, the learning and congestion models indicate progressive decreases in allotted times as server behavior effects became more influential to service times. The decrease in allotted time suggests that scheduling patients more closely together as the day goes on may be an effective strategy. The fatigue
model returns longer allotted times for patients as the learning effect becomes more prominent, suggesting that patients should be scheduled further apart to help decrease waiting times.

5.3 Limitations and Future Work

There some limitations associated with the models produced in this research. One main limitation that should be examined in depth in the future is the structure of the congestion model formulation. In an attempt to understand time constraints on solving the congestion formulation it was discovered that the model may not be required to be defined as a MIP. When the integer restrictions on particular variables are removed, the model will still return integer values for these variables. Since it is unnecessary to define integer variables it suggests that the model has some properties that have not currently been recognized. Determining the model properties provides a better understanding of its results and potential ways to decrease computation time.

Another limitation that appeared during this set of research involved the fatigue effect function. It seemed that the values used in the fatigue effect function may not allow us to easily evaluate the model because they are too similar to their being no effect, see the discussion at the end of Section 4.4, Different, more significant versions of the fatigue effect function should be applied in the future. A final limitation of this work is the support behind the heuristics created from the learning, fatigue, and congestion models. Each of the heuristics were evaluated against their respective models and none of the heuristics were found to be a decent estimate of their respective model. Some future work should be dedicated to finding better heuristics for each of the models.

Our research also leaves room for a large amount of future work that could be done. One focus could be to also include patient responses, in combination with server behavior. Several patient behaviors were cited in Chapter 2 that could be interesting additions, including: probability of patient no-shows, balking, change in needs because of increased or decreased waiting time, or effect of patient age or gender on service time. Another future focus could be on the cost of idle time. Each model has idle time cost built into the objective function, but very little of our research was focused on this aspect and analysis may return interesting results. Finally some future work that would be very interesting and would help to indicate the validity of this work is proof of true server behavior. Analysis of real health care facility data that shows a shift in service time as a result of environmental effects would provide a wonderful platform for this research. The work presented in this thesis provides a basis for these future extensions.
REFERENCES


