ESSAYS IN SCHEDULING: APPLICATIONS IN HEALTH CARE AND MANUFACTURING

by

SUBHAMOY GANGULY

B.E., Bengal Engineering College, India, 1996

M.B.A., Michigan State University, 2005

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Operations and Information Management

2013

This thesis entitled:

Essays in Scheduling: Applications in Health Care and Manufacturing
written by Subhamoy Ganguly

has been approved for the department of Operations and Information Management

_____
Stephen R Lawrence

_____
Manuel Laguna

*Date*_____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

ABSTRACT

Ganguly, Subhamoy (Ph.D., Operations and Information Management)

Essays in Scheduling: Applications in Health Care and Manufacturing

Thesis directed by Associate Professor Stephen R. Lawrence

The problem of finding an optimal schedule is ubiquitous in human endeavor. Airlines strive to find the best schedule for their aircrafts and flight-crews, academic scheduling offices at universities schedule classes, and business organizations prioritize their jobs or customers and schedule them accordingly. Even in our day to day lives, whether consciously or not, we decide on our daily schedule and go about our sequence of tasks, prioritizing certain tasks, and procrastinating on others. Not surprisingly, scheduling of people, tasks, and manufacturing jobs constitute a significant proportion of the operations management literature. Despite all the research that has been published on scheduling, we encounter scheduling problems that call for novel approaches, and when treated with appropriate tools and techniques, provide interesting managerial insights. The dissertation consists of three such scheduling problems, two from the health care industry, and one from manufacturing. The first health care paper (Ganguly *et al*. 2013) employs mixed-integer linear programming to schedule physicians for a medical emergency department (ED). The second paper (Ganguly and Samorani 2013) uses an analytical approach for making real time updates to patient schedules in outpatient clinics that primarily operate on an appointment basis. The paper on manufacturing (Ganguly and Laguna 2013) develops heuristic techniques for a particular type of job-sequencing problem that involves two types of setups.

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Prof. Stephen Lawrence, who has guided me all along this challenging but rewarding path that leads to a Ph.D. Next, I would like to thank my committee members and the faculty at the Leeds School of Business. I am lucky to have such accomplished and supportive people as my professors and committee members. Without your help and guidance, I would not have accomplished this mission.

I am indebted to my co-authors: Profs. Stephen Lawrence and Manuel Laguna at the University of Colorado, and Prof. Michele Samorani, who was a fellow Ph.D. student, and is now a faculty at the University of Alberta. From Steve, I have learned how to write a research paper that makes a contribution to a real world business problem; from Manuel, I have learned the science of metaheuristics and the art of applying it; and from Michele, I have learned skills of programming and simulation.

Whatever I have learned in graduate school here in Colorado built upon the foundation of education that I received since childhood. I am thankful to all the wonderful teachers that I have had since elementary school through high school and into college. Some of them were truly great teachers who have inspired me to join the academic profession. I should make a special mention of my maternal grandfather Late Shyamapada Mukherjee, who was my first teacher of critical reasoning; he was the socratic teacher who taught me to question the rationale behind all conventional wisdoms, whether they are based on textbooks, scriptures, or social customs.

Whatever accomplishments I have made in life are all attributable to the care and sacrifice of my parents. My mother, Dr. Jayasree Mukherjee (Ganguly), started her graduate studies while raising a family. She went on to top her M.A. class, complete her Ph.D. in philosophy, and work as a faculty member at two colleges for three decades. She has definitely been an inspiration for my graduate studies. If I had her perseverance, I would definitely have had a better academic track record, and would possibly have completed my Ph.D. sooner in life.

My father, Mr. Ramendu Lal Ganguly, is a compassionate and friendly human being, one with a terrific sense of humor and a presence of mind; he is happiness personified. I wish my personality were more like his – that would have made me a more affable individual and a better human being. I am glad to have such supportive parents who have always compromised on their pleasures to enable me to obtain the best education and life experience possible.

My ventures into graduate school would not have been successful without the support and cooperation of my wife Ms. Rituparna Ganguly. While raising our children, she has worked and supported the family while her husband remained a graduate student for the most part of our married life. I thank her for her unflinching support to my academic endeavor.

I am thankful for our two beautiful daughters, Kojagoree and Prateeti. Amid the uncertainties of research, dissertation, and job search, the pleasure of parenting has been an oasis that has kept me revitalized.

I am thankful to the wonderful friends that I have had over the years – in high school, college, work, and beyond. My interactions, arguments, and debates with those bright and thoughtful individuals have improved my interpersonal skills, made me more articulate and open to ideas. My friends have always been, and still are, a very important part of my life. I owe much of my maturity and development to my friends.

Last but not the least, I am thankful to God, to luck, to the circumstances in my life that have brought me this far. I am grateful to the wonderful higher education system of the United States of America which provides the opportunity and resources to obtain world class education, not just for its citizens, but for foreign nationals as well. Had I not been exposed to the US university system, I would possibly never have discovered my interest in graduate education.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 - INTRODUCTION

The problem of finding an optimal schedule is ubiquitous in human endeavor. Airlines strive to find the best schedule for their aircrafts and flight-crews, academic scheduling offices at universities schedule classes, and business organizations prioritize their jobs or customers and schedule them accordingly. Even in our day to day lives, whether consciously or not, we decide on our daily schedule and go about our sequence of tasks, prioritizing certain tasks, and procrastinating on others. Not surprisingly, scheduling of people, tasks, and manufacturing jobs constitute a significant proportion of the operations management literature. Despite all the research that has been published on scheduling, we encounter scheduling problems that call for novel approaches, and when treated with appropriate tools and techniques, provide interesting managerial insights. The dissertation consists of three such scheduling problems, two from the health care industry, and one from manufacturing. The first health care paper (Ganguly *et al*. 2013) employs mixed-integer linear programming to schedule physicians for a medical emergency department (ED). The second paper (Ganguly and Samorani 2013) uses an analytical approach for making real time updates to patient schedules in outpatient clinics that primarily operate on an appointment basis. The paper on manufacturing (Ganguly and Laguna 2013) develops heuristic techniques for a particular type of job-sequencing problem that involves two types of setups.

One of the key issues in operations management is about improving operational efficiency by balancing supply, or capacity, with demand. On the one hand, businesses with uncertain demand must make prudent decisions to manage their capacity - overcapacity leads to higher operating costs while shortage of capacity leads to lost opportunity and unmet customer demand. On the other hand, for a given capacity, a firm has to: manage its demand; prioritize its customers; and schedule and serve its jobs and clients in the most efficient manner. This need of capacity planning on one side and demand management on the other permeates across manufacturing and service sectors. In this dissertation, we address specific problems pertaining to both sides of this dichotomy. In ED staff planning (Chapter 2), we deal with the supply side; we investigate the issue of personnel capacity planning at a medical emergency department

that faces stochastic demand. The remaining papers deal with managing the demand side; in Chapter 3, we deal with patient appointments in an out-patient clinic to balance clinic overtime costs and patient dissatisfaction caused by long wait times; and in Chapter 4, we try to arrive at the best sequence of manufacturing jobs that involves two different types of sequence dependent setup costs.

My research makes a humble contribution to the body of knowledge of operations management that deals with balancing supply and demand, through planning of personnel capacity and managing schedules of customers and jobs. A brief outline of the three papers is provided in the following sections.

1.1. Emergency Department (ED) Staff Scheduling

It is easy to see why personnel scheduling is one of the dominant streams of literature in operations management. While it is vital for most organizations to have the "right staff on duty at the right time" to satisfy customer needs and be successful as a business, personnel costs typically constitute one of their largest expenditures. Thus, it becomes important to have the right number of appropriately skilled staff on duty. Furthermore, for staff planning and scheduling, there are legal constraints and company policies that need to be adhered to. Ernst *et al*. (2004), in their comprehensive annotated bibliography, list about 700 papers on personnel scheduling and rostering, and categorize them based on (i) the stage of the scheduling process that it addresses; (ii) application area; and (iii) solution method. Among application areas, personnel scheduling in health care settings account for almost 20% of the papers listed, 80% of which is devoted to nurse scheduling. Compared to the abundance of literature on nurse scheduling, paucity of papers on physician scheduling is noticeable (Brunner *et al*. 2009).

Among physician scheduling problems, staffing emergency departments (EDs) pose an interesting problem that involves stochastic arrival of patients of varying acuity, who have to be treated and stabilized by providers of varying skill profiles, working solo or in *ad hoc* teams. Of papers devoted to ED physician scheduling, most presuppose that required provider staffing levels are known and therefore focus on creating a schedule or roster of personnel assignments (Beaulieu *et al.* 2000; Carter & Lapierre 2001; Ernst *et al.* 2004; Ferrand *et al.* 2011). Another stream of research has used work

sampling methods to assess physician activity and thereby decide on the appropriate level of hospital staffing (Oddone *et al.* 1993; Ben-Gal *et al.* 2010). While work sampling methods are a reliable tool to obtain good estimates of physician time requirements needed to treat a variety patients, they are not intended to create a planning schedule of providers and teams required to meet demand on an hour-by-hour basis.

In contrast to the extant literature on ED physician scheduling, the essay in Chapter 2 (Ganguly *et al*., 2013) presents an ED staff planning model that takes into account the distributions of patient-care demand and makes the best staffing and scheduling decisions to minimize overall cost, while fulfilling service level constraints and scheduling rules. Using empirical data collected from three active EDs, we develop an analytic model to provide an effective staffing schedule for EDs. Patient demand is aggregated into discrete time buckets and used to model the stochastic distribution of patient demand within these buckets, which considerably improves model tractability. This model is capable of scheduling providers with different skill-profiles who work either individually or in teams, and with patients of varying acuity-levels. We show how our model helps to balance staffing costs and patient service levels, and how it facilitates examination of important ED staffing policies.

This paper makes several contributions. First, it shows how patient workload demand by acuity level can be aggregated into discrete time buckets to estimate the stochastic distribution of aggregate demand. Second, the model allows ED administrators to effectively schedule providers with a range of skill-levels. In an ED setting, patient demand distributions apparent to medical providers are functions of their skill-levels. Providers can treat any patient with acuity at or below their skill-level, which is similar to the nursing skill classes of Warner and Prawda (1974), a characteristic which Bard (2004) termed as "downgrading". Third, our model accommodates the use of provider teams in an ED, in which a provider with lesser skills can treat high-acuity patients when supervised by a more qualified provider. This adds complexity to the problem since the time demands placed on supervising providers must be accommodated within the model. The results provide an optimal staffing plan by provider type and

facilitate the creation of efficient and effective staff schedules. To our knowledge, we are the first to employ this approach in the context of ED staff planning in particular and to staff planning in general.

1.2. Provider's Wait-Preempt Dilemma

If personnel capacity planning is management of the supply side, scheduling of jobs, clients, and patients involves management of the demand side. With limited capacity of personnel and equipment, it becomes necessary to manage demand in the most efficient manner. In the manufacturing sector, this involves scheduling and sequencing of jobs; in service industries, it often involves scheduling client appointments. The next essay, presented in Chapter 3 (Ganguly & Samorani, 2013), also deals with the health care industry, but rather than schedules of providers, it deals with appointments of patients.

Over the past several years, skyrocketing healthcare costs and lack of easy access to healthcare providers have generated a great deal of research focused on clinic productivity and accessibility. It has been firmly established that for most non-urgent care clinic settings, scheduling appointments is an effective way to manage clinic resources and to keep patient wait times within reasonable limits (Gupta and Denton 2008). While appointment scheduling works reasonably well for many clinics, it does create two unintended consequences: patient no-shows and patient unpunctuality, both of which adversely affect clinic performance (LaGanga and Lawrence 2007, White and Pike 1964). A scheduled patient not showing up at all is considered a no-show, while unpunctuality refers to the early or late arrival of scheduled patients.

While significant work has been done on patient no-shows (LaGanga and Lawrence 2007, 2012), there exists little research that addresses the issue of unpunctuality in depth. The phenomenon of patient unpunctuality and its ill-effects were first documented in White and Pike (1964); late arrival of a patient may negatively affect the clinic overtime and waiting time of the following patients. Unpunctuality of patients can also result in patients arriving out of appointment order, i.e. the patient with a later appointment arriving sooner. This can result in a dilemma if an idle provider faces a situation where a later scheduled patient has already arrived, but the patient supposed to be seen next has not arrived yet.

We investigate the management of such patients, who arrive early and out of turn for appointments in outpatient healthcare clinics. Although existing work shows that some patients arrive very early for scheduled appointments, little research has been undertaken to identify optimal rules to deal with them. Chapter 3 formally defines the provider's wait-preempt dilemma and develops an analytic method that helps the clinic to make the optimal decision when faced with the dilemma, whether to wait for the patient scheduled next, or to preempt and see the patient who has already arrived. In previous works, this dilemma has typically been addressed by either seeing the waiting patient right away, or by waiting for an arbitrarily predetermined time. We show that for reasonable assumptions regarding the arrival pattern of patients, it is possible to analytically determine that in order to maximize the expected utility of the clinic, how long should the provider remain idle while waiting for the next scheduled patient, if a patient scheduled later has already arrived. The problem takes into account the cost of keeping patients waiting and the cost of clinic overtime. The results provide insights into real-time schedule management at a clinic, and depending on clinic specific parameters, allow clinic managers to come up with optimal preemption policies.

This research makes the following contributions to the literature regarding real-time management of patient schedules. First, we optimally solve the wait-preempt dilemma for the two patient case. Next, by identifying structural properties of the cost function, we prove that a first-come-first-served policy is never optimal, and come up with managerial insights as to when it is optimal to wait and when to preempt. We also provide a heuristic solution method for the N-patient case and provide a software program that clinics can readily use to arrive at optimal policies of dealing with the dilemma. Our work is not only applicable to outpatient health clinics, but also to any other appointment-based activity, such as lawyer offices.

1.3. Dual Setup Scheduling

Just as appointment scheduling of patients or clients is a common way of demand side management in service industries constrained by personnel capacity, manufacturing industries with limited manufacturing

capacity use scheduling and sequencing of manufacturing jobs as a way of managing and coordinating demand. Given a set of production jobs and a set of resources needed to complete the jobs, it is important to schedule them in a manner that fulfills objectives such as utility maximization or cost minimization, while adhering to the operational constraints such as resource availability.

Manufacturing job scheduling has been an extensively researched area in operations management - entire textbooks have been devoted to the mathematics of job-shop scheduling and sequencing (French, 1982). A subset of the manufacturing job scheduling literature deals with scheduling of jobs that involve setup costs (Cheng *et al*. 2000, Zhu and Wilhelm 2006, Allahverdi *et al*. 1999, Allahverdi *et al*. 2008). Scheduling of jobs involving setup costs become further complicated when the setups are sequence dependent, e.g. the setup involved in processing *Job B* after *Job A* can be different from the setup involved in processing *Job C* after *Job A*. While there are several papers that consider sequence dependent setups, the extent of sequence dependence considered in these works are limited to adjacent jobs (Szwarc and Gupta 1987, Laguna 1999, Rajendran and Ziegler 2003, Ruiz *et al*. 2005 and Jungwattanakit *et al*. 2009). The essay in Chapter 4 (Ganguly & Laguna 2013) introduces and addresses a manufacturing job sequencing problem that involves two types of setups, and provides alternative solution approaches.

The paper addresses a problem of sequencing batches in consecutive loops that is often encountered in production systems with closed-loop facilities. We study a problem encountered in a production facility in which plastic parts of several shapes must be painted with different colors to satisfy the demand given by a set of production orders. The shapes and the colors produce a dual-setup problem that to the best of our knowledge has not been considered in the literature; adjacent items of different colors require a paint setup, and items of different shape that are separated exactly by the length of the loop require a tool setup. We formulate the problem as a mixed-integer program and discuss the limitations of this approach as a viable solution method. We then describe two alternative solution approaches that are heuristic in nature: one specialized procedure developed from scratch and the other

one built in the framework of commercial software. Our computational experiments are designed to assess the advantages and disadvantages of both approaches.

This paper makes the following contributions. Earlier works involving sequence dependent setups have typically assumed that the setup cost incurred by the current job is dependent only on the current job and its immediate predecessor. In contrast, we introduce a real world problem that considers not only the horizontal setup cost due to the current job and its immediate predecessor, but also the vertical setup cost incurred by the current job and the job that occupied the same position in the prior loop. We show that consideration of these two types of setup costs adds significant complexity to the problem. Further we present three alternative solution methods: a mixed integer programming (MIP) solution method and two different heuristic approaches. The MIP model works well for small problem instances and the solutions obtained from it serve as benchmarks for the heuristic procedures. One of the heuristic procedures utilizes a metaheuristics based commercial software, which provides the advantage of faster development time. The other heuristic is built from scratch, is tailored to the particular problem at hand, and provides superior results for larger size instances.

1.4. Organization of the Following Chapters

The rest of the dissertation is organized as follows. Chapter 2 discusses the personnel capacity planning for a medical ED with stochastic demand for patient care. Chapter 3 defines a provider's wait-preempt dilemma, the predicament faced by an idle provider when the patient scheduled next has not shown up yet, but a patient scheduled for a later slot has already shown up. The paper investigates the best way of managing arriving patients by making optimal decisions of wait and preempt. Chapter 4 introduces a manufacturing job sequencing problem with two types of setup costs. For a given production capacity and a set of production orders, we present alternative solution methods that minimize total setup costs. We conclude in Chapter 5 by summarizing contributions of the three papers and by outlining possible avenues of further research.

Chapter 2 - EMERGENCY DEPARTMENT STAFF PLANNING TO IMPROVE PATIENT CARE AND REDUCE COSTS

In Ganguly *et al*. (2013), we develop a mixed-integer linear programming model to schedule physicians for a medical emergency department.

2.1. Introduction

Medical emergency departments (EDs) are the front line of acute health-care defense in most medical systems where patients arrive with a variety of ailments, illnesses, and injuries, which can vary from trivial (e.g., runny noses and indeterminate aches) to life threatening (e.g., heart attacks and gunshot wounds). The role of emergency department staff is to quickly assess the acuity of an arriving patient and then to either treat the patient for discharge or to medically stabilize the patient for admission to a regular hospital for further treatment.

A challenging problem for ED administrators is to determine the best mix of medical staff so that arriving patients are promptly treated by qualified personnel while minimizing overstaffing expenses. Patients must be seen quickly, especially those with higher acuity levels, and costs must be contained. When too many ED personnel are deployed, utilization suffers and ED costs increase. When too few personnel are available or when the mix of providers is wrong, patient service suffers, wait times increase for low acuity patients who may renege, urgent care cases may be directed to other EDs, revenues decline, and both patient and provider satisfaction declines.

Among staff planning problems, emergency departments are especially complex and difficult for several reasons. First, an ED must accommodate highly stochastic patterns of patient arrivals and the random nature of their acuity, which may vary by time of day, day of week, and season of the year. Second, ED medical providers have a range of skill levels, which means that highly skilled providers can treat any arriving patient, but providers with lower skills can only treat patients with acuities at or below their skill level. Third, ED providers can form *ad hoc* teams where a provider with lesser skills can treat high-acuity patients by teaming with a higher-skilled supervising provider.

In this paper we develop a model and methodology for *staff planning* in emergency departments that accommodate these complexities to achieve target patient service levels while minimizing personnel expenses. A major contribution of our ED staff planning model (EDSP) is to incorporate the complex interaction of ED providers who have different skill levels and who work in *ad hoc* teams of skilled and less-skilled providers, all for the purpose of serving the highly variable demands of arriving patients with very different treatment needs. The development of our planning model was informed by data collected from three active emergency department facilities located in a metropolitan region of 2.5 million people – all HIPAA-sensitive data were removed from the database prior to its release for use in this study.

Following Cezik and L'Ecuyer (2008) and Robbins (2010), we define *staffing* to be the determination of the number and mix of skill levels to employ in any given time slot of a day; *scheduling* to be the two step process of first specifying a set of admissible shift schedules and then deciding on the number of providers who should be working during each of those shifts; and *rostering* to be the assignment of individuals to selected schedules. Using these definitions, our model combines ED *staffing* and *scheduling*, which we call *staff planning.* We demonstrate that ED staffing and ED scheduling cannot be separated from one another due to the interactions of provider qualifications and *ad hoc* team formation. The purpose of an ED staff plan is to determine the resources (medical personnel) that must be deployed in order to meet customer (patient) demand during aggregate blocks of time. The output of ED staffing is a plan that shows how many providers with specified qualifications are needed by hour over a planning horizon.

From an ED staffing plan, ED administrators can determine the appropriate mix of medical personnel, select appropriate master shift schedules, and maintain the most effective mix of ED medical personnel. Once these planning decisions are made and implemented, *rostering* decisions are then made on a weekly, biweekly, or monthly schedule at which time individual providers are assigned to shifts in the staffing plan, taking into account personnel vacation schedules, time-on time-off constraints, shift preferences, and all of the other factors which make rostering difficult (e.g., Beaulieu, Gendron, & Michelon, 2000). In this paper, we focus on ED staff planning (staffing and scheduling), not on rostering.

2.1.1. Supporting Literature

Work on ED staffing and scheduling is relatively sparse. Brunner, Bard & Kolisch (2009) note that compared to the large literature on nurse scheduling, relatively little research has been published on the problem of physician scheduling. Of papers devoted to ED physician scheduling, most presuppose that required provider staffing levels are known and therefore focus on creating a schedule or roster of personnel assignments (Beaulieu, Gendron, & Michelon, 2000; Carter & Lapierre 2001; Ernst, Jiang, Krishnamoorthy, & Sier, 2004; Ferrand, Magazine, Rao & Glass 2011). Similarly, there is a large literature on nurse rostering (Cheang, Li, Lim, & Rodrigues 2003), which also largely assumes that required staffing levels are known. In contrast, Green, Soares, Giglio & Green (2006) use queuing theory to determine staffing levels at EDs to reduce patient reneging.

There is a large literature on the broad topic of "hospital staffing." Some studies have used work sampling methods to assess physician activity and thereby decide on appropriate level of hospital staffing (Oddone, Guarisco, & Simel, 1993; Ben-Gal, Wangenheim, & Shtub 2010). While work sampling methods are a reliable tool to obtain good estimates of physician time requirements needed to treat a variety patients, they are not intended to create a planning schedule of providers and teams required to meet demand on an hour-by-hour basis. Lipscomb (1991) uses a different approach to estimate physician requirements, that uses both empirically based and expert judgment models for physician staffing. This approach is focused on determining the full-time-equivalent (FTE) employees of physicians that need to be on the payroll, and does not address the staffing needs required during particular shifts or hours. We are not aware of other research in the medical care planning literature that develops hour-by-hour staffing plans in the presence providers with different skill levels, *ad hoc* formation of provider teams, random arrival of patients with different acuity levels, and the necessity of meeting temporal service levels.

The last decade has seen a growth in research focused on call-center staffing and scheduling, which has some similarities with ED operations such as stochastic arrivals and stochastic service times. Cezik and L'Ecuyer (2008) examine the creation of staff schedules for large multi-skill call centers in which incoming calls are routed to groups of agents with specialized skills. Green, Kelesar, and Whitt

(2007) use queuing theory to study single-period staffing requirements (versus staff planning) in service systems where customer demand varies in a predictable pattern over time. They show how their results apply to a variety of service systems including call centers and emergency departments.

While there are many similarities between call-center and ED operations, there are limitations to applying call-center staff-planning methodologies to ED staff planning. First, providers in an ED often serve multiple patients simultaneously, unlike call-centers where an operator stays with a caller until the call concludes. Second, patients at EDs are often treated by a team of providers, where a less qualified attending provider is supervised by a more qualified provider, unlike call centers. An "attending provider" is the person assigned to treat an individual patient with appropriate medical care. These factors add complexity to ED staff planning and differentiate it from call center operations.

2.1.2. Contributions

Our chance-constrained EDSP model takes into account the hourly distributions of highly variable patient-care demand and makes the best staffing and scheduling decisions to minimize overall cost, while fulfilling service level constraints and scheduling rules. Our research makes several contributions. First, we show how patient workload demand by patient acuity level can be aggregated into discrete time buckets during an ED day to determine the stochastic distribution of aggregate patient demand within these buckets, which considerably improves the tractability of the ED scheduling problem. Second, our model permits ED administrators to effectively schedule providers with a range of skill-levels. In an ED setting, patient demand distributions apparent to medical providers are dependent upon their skill-level. ED providers can treat any arriving patient with acuity at or below their skill-level, which is similar to the nursing skill classes of Warner and Prawda (1974), a characteristic which Bard (2004) termed as "downgrading". Third, our model accommodates the use of ED provider teams, in which a provider with lesser skills can treat high-acuity patients by partnering with a more skilled provider. This adds complexity to the problem since the time demands placed on supervising providers must be accommodated within our model. Our results provide an optimal staff plan by provider type and facilitate

the creation of efficient and effective staff schedules. To our knowledge, we are the first to employ this approach in the context of ED staff planning and to staff planning in general.

2.1.3. Organization of Paper

The balance of the paper is organized as follows. In the next section, we formally define the ED staff planning problem and develop a chance-constrained mixed integer linear program model to solve it. In Section 3, we develop and employ a staff-assignment algorithm that makes patient-provider allocations depending on workload demand that arises, which demonstrates the efficacy of our model in developing staff plans that meet target service levels. Next, using a computational example, we investigate the model's sensitivity to several factors: scheduling horizon; supervision requirement; set of feasible shift schedules; and variability of service times. In Section 4 we discuss managerial implications of our analyses, including investigating the impact of limiting the available classes of provider, the results of enforcing a certain service level, and the consequence of allowing *ad hoc* provider teams to care for patients. We conclude in Section 5 with a summary of our contributions and a discussion of future research.

2.2. The ED Staff Planning Problem

In the next subsections we describe the ED staff planning problem (2.1), develop analytic notation to represent the problem (2.2), and construct an optimizing model to solve the problem (2.3).

2.2.1. Problem Description

We consider a typical emergency department where unscheduled patients arrive and are treated on a short-term basis only. Patients with conditions that require only short-term care are treated and released, while those with more serious or life-threatening problems are diagnosed, medically stabilized, and then are admitted or transferred to an inpatient hospital bed for ongoing care.

When patients arrive for treatment, they are first seen by a triage nurse, who assesses the severity and urgency of their condition and assigns them an acuity level, where higher acuity levels indicate

greater urgency. Waiting patients are usually treated in the order of their acuity, where higher acuity patients are seen first.

A patient with a particular acuity is treated by a qualified provider, assisted by ED staff. In the event of multiple patients arriving with acute and life-threatening conditions, providers often divide their time across critical patients, working on several patients simultaneously to stabilize each of them (Garmel 2005). In instances of extreme congestion, a single patient may be treated by multiple providers (Kachalia *et al* 2007). The workload requirements for patients of different acuity levels varies widely, with low acuity patients demanding less provider time with less urgency, while critically acute patients require much more provider time with great urgency. For this reason, we focus on *patient workload demand* by patient acuity level, derived from patient service times and their acuity levels.

The workload of a provider includes activities that provide direct and indirect care for patients, and exclude ancillary services performed by supporting staff. For example, examining a patient, interpreting test results, or performing a procedure all add to the provider's workload. In contrast, drawing blood, performing lab tests, and taking an X-ray are done by supporting staff and so are not counted in the workload of a provider.

Emergency departments are staffed by medical personnel with different skill levels, who cannot treat patients with acuities greater than their certified skills unless supervised by a provider with more advanced skills. For example, board certified Emergency Medicine (EM) physicians are specifically trained to treat and stabilize even the worst cases of trauma and illness that arrive at an emergency department and so can handle any acuity level. When working without supervision from a higher qualified provider, Family Practitioners (FPs) and General Practitioners (GPs) can treat patients up to a midrange acuity level, and Physician Assistants (PAs) can only treat patients assessed with the lowest acuity levels. The cost of employing a medical provider typically increases with skill level and training. In any given hour, the set of providers on station have an aggregate *workload capacity* that set a limit on the service that can be provided to each patient acuity level.

In order to provide satisfactory service to its patients, the ED stipulates a target *service level* based on experience and observation. This service level is the probability that provider *workload capacity* will meet or exceed arriving patient *workload demand* during any given hour of the day, and is similar in concept to the service levels employed in classic newsvendor models (Hillier & Lieberman 2001, 965-967). Note that setting an ED service level does not mean that patients are abandoned or turned away. Patients rarely leave untreated from an ED unless they choose to leave on their own (Green *et al.* 2007). Instead, on those occasions where patient demand exceeds physician capacity, low acuity patients may need to wait for one or more periods until medical personnel become available to treat them. In extreme cases of excessive demand, arriving patients may be diverted to other EDs or off-duty providers may be called in on an emergency basis. Also, we assume that beds are available for ED patients who need to be admitted to hospital and that a lack of beds in the ED itself is not a constraint on ED staff planning.

To facilitate the creation of staffing plans, the ED maintains a set of admissible "shifts" to which its provider personnel may be assigned, such as an eight-hour shift starting at 7:00 AM, a ten-hour shift starting at 9:00 PM, and so forth. To generate a staff plan, ED administrators specify solid blocks of time (say 10 hours) without scheduled breaks or lunch hours, which are taken during slack periods, which mean that the number of planning shifts in an ED may be relatively small compared to other settings such as call centers. Other constraints such as required days off, personal scheduling preferences, vacation and personal time off, and similar short-term constraints are accommodated in a weekly or biweekly rostering process, which implements the staffing plan generated by our model.

The ED staff planning problem is two-fold. First, the distribution of patient workload demand by acuity must be estimated based on historical data or other estimates. Second, ED providers must be assigned to admissible shifts so that service levels are met throughout the planning horizon and costs minimized, all while taking into account the training qualifications and teaming abilities of the various medical staff providers.

2.2.2. Notation

The ED divides time into N time-slots or time-buckets of an appropriate duration d (perhaps an hour or half-hour) so that $H = Nd$ is the time horizon of the scheduling problem for the ED. Arriving patients are first examined by a triage nurse who makes initial evaluations of the severity of their conditions and assigns each an acuity level ℓ L used by the ED, where higher values of ℓ represent increasingly urgent patient conditions requiring more immediate attention.

We bifurcate the formal problem definition into two parts. The first examines hourly arriving patient workload demand by acuity level and determines minimum provider capacity required to meet ED target patient service levels. The second part defines the combined provider staffing and scheduling problem confronting emergency departments. We show in the subsequent subsections that for EDs with multiple provider skill levels and with the ability to form *ad hoc* teams, both provider staffing and provider scheduling must occur simultaneously, since the interaction between provider capacities, multiple skill levels, and teaming cannot be decomposed.

Table 2-1. Notation Summary

Parameters	
N	Number of time-slots
d	Duration of each time-slot
$H = Nd$	Time horizon of scheduling problem
L	Number of acuity levels
ℓ	Index of Acuity Level, from 1 to L; higher acuity indicates higher urgency of patient condition
ℓ'	Lowest training-level of provider who, when supervised, may attend patient of acuity ℓ
$\hat{\ell}$	Highest acuity-level that an ℓ trained provider may attend when supervised
$\alpha_{i\ell}$	Supervision requirement when acuity ℓ is treated by an i-trained provider; expressed as a fraction of attending work content
$w_{t\ell}$	Random work content of attending to ℓ-acuity patients during time-slot t
$z_{t\ell}$	Random work content of attending to all patients of acuity ℓ or greater; follows distribution $G_{t\ell}(z_{t\ell})$
$G_{t\ell}$	Distribution of $z_{t\ell}$ (work-content of attending patients of acuity ℓ or above during time-slot t)
ρ	Target service-level
$\pi_{t\ell} = G_{t\ell}^{-1}(\rho)$	Attending staff capacity required to meet service level ρ for patients of acuity ℓ or above during time-slot t
$\mathbf{Q}$	Matrix representing admissible shifts
q_{jt}	Binary element of $\mathbf{Q}$ matrix; 1 if shift j includes time-slot t, 0 otherwise
Q	Number of available shifts
$c_{t\ell}$	Cost of employing an ℓ-trained provider during time-slot t
$C_{t\ell}$	Cost of employing an ℓ-trained provider on shift j
Variables	
n_{ℓ}	Number of ℓ trained providers assigned to shift j
$p_{t\ell}$	Number of ℓ trained providers on-duty in time-slot t
$x_{i\ell t}$	Work content assigned to i- trained providers, treating patients of acuity ℓ during time-slot t

2.2.2.1. Patient Demand

We first examine patient demand distributions for each time period *t* of a clinic, and define w_ℓ as the random work content required of attending providers to meet the service needs of ℓ-acuity patients in period *t.* Without loss of generality, we express work content as a fraction of the duration of a time slot. For example, if w_ℓ ^ 5, then acuity-ℓ patients would require the attention of a qualified provider for 50% of a slot duration. We assume that the distribution F_ℓ (... ℓ) and density f_ℓ (... ℓ) of work content are known from historical data or can be estimated from ED patient arrival and treatment records. Based on conversations with our reference ED, we assume that the required work content of an acuity-ℓ patient is the same regardless of the training level of the attending provider.

Since patients of acuity ℓ can be served by any ℓ-qualified provider or provider team (level ℓ or greater), it will be useful to define z_ℓ as the random work content of all patients of acuity-ℓ or greater in period *t* :

$$z_\ell \quad \sum_{J=\ell}^{L} w_{it} \tag{1}$$

with distribution G_ℓ (− ℓ) and with density g_ℓ (− ℓ). The latter is the convolution of the work content densities f_ℓ (... ℓ) for individual acuities levels greater than or equal to ℓ in period *t* :

$$g_\ell (- _\ell) \quad f_\ell (... _\ell) * f_\ell \quad (... _\ell) * \quad * f_{Lt}\left(w_{Lt}\right) \tag{2}$$

where $*$ is the convolution operator defined as $\left(f_a * f_b\right)(x) \doteq$ $)f_b\left(x-y\right)dy$, which assumes that random variables w_ℓ are independently distributed. This combined pool of work content across acuity levels provides the joint distribution of work content for patients with acuity level ℓ and above, and represents the combined workload for all providers who are at least ℓ-trained.

In the case of teams, the work content required of an attending provider treating an ℓ-acuity patient is w_ℓ, the same as that for individual providers. However, the supervising physician also must allocate time to work with the attending provider, and the supervising time required will depend on both the training level of the attending provider and the acuity of the patient. Based on observations of our reference EDs, we assume that the oversight time requirement of a supervising physician, inclusive of the inefficiency of "catch-up" time, is $\alpha_{i\ell} w_\ell$, where $0 \le \alpha_{i\ell} < 1$ is a fixed fraction of the attending provider's work content w_ℓ, index ℓ is the acuity of the patient, and i is the training level of attending provider. Thus, in case of a team provider, the total labor requirements are given by $w_\ell (1 + \alpha_{i\ell})$.

Since acuity-ℓ patients can be served either by ℓ-qualified providers or by ℓ-qualified teams, define intermediate variable $x_{i\ell}$ that holds the attending work content assigned to each feasible attending provider class, where index i is the training level of the attending provider, ℓ is the acuity level of the patient, and t is the time period. When $i < \ell$, i.e. when provider training-level is lower than the acuity of the patient being treated, supervision by an ℓ-qualified provider is required. Otherwise, when $i \ge \ell$, the attending provider works independently. We use ℓ to denote the lowest training level of provider who, when supervised by an ℓ qualified provider, may attend to a patient of acuity ℓ, where $\ell \quad \ell$. Conversely, when supervised by an adequately-qualified provider, the highest patient acuity level that an ℓ qualified provider can attend to is denoted by ℓ, where $\ell \quad \ell$. While ℓ and ℓ are both dependent on ℓ and can therefore be expressed as $\ell \ \ell$ and $\ell \ \ell$ respectively, we use the notation ℓ and ℓ for simplicity.

2.2.2.2. Provider Staffing and Scheduling

An ED is usually staffed with medical providers of different skill levels, where each provider has been trained to a skill level ℓ corresponding to acuity level ℓ. An individual provider with skill level ℓ can independently treat arriving patients with acuity levels of ℓ or below, but not above. It will be convenient

to define "ℓ-trained" providers as those with the maximum skill-level of ℓ, and "ℓ-qualified" providers as those who have a skill-level of ℓ or higher and who thereby can treat acuity-ℓ patients.

Providers can also work in teams where "attending" providers with skill-levels less than ℓ work with ℓ-qualified "supervising" providers to service acuity-ℓ patients. In the case of teaching hospitals, residents may serve as attending providers under the supervision of experienced supervising physicians (Sox *et al.,* 1998). Supervising providers usually spend much less time with patients than attending providers, but the total time spent by a provider team is greater than that required by an ℓ-qualified provider working alone. We show that the use of provider teams affords an ED with greater flexibility and lower costs than using single providers alone, even though total provider time efficiency declines (Section 4.1).

Providers are assigned to work predetermined shift schedules, which can take into account minimum and maximum work durations, work rules, labor laws, and so on. In the standard way of constructing personnel scheduling models (for example, Hillier & Lieberman 2001, 57-59), we assume that ED administrators have previously identified Q feasible or admissible shifts that together form the Q rows of shift matrix **Q,** in which its binary elements q_{jt} indicate whether or not time slot t is included in shift j:

$$q_{jt} = \begin{cases} 1 & \text{if time period } t \text{ is included in shift } j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Denote n_ℓ as the number of ℓ-trained providers assigned to shift j, so the number of ℓ-trained providers on duty in time slot t is p_ℓ [illegible]. Define C_ℓ [illegible] to be the cost of employing one ℓ-trained provider on shift j.

2.2.3. An ED Staff Planning Optimization Model

Our approach to modeling the ED staffing problem is to employ chance constraints (Charnes and Cooper 1959; Henrion 2004; Birge and Louveaux 2010, Chapter 3) to insure that service levels are met for each

acuity-level of demand. Let □ be the target service level exogenously established by the ED, where service level is defined as the probability that workload of arriving patients in a given time period does not exceed the workload capacity of providers on station. The service level constraint of our ED staff planning problem can then be expressed as follows (Henrion 2004, Birge and Louveaux 2010):

$$\Pr\left(\sum_{i=\ell}^{L}\sum_{\ell}^{L} x_{ikt} \geq \xi_{\ell}\right) \geq \rho \qquad \forall \ell \qquad , t \in \{1...H\} \tag{4}$$

where ξ_{ℓ} is a random variable representing the actual work content of arriving patients with acuity greater than or equal to ℓ in period *t,* and $0 \leq \rho \leq 1$ is the target service level set by the ED administration. This constraint asserts that the capacity of qualified providers must be at least equal to arriving work content ξ_{ℓ} with probability ρ. If an ED wishes to vary target service levels by time of day, then target ρ can simply be subscripted with time slot index *t.*

Using work arrival distribution $G_{\ell}(\cdot)$ developed earlier, we can rewrite constraint set (1) as the "deterministic equivalent" set that captures the stochastic requirements of a minimum service level (Henrion 2004, Birge and Louveaux 2010):

$$\Pr\left(\sum_{i=\ell}^{L}\sum_{\ell}^{L} x_{ikt} \geq \xi_{\ell}\right) \geq \rho \qquad \sum_{\ell}^{L}\sum_{\ell}^{L} \geq G_{\ell}^{-1}(\rho) \quad \forall \ell \qquad , t \in \{1...H\} \tag{5}$$

The decision variables of the ED staffing problem are the number of ℓ-trained medical providers n_{ℓ} to assign to each admissible staff schedule $j \in \mathbf{Q}$. Mathematically, the ED staff planning problem is expressed as:

$$\text{Min } C = \sum_{\ell}^{L}\sum^{Q} C_{\ell} \tag{6}$$

Subject to

$$p_{\ell} \qquad \sum_{J=1}^{Q} \qquad \geq \qquad \forall \ell \qquad , t = \{1...H\} \tag{7}$$

$$\sum_{i=k'}^{L}\sum_{k=\ell}^{L} x_{ikt} \geq G_{\ell}^{-1}(\quad) \qquad \forall \ell \qquad , t \in \{1...H\} \tag{8}$$

$$\sum_{i=\ell}^{L} p_{it} \geq \sum_{\ell}^{\hat{i}}\sum_{\ell}^{L} x_{ikt} + \sum_{\ell}^{k-1}\sum_{\ell}^{L} \alpha_{ik} x_{ikt} \qquad \forall \ell \qquad , t \in \{1...H\} \tag{9}$$

$$p_{Lt} \geq 1 \qquad \forall t \in \{1...H\} \tag{10}$$

$$x_{i\ell} \geq 0;\ p_{\ell} \geq 0 \text{ and integer} \tag{11}$$

The objective of the problem is to minimize total staffing costs C for all provider levels and shift assignments for a given target service level ρ. Constraint set (7) defines p_{ℓ}, the number of ℓ-trained providers available in time-slot t. Inequalities (8) are chance constraints which ensure that there are a sufficient number of qualified attending providers available in every time slot to meet service level targets for each acuity level. It represents the ability of highly trained medical personnel to provide service to patients with lower acuity levels. For example, this constraint set ensures that there are adequate ℓ-qualified personnel to serve acuity-ℓ patients, but also permits ℓ-qualified personnel to serve lower acuity patients as well.

Constraint set (9) ensures that a sufficient number of ℓ-qualified providers are assigned to work in each time period. The first term of the constraint is the sum of the work-content for ℓ-qualified providers working as attending physicians in a team, or as individual providers when $k \leq i$. The second term is the sum of the work-content for ℓ-qualified providers working as supervising providers. Constraint set (9) in conjunction with constraint set (8) demonstrates why the problems of ED staffing and ED scheduling cannot be decomposed. While it is necessary to have attending staff capacity to meet the service levels (constraints 8), it is also essential that qualified supervising personnel are available (constraints 9).

Constraints (10) require that a provider of highest qualification, e.g. an MD physician trained in emergency medicine be scheduled at all times, which is a requirement in most emergency departments. Since high acuity patients can only be treated by ED physicians, constraints (10) will be automatically

satisfied by constraints (8), but is included in the model for transparency. Finally, constraint set (11) enforces the integrality and non-negativity of the number of providers scheduled. Depending on specific scheduling rules or labor laws followed at a particular ED, there may be additional constraints included.

Note that p_ℓ is always integer and non-negative since it is defined to be the product of two non-negative integers; hence it is not necessary to explicitly declare its integrality in the problem formulation. Variables $x_{i\ell}$ represent the workload from acuity ℓ patients assigned to attending providers of skill-level *i,* in time period *t.*

When modeling a stochastic problem, the use of a chance-constrained program is appropriate, versus other stochastic methods, when the following conditions are met (Birge & Louveaux 2011, Ben-Tal et al. 2009, Henrion 2004):

1. *Constraint violations cannot be avoided with certainty.* In EDs, demand beyond capacity can never be entirely prevented due to normal variation in demand patterns, and especially due to the possibility of larger-scale emergency situations such as fires, multi-vehicle auto accidents, industrial accidents, and other incidents that involve multiple victims.
2. *Compensation or recourse does not exist.* Emergency departments have very limited recourse to adjust capacity when demand exceeds supply during normal operations since demand arrives unannounced and short term staff adjustments are usually not possible.
3. *It is difficult or impossible to assign costs to violated constraints.* The costs of ED services delayed or denied are not usually monetary, but are measured in sickness, injury, and lives. These costs are difficult to represent in an optimizing model.
4. *Approximate distribution(s) of the random parameter(s) are available.* In the case of our three reference EDs, we have abundant demand data for over two years, which provides us with the necessary approximations of hourly demand distributions for the six acuity levels of arriving patients.

An alternative approach to the EDSP problem might be to develop a two-stage stochastic programming model (for example, Campbell 2011 and Easton 2011) where staffing and scheduling constitute the first stage decision, and allocation of providers to patients is made upon realization of demand. However, we determined that a two-stage stochastic programming approach would require more constraints, more variables, and new binary variables leading to an unnecessarily complex model requiring excessive computational resources. Further, model fidelity would be severely diminished since we would be limited to a small number of scenarios relative to the large solution space represented by the joint probability of the six patient acuity demand streams. And finally, it would be difficult to identify representative scenarios that capture low-probability high-demand values with precision.

2.3. A Computational Example

To investigate the practical utility of our model for ED staffing, we developed an example problem abstracted from data gathered over two years from three active general emergency departments in a metropolitan region of 2.5 million people. This example problem is not intended to replicate any particular ED, but is representative of typical ED operations and serves to illustrate how our model can be used to inform ED practice and administration. More specialized EDs, such as pediatric units, may have different patient demand and provider operating characteristics than those of our reference EDs. We use this example to validate our ED staff planning model, to perform sensitivity analyses, and to develop managerial insights valuable for ED administrators.

2.3.1. Computational Design

Statistical analysis of arriving patient demand data for our reference EDs shows that workloads vary significantly throughout the day (Figure 2-1a), but that variability across days of the week and across months of the year is insignificant. Consequently, our initial computational example focuses on scheduling a single day, so time horizon H was chosen as one day (24 hours) – a longer horizon is investigated in Section 3.2. The duration of a scheduling time slot d was set to be 1 hour (d=1), which comports with the scheduling practices of the reference ED. We note that other EDs report significant

demand variation by day-of-week (Moloney et al, 2006; and Leo & Bove, 1996) and that our EDSP model is not limited by any assumptions regarding patterns of workload demand.

Figure 2-1. Two Year Average of Patient Arrivals by Acuity Level (A1-A6)

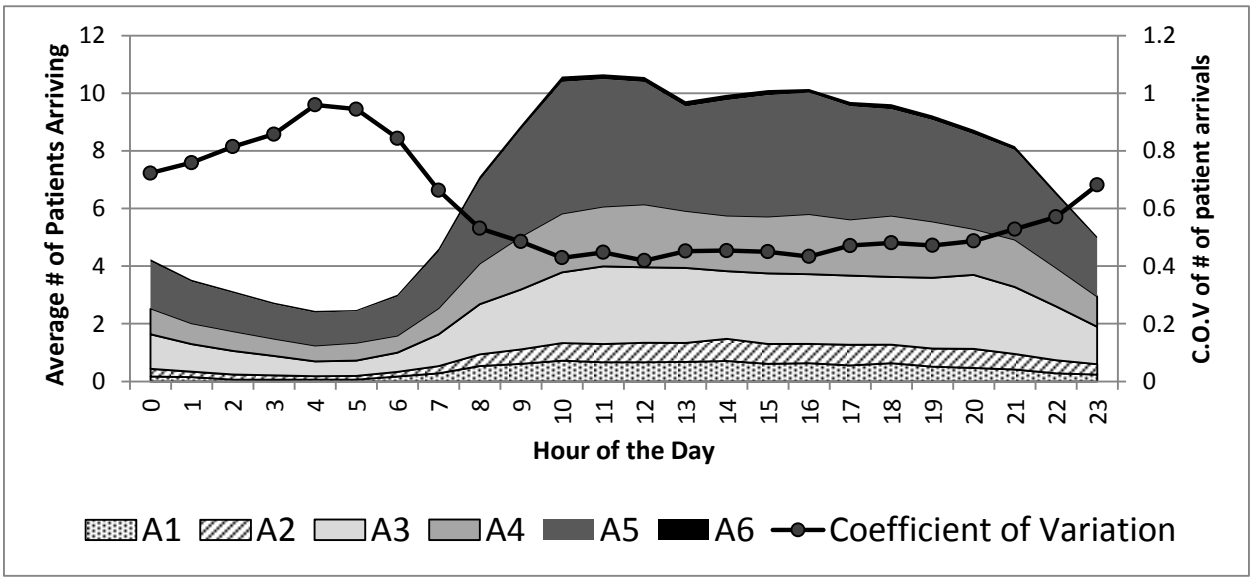


Fig 2-1a – Arrivals by acuity level and hour, with coefficient of variation

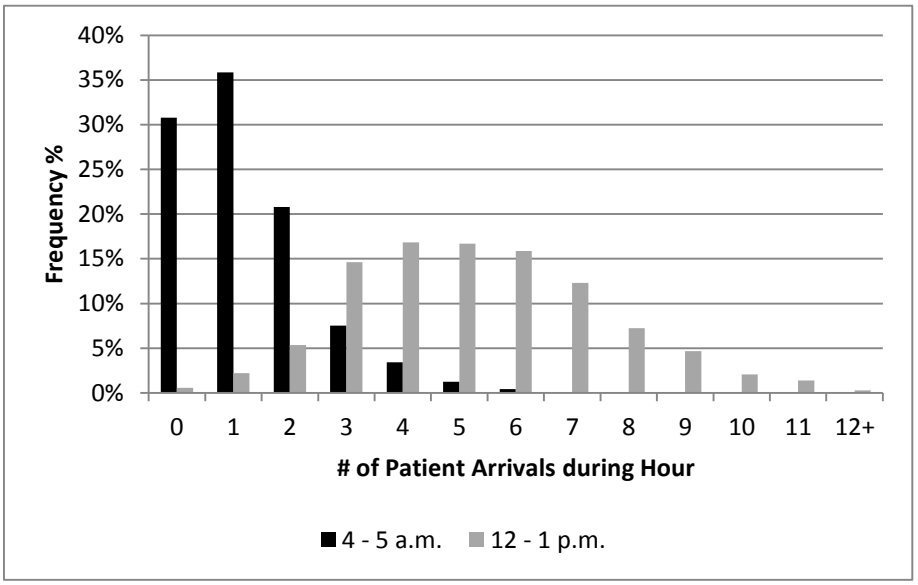


Fig 2-1b – Distribution of arrivals for two representative time blocks

In our example, patients arrive throughout the day and are sorted into one of 6 acuity levels (A1, A2 … A6) depending on the severity of their conditions (Figure 2-1), where the demand for care during each hour follows an empirical distribution obtained from historical demand during that hour. Interestingly, the coefficient of variation of patient arrival tends to be higher during the less busy hours

(early mornings), and lower during the busy hours (midday and afternoon) (Figure 2-1a). Figure 2-1b compares the frequency distribution of patient arrival during the least and most busy hours. This observation suggests that it will be relatively more expensive to fulfill service level during the midnight and early morning hours, and for a uniform service level enforced around the clock, providers working during the lean hours will enjoy more idle time.

Once admitted, patients require care from medical providers for amounts of time that depend on their acuity level. Our example ED employs medical providers with six skill levels (P1, P2 … P6), each with a different hourly cost **(**Table 2-2**)**. A provider of a certain skill level is limited by the highest patient acuity level that s/he is qualified to attend without supervision.

Table 2-2. Patient and Provider Characteristics

Patient Acuity	Min Rqd Provider Skill Level	Avg Provider Hours Rqd/Patient	Relative Provider Cost per Hour
A1	P1	0.13	15.0
A2	P2	0.13	30.0
A3	P3	0.30	50.0
A4	P4	0.42	66.5
A5	P5	0.42	83.5
A6	P6	0.58	100.0

In the case of teams, a provider who is not ℓ-qualified may still attend to an acuity-ℓ patient if partnered with an ℓ-qualified supervisor. The time required by the attending provider is the same as for an independent provider (Table 2-2), and the time required of the supervising provider is a fraction of the attending provider's time (Table 2-3). These figures were derived from estimates provided by physicians associated with the EDs. Sensitivity of the model's result to the estimates of supervising time requirement is investigated in Section 3.4.

Table 2-3. Provider Supervision Hierarchy

Attending Provider Type	Patient Acuity	Supervision Required	Can be Supervised by
P1	A2	10%	P2, P3, P4, P5, P6
P1	A3	30%	P3, P4, P5, P6
P1	A4	50%	P4, P5, P6
P2	A3	10%	P3, P4, P5, P6
P2	A4	30%	P4, P5, P6
P2	A5	50%	P5, P6
P3	A4	10%	P4, P5, P6
P3	A5	30%	P5, P6
P3	A6	50%	P6
P4	A5	10%	P5, P6
P4	A6	30%	P6
P5	A6	10%	P6

Empirical data gathered from our reference EDs showed that the average service-level across all acuities and all hours was $\rho = 0.97$, which is the service-level we therefore chose to use in our analysis. Empirical workload distributions were developed for each of 6 acuity levels for each hour of the day using two years of data (730 data points per distribution) collected from the reference ED.

Workload distributions $G_{\ell}(z_{\ell})$ were numerically calculated by accumulating observed workloads for each hour of the day over two years. For example, workload distribution $G_{3,8}(z_{3,8})$ was determined by observing all arrivals in the 8th hour of 730 days, adding the arriving workloads for patients with acuities of 3, 4, 5, and 6 in each of those days to determine aggregate workload z_{ℓ} (equation 1), and then sorting the results to form an empirical representation of workload distribution $G_{3,8}(z_{3,8})$. This distribution was then consulted to determine the 97th percentile workload that would need to be serviced to provide a 97% service level. While we were fortunate to have an abundance of empirical data to use, with more sparse data, other methods such as queuing analysis or Monte Carlo simulation could be employed to estimate workload distributions $g_{\ell}(z_{\ell})$.

Based on the acceptable shift schedules identified by our reference ED, we studied shifts of three different lengths – 8, 10 or 12 hours, beginning at the start of any hour between 6 a.m. and 12 midnight; a

total of 57 allowable shifts in a 24 hour day. Shifts starting later in the planning horizon wrap around to cover early hours; e.g. for a 24-hour planning horizon, an 8 hour long shift beginning at 11 p.m. provides coverage up to 7 a.m. The problem of the ED is to develop a staffing plan that will minimize total ED provider costs. We note that our model is not limited by the construction of allowable shift schedules, which could include required breaks, maximum shift lengths, minimum days off, and other staff planning constraints for a particular ED.

We analyzed the resulting example problem using a commercial constrained-optimization solver running on a fast research server.[1] For problems without teams (where providers always work alone), optimal schedules were identified within 1 second of computing time. Computation increased significantly for problems with provider teams, indicating the complexity that teaming brings to the EDSP problem. For these problems, feasible schedules with an optimality gap of less than 5% were obtained within 7 minutes. On test problems, additional runtimes of up to 5 hours reduced the optimality gap to about 3%, suggesting that optimal or near optimal solutions are found early, but proving optimality is difficult and time-consuming. We therefore limited the runtime of our experiments to 15 minutes, which provided optimality gaps ranging from 0% to 4% withan average gap of 2%.

We conjecture that the slow convergence for problems involving teams is attributable to degeneracy and alternate optima caused by work allocation variables $x_{i\ell}$. Given our focus on improving ED performance and informing managerial decisions, we considered provably good (near-optimal) schedules to be satisfactory for the purposes of aggregate staff planning. We will study techniques for improving the computational performance of our model in future work using proven methods such as branch and price (Brunner, Bard, & Kolisch 2010) and column generation (Brunner & Edenharter 2011).

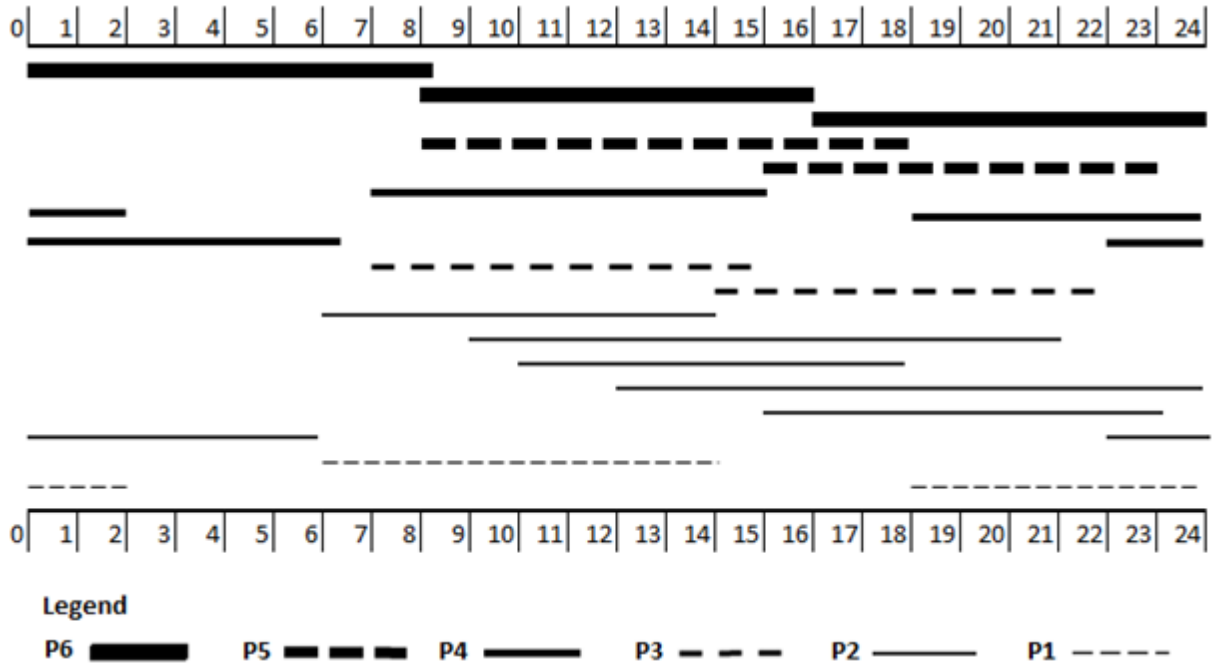
2.3.2. Computational Results

Computational results indicate that our model can improve staffing schedules over the incumbent schedules used by the reference EDs, by lowering staffing costs while maintaining consistently high

[1] CPLEX 11.2.1 solver running on a Windows Enterprise Server with 3.16 GHz dual processors and 32 GB RAM.

patient service levels (Figure 2-4). For the example problem, the optimal staffing schedule uses 15 of the 57 allowable shifts (Figure 2-2).

Figure 2-2. Staff Schedule for Example Problem



Scheduled providers spend their time attending to patients, supervising other providers, or are idle (Figure 2-3). The most qualified providers spend more active time in supervising duties (88% for P6 providers), while less qualified providers spend most of their active time attending to patients (100% for P1 providers). All provider types are idle (*i.e.*, not attending to patients or supervising other providers) for much of their assigned shift. This can be attributed to the high service levels required in emergency departments – we investigate the sensitivity of the model to ED service levels in Section 4.3.

When tested against the 17,520 hours of sample demand data used to build the model, the optimal staff schedule meets or exceeds the 97% service level target for each of the 24 hours of the day (Figure 2-4) for every acuity level, as required by the underlying model. In contrast, average service levels observed in the reference EDs vary widely from a low of 77% in early morning to almost 100% in the afternoon and evening. The optimal staff schedule would thus allow the ED to operate at a more consistent level of service and with a more even workload throughout the day.

Figure 2-3. Time Distributions of Labor by Provider Level

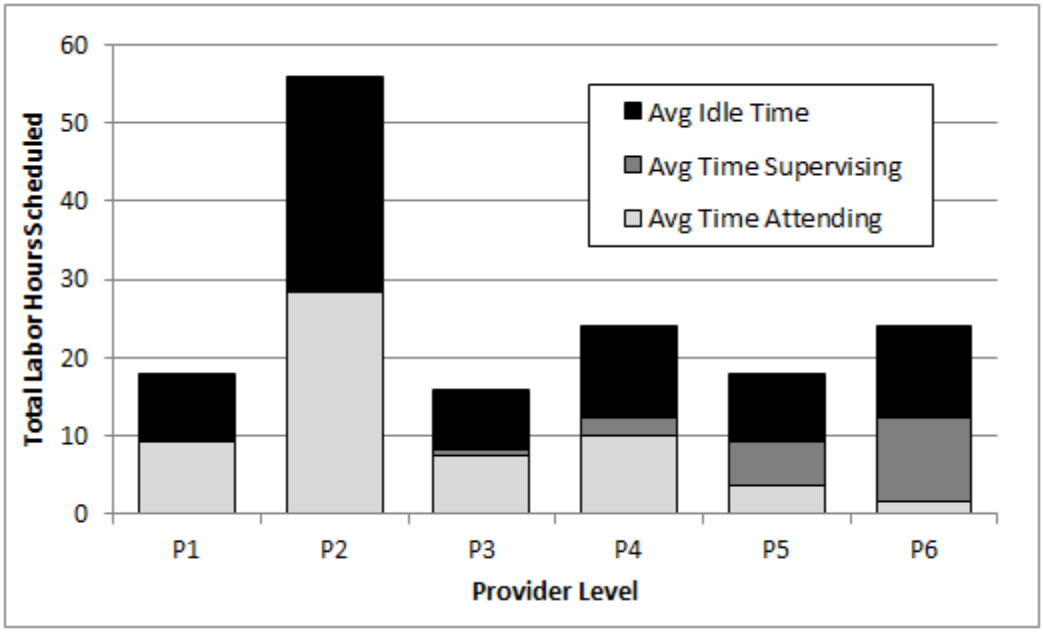
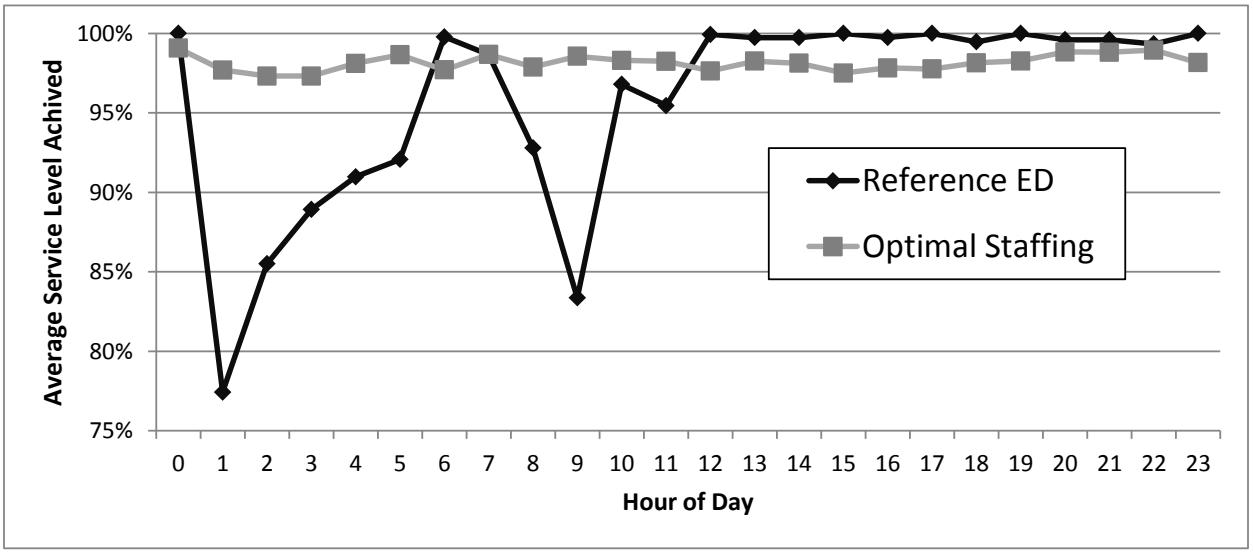


Figure 2-4. Reference ED vs. Optimal Schedules – Average Service Levels Achieved by Hour



2.3.3. Model Validation

To validate our results and to facilitate further managerial analyses, we constructed a two-stage staff assignment algorithm, which first determined the optimal ED staffing plan using the model developed in this paper and then assigned providers and teams to arriving patient workload for each hour of the planning horizon (details in Appendix 2.1). The algorithm demonstrates how a staffing plan can be

implemented in practice to accommodate several scheduling issues that are not incorporated in the staff planning model: (1) the order in which waiting patients are seen by providers, and (2) what to do when demand exceeds provider capacity. First, the algorithm accords treatment priority to high acuity over low-acuity patients, which is the order in which patients are treated according to our reference EDs. Second, the algorithm calculates workload within an hour as the sum of newly arriving workload plus any unfulfilled demand from the prior hour, so that unmet demand "spills over" from one hour to the next. We assumed that hourly aggregate patient workload by acuity level was known at the start of each hour, consistent with the time-granularity of our staff planning model.

The assignment algorithm was tested on two data sets, both with a target service levels of ρ = 97%. The first data set was the two years (17,520 hours) of workload demand data used to develop the optimal staff planning schedule. The second was a "hold-out" sample of 1,680 hours (10 weeks) of data, which immediately follows the first data set in time. The staff assignment algorithm achieved a service level of 98.69% for the first data set and a service level of 98.37% for the hold-out set. We conjecture that the reason that service levels were higher than the 97% target is due to the integrality constraints on scheduled providers, which provides some slack capacity beyond the target service level. These results suggest that our staff planning model achieves its intended result of constructing a staff plan with capacity to meet or exceed arriving workload demand for a specified level of service. Further discussion of these results is included in Section 3.4.4 below.

2.3.4. Model Sensitivity

Following the main computational experiment, we undertook several additional studies to understand the sensitivity of our model to its assumptions and parameters, including scheduling horizon, supervision time requirements, duration of shift schedules, and service time variability.

2.3.4.1. Sensitivity to Scheduling Horizon

A planning horizon of one day was used in the base case example, but staff planning typically occurs over longer horizons. To understand the sensitivity of our model to scheduling horizon, we tested a week-long

schedule and compared the results to a daily schedule. Using our two-year base data set, we recalculated the workload distributions for each hour of the day for each day of the week using 104 empirical workload demand observations for each distribution. Minimum service levels $G_{\ell}^{-1}(\quad)$ were recalculated for each of the 168 hours in a week, and in the model, inequality set (6) was expanded from 24 to 168 constraints, and the problem was rerun.

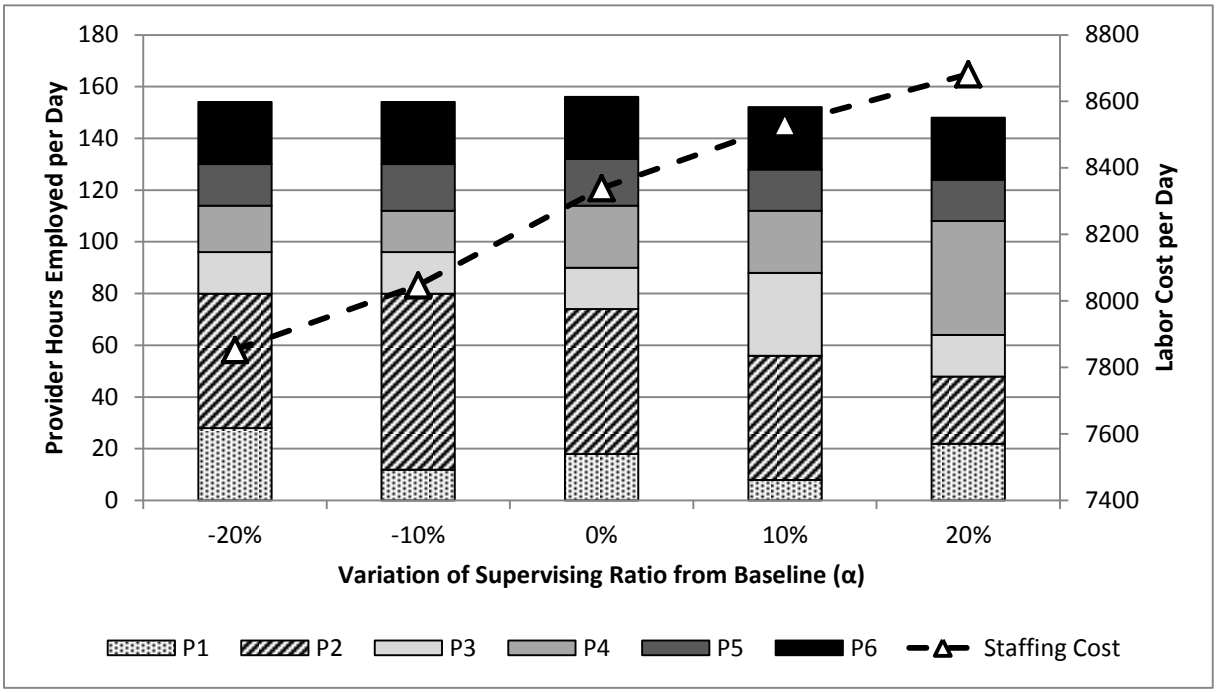
Staffing costs for the 7-day schedule were about 2% lower compared to the 1-day schedule, because a longer planning horizon (e.g. weekly instead of daily) provides more degrees of freedom to find schedules which meet the stipulated service levels. If there are variations in demand across days of week, then on a per-day basis, staffing costs may be lower for weekly schedules than for daily schedules. These results indicate that scheduling costs are relatively robust to changes in scheduling horizon, at least for our example data set, which suggests ED administrators can choose a scheduling horizon that fits the personnel needs and policies of the ED without fear of significant cost penalties. Because our staff planning model does not include rostering, here we are assuming that depending on the providers employed by the ED, there are feasible rosters that fit the recommended shift schedules and conforms to requirements of rest periods and days off.

2.3.4.2. Sensitivity to Required Supervision Time

Supervision time parameters (α) used in our base experiments were based on estimates provided by physicians associated with our reference ED. To investigate the sensitivity of staffing costs to α values, we experimented by varying the α parameters by up to 20%.

Lower values of supervision ratio α make team-providers more cost efficient, engages more of the lower qualified providers, and hence allows for lower staffing costs (Figure 2-5). For a 10% deviation in supervising ratio, staffing costs were impacted by a corresponding change of about 2% to 3%, which suggests that staffing costs are not overly sensitive to relatively minor errors in estimating the supervising ratios.

Figure 2-5. Sensitivity to Supervising Ratio (α)



2.3.4.3. Sensitivity to Shift Schedules

To understand how the set of shift schedules **Q** impact ED staffing costs, we ran our model with 24 1-hour shifts to represent perfect flexibility in scheduling providers and teams in order to provide a lower bound on staffing costs. While impractical to implement, this perfect flexibility provides an indication of the costs imposed on staff planning constraining flexibility with a finite set of shift schedules. With perfect shift flexibility, total staffing costs were reduced by 3.5% compared to the best staffing plan using 57 allowable shifts. This relatively small reduction in costs provides assurance that the practical necessity of imposing shift schedules on the system does not cause dramatic declines in cost performance, at least in this instance.

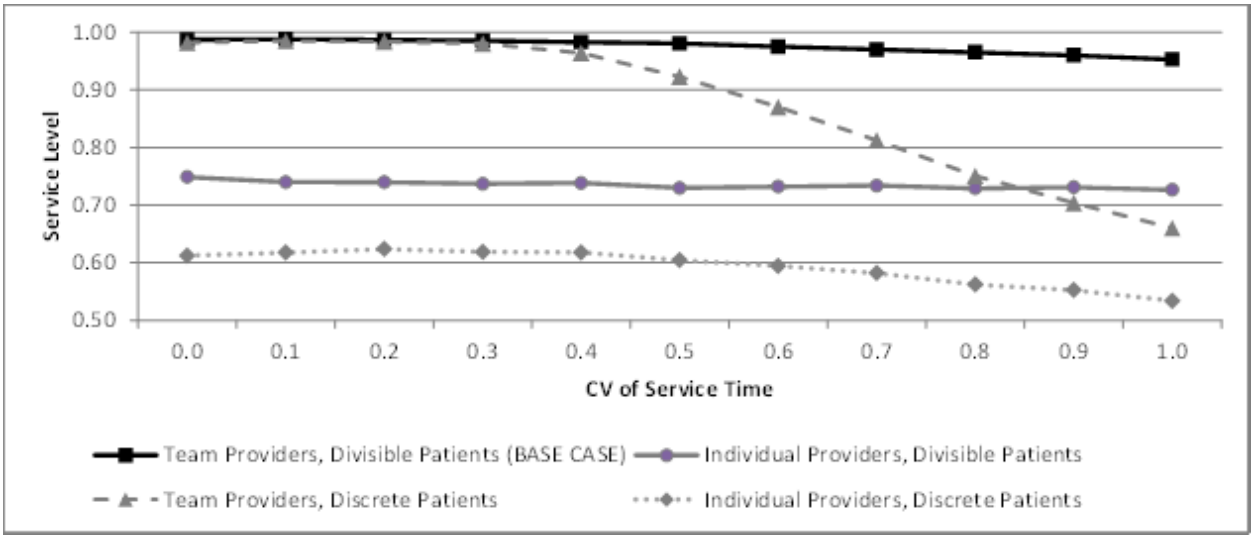
2.3.4.4. Sensitivity to Service Time Variability

Since our base model assumes that the workload for each acuity class is known with certainty, we investigated the sensitivity of our model to variability of service times with a simulation that used the staff assignment algorithm introduced in Section 3.3. The simulation started with a 97% service level

staffing plan developed using deterministic service times (Table 2-2). The staff assignment algorithm was then run against this staffing plan using stochastic service times randomly drawn from gamma distributions with mean service times equal to the deterministic service times. The coefficient of variation (*cv*) of the gamma distribution was varied from *cv*=0.0 to *cv*=1.0 in increments of 0.1, where *cv*=0.0 corresponds to deterministic service times and *cv*=1.0 corresponds to exponentially distributed service times. A gamma distribution was chosen to model service times because its range is strictly non-negative and because it is skewed to the right, which allows for the possibility of very long treatment times.

Our results indicate that the realized service levels are quite robust to variability in service times. For example, the base-case staffing plan (*cv*=0.0) with deterministic service times realized a service level of 98.7%, while with exponential service times (*cv*=1.0), a service level of 95.3%, was achieved (Figure 2-6). For all experiments with *cv* < 0.7, the target service level of 97.0% was achieved despite significant service time variability, since the longer than usual service times required by some patients are compensated by the shorter than usual service times required by other patients.

Figure 2-6. Impact of Service Time Variability



As Figure 2-6 shows, the benefit of using teams and the need for multiple providers attending to a single patient is impacted differently by service time variability. As service time variability increases beyond *cv*=0.4, the service level attained by teams attending to individual (nondivisible) patients declines

rapidly in comparison to the other three cases, suggesting that with high service time variability, achieving high service levels is possible only if individual patients are allowed to be treated by multiple providers and/or teams.

2.4. Managerial Analysis

In addition to generating cost-minimizing schedules that meet patient service level targets, our model also offers important managerial insights into the effective and efficient administration of emergency department operations. In this section, we employ our model to examine three significant issues in ED management: whether or not to use provider teams; restricting the number of provider types; and the impact of patient service levels on staffing costs. We conclude the section with a discussion of implementation issues.

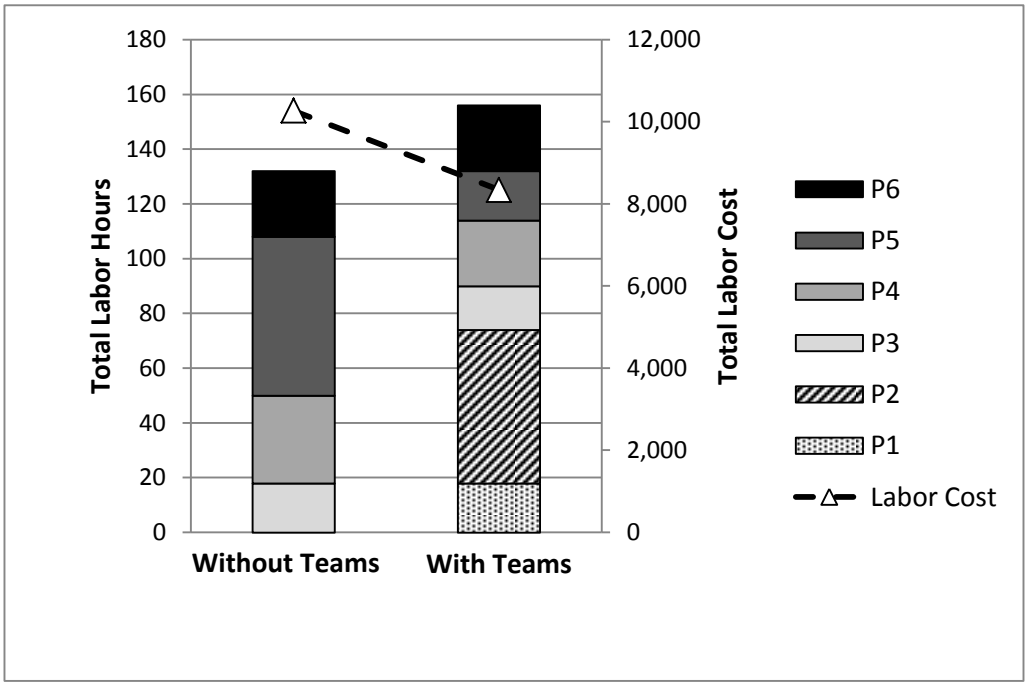
2.4.1. Teams vs. No-Teams

To understand the impact of teams on the operations of an ED, we compared the cost of the base case schedule (with teams) against the cost of a schedule without teams (Figure 2-7). Without teams, total labor hours are 30% lower than schedules with teams, but counter-intuitively, total staffing costs are 20% higher. This can be explained as follows. Labor hours are lower for the no-team schedule because no supervisory hours are used. For with-team schedules, supervisory hours are significant for highly-skilled providers (Figure 2-3), but the total labor requirement from highly-skilled providers declines because their time is now leveraged through the use of low-skill providers to treat high-acuity patients under supervision. For example, a P3 provider can independently treat an A3 patient at a relative cost of ¤50/hr. However, the same A3 patient can be served by a P1 provider (¤15/hr) who is supervised 30% of the time by a P3 provider (¤50/hr), resulting in a composite hourly cost of ¤30/hr. So, the overall effect is for total labor hours to increase, but for total costs decline.

A second and less obvious reason for lower costs is that using teams improves overall utilization of providers. Teams allow high level providers to serve as supervisors and leverage their skills across

lower level providers. Our experiments found that overall provider utilization was 51% with teams versus 46% without teams.

Figure 2-7. Optimal Staff Mixes With and Without Teams



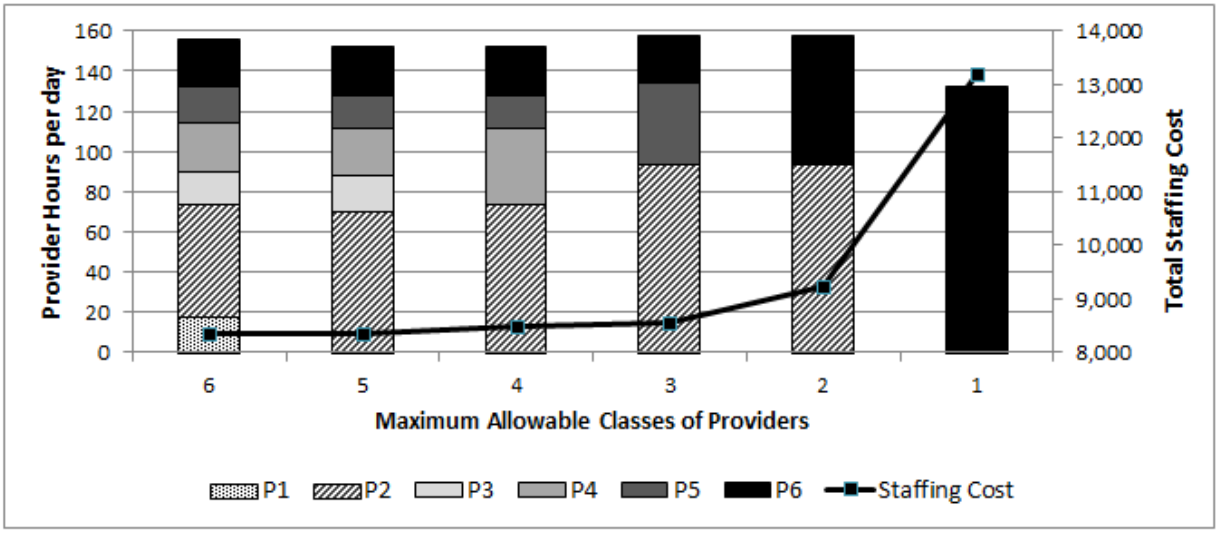
2.4.2. Restricting Provider Types

The base example problem assumed that the ED could employ any of six levels of trained providers. However, a specific ED may be limited to employing fewer classes of providers, either because of availability or for other administrative reasons. The question then arises as to how staffing costs of an ED are affected if the set of available provider levels is reduced. To understand this issue, we successively constrained the number of allowed provider skill levels from 6 to 1.

The results are illustrated in Figure 2-8, which shows the value of provider flexibility. As the number of provider-types is restricted, costs increase slowly and then accelerate. Restricting the number of provider types to three makes very little difference in terms of staffing costs, leading to a cost increase of only about 3%. If only two types are allowed, costs increased by over 10%; and if only a single provider-type is permitted, then costs increase by 58% despite a corresponding decline in total labor hours employed, since the entire staff is comprised of expensive P6 providers. Figure 2-8 also demonstrates the

diminishing marginal benefit of additional flexibility, which has been observed in other contexts such as manufacturing flexibility (Jordan and Graves, 1995) and employee cross-training (Brusco and Johns, 1998; Campbell, 1999).

Figure 2-8. Effects of Restricting the Number of Providers (with Teams)



In the reference EDs, only two levels of providers were regularly employed – P2 providers to attend to low and medium acuity patients and P6 providers to treat high-acuity patients and to supervise P2 providers. The optimal base case schedule with only P6 and P2 providers delivered staffing costs that were 7.8% lower than those incurred by the reference EDs schedule (Table 2-4). If the ED were to employ P4 providers in addition to the current mix of providers (P5 providers were not immediately available to the ED), costs would be further reduced by 5.5%. This result illustrates that our model can be effectively used both to create staff schedules, and to evaluate and compare alternative staffing policies.

Table 2-4. Staff Schedule Cost Comparisons

	Available Provider Types	Total Staff Costs (¤)	Total Staff Hours Scheduled
Reference Schedule	P2, P6	10,000	142
Optimal Schedule	P2, P6	9,220	158
Optimal Schedule	P1-P6	8,339	156

2.4.3. Impact of Patient Service Levels

An important parameter in our model is the target service-level for an ED. In this section, we examine (1) the impact of service-level on staffing costs, (2) patient wait times, (3) service cascading, and (4) schedules with differential service levels for different acuity levels.

2.4.3.1. Target Service-level and Costs

ED administrators will want to select a service level that not only provides excellent service to patients, but is also economically viable for the hospital. To understand the impact of service levels on staffing costs, we varied service levels between 99% and 60% and determined the resulting staff costs for the optimal schedule (Figure 2-9a). When the target service-level is reduced from 99% to 95%, staffing costs drop by 18%. With further decreases in service-levels, cost savings continue to decline but with diminishing returns. A reduction in target service levels increase staff utilization since fewer providers were required to meet lower target service levels. For an extremely high service-level of 99%, utilization is only 45%, but it increases as service-level was reduced.

Figure 2-9. Sensitivity to Patient Service Levels

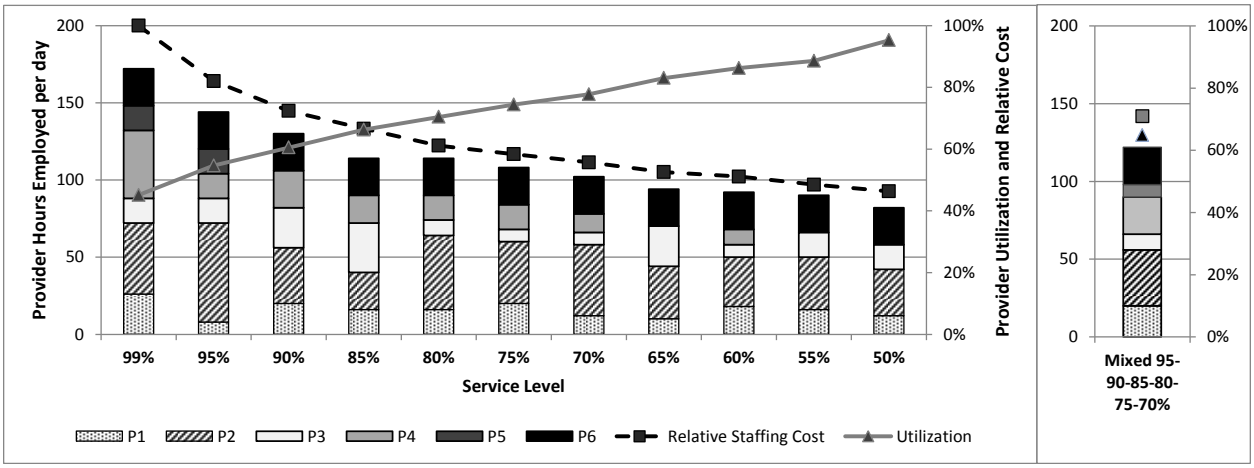


Figure 9a Figure 9b

As target service-levels declined below 99%, highly trained P6 providers accounted for a progressively larger share of the provider hours to be employed, since a minimum of one P6 provider was required to be on-staff at all hours. While this causes some degree of P6 overstaffing, this is partly

compensated by understaffing of the lower level providers, resulting in P6 personnel accounting for a larger share of provider hours employed for lower service-levels.

2.4.3.2. Long Patient Waits

We reiterate that during periods when demand exceeds capacity, patients are not turned away from an ED. Rather, providers attend to patients in the order of their acuities and some low acuity patients will wait longer to be treated. Therefore, it is important to examine occurrences of long patient waiting times.

Using the validation study described in Section 3.3, we collected the results of long wait times, i.e. instances when demand is not fulfilled during the hour during which it arrived, but is fulfilled during a subsequent hour. The validation study assumed that during any hour, total workload demand was comprised of unmet demand from the prior hour plus new demand arriving during that hour. Further, when provider capacity was insufficient to meet workload demand, priority is accorded to the most acute patients, and for patients belong to the same acuity level, priority is given to those who arrived first.

Figure 2-10. Occurrence of Long Wait Times

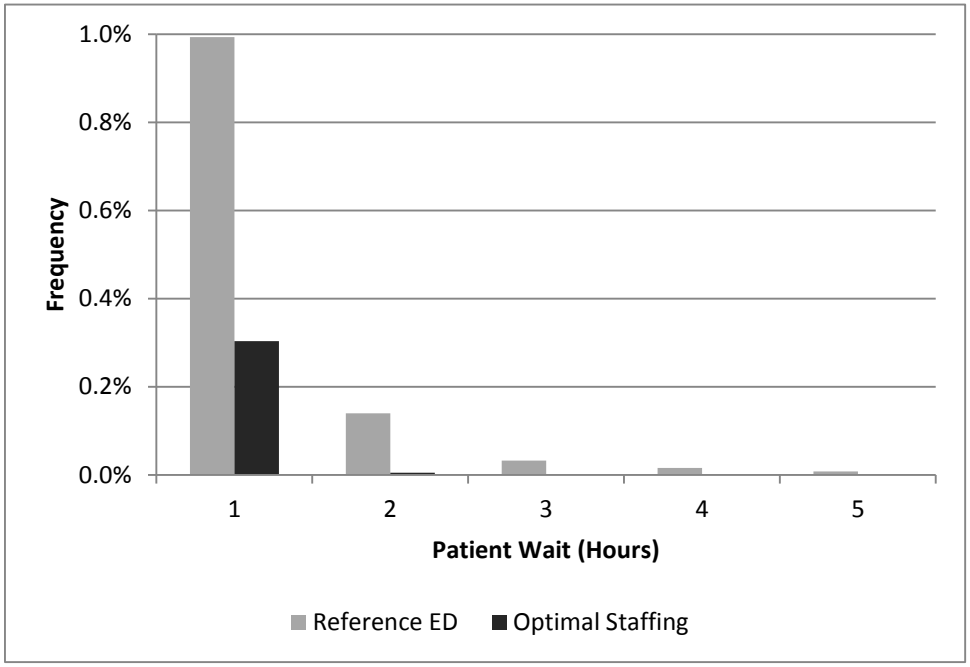
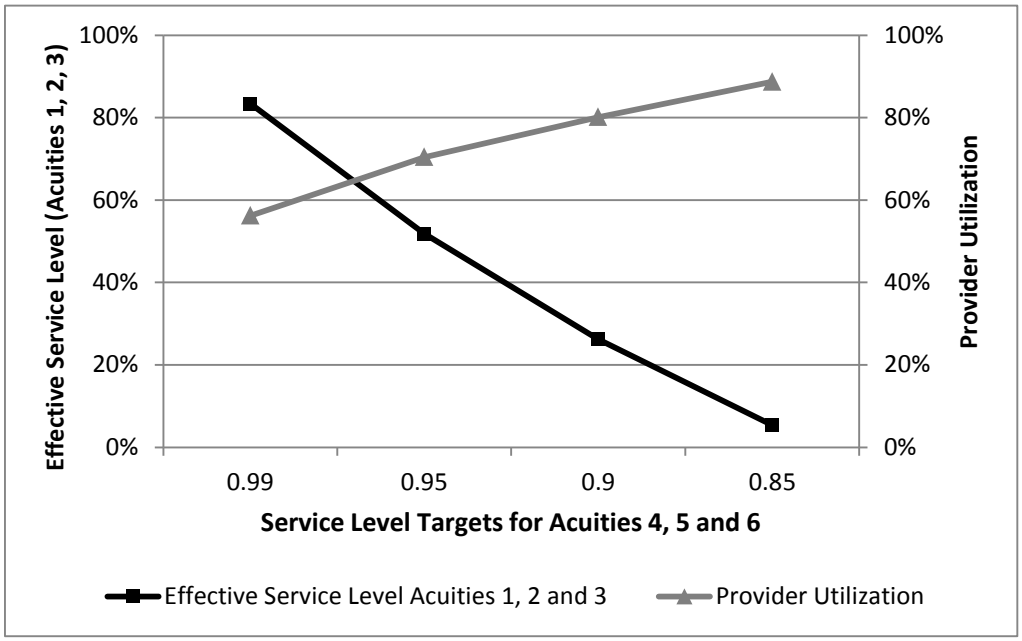


Figure 2-10 shows that the optimal staffing schedule results in significantly fewer instances of long patient waits compared to the reference ED. This result is encouraging since it demonstrates that effective ED staffing can provide both lower costs as well as improved patient care.

2.4.3.3. Service Cascading

Figure 2-11. Service Cascading to Lower Acuity Patients (Target Service Level for acuities 1, 2, & 3 = 0%)



An interesting feature of the ED staff planning problem is that even if an ED is staffed to fulfill a high service-level for only high acuity patients and does not put any additional staff to care for low acuity patients, lower acuity patients may still experience a reasonable degree of service. This phenomenon occurs because high skill-level providers, when not fully utilized by high-acuity patients, can treat low-acuity patients. Based on the empirical distribution of hourly demand at the example ED of Section 3, if the ED is only staffed to meet a service-level of 99% for acuity levels ℓ=6, 5 and 4, without considering demand from lower acuity-levels ℓ= 1, 2 and 3, then these lower acuity patients the will none-the-less experience an effective service-level of 83%, *without any additional staffing*. However, because this implicit service-level experienced by the lower-acuity patients is dependent upon low utilization of highly-skilled providers, it is very sensitive to the service-level target for the higher acuities (Figure 2-

11). This observation implies that if service levels are high for high-acuity patients, there is opportunity to reduce staffing costs by decreasing service-levels for lower acuity patients, and that lower-acuity patients are likely to receive good service even if their target service levels are lower than those for higher-acuity patients.

2.4.3.4. Differential Service-levels

The phenomenon of service cascading suggests that ED staffing costs might be further reduced by assigning lower target service levels to less severe levels of patient acuity. Since EDs are primarily designed to treat critical patients, that is those with high acuity levels, and given the high risk of being understaffed for critical patients, an ED may well prefer to assign high service-levels to the highest acuities and progressively lower service-levels for the lower acuities. We investigated a scenario using acuity-specific service-levels of 95, 90, 85, 80, 75 and 70% for acuity-levels $\ell = 6, 5, \ldots 1$ respectively which resulted in a weighted average service level of about 86% (Figure 2-9b). Results of this test indicate that setting lower service level targets for lower acuity patients can significantly reduce costs compared to setting a single target level for all acuity levels. For example, the mix of target levels above, staffing costs are 87% of the costs incurred with a uniform 95% service target – somewhat less than the costs incurred with a uniform 90% target, while providing higher service levels for the most acute patients.

2.4.4. Implementation Issues

Several issues have important practical implications for the implementation of our model, including selection of a time bucket duration and estimation of demand distributions for each bucket; determination of an appropriate service level, and the decision to use (or not use) different service levels for different patient acuity levels. Each of these are discussed below,

Time Bucket Duration. It is important for ED administrators to develop accurate distributions of patient demand for ED services. Time buckets should be small enough to capture granularity of deviation (in our example problem we have used hourly time-buckets). Smaller time-buckets generally lead to

higher fidelity, but require more data and computation time and may not comport with the start of clinic personnel shifts. If historical data is not available from which to estimate demand distributions, practitioners can estimate demand distributions based on experience and intuition, and these distributions should be updated and refined as time unfolds and new data is collected. To test the sensitivity of staffing costs to the demand distribution, we experimented with variations in mean and standard deviation of demand. The results, which we do not report in detail for the sake of brevity, indicate that while good estimate of demand distribution is important for effective use of our model, the model is not disproportionally sensitive to specification errors either in the mean or the variability of demand. We also observe that the recommended staffing plan is robust to variability of service times, though highly variable service times increase the need for some patients to be treated by multiple providers (Figure 2-6).

Patient Service Level. Also important for implementation is the choice of appropriate patient service level, since it impacts the quality and availability of care (by definition), as well as staffing costs, provider-utilization, and the mix of providers scheduled. A first consideration is to understand the trade-off between patient service level and staffing cost. The appropriate trade-off will vary widely for different EDs depending on patient mix, ED budget, mix and availability of providers, and average patient acuity, among many other considerations. However, our results show that there are declining returns to reductions in patient service levels, which suggests that relatively high service levels can be sustained without inordinate increases in staffing costs. In our test problem, for example, cost savings of about 28% are obtained by cutting service levels from 99% to 90%, but further cuts to 80% only save another 11% in staffing costs (Figure 2-9a).

Variable Service Levels. Another service level consideration is whether or not to vary service targets by acuity level. We have shown that reducing service levels for lower acuity patients can significantly reduce staffing costs (Figure 2-9b), but that the actual service level experienced by low-acuity level patients is often greater than the specified target (Figure 2-11). These two results together suggest that assigning lower service level targets for lower acuities may be a good way to reduce staffing costs without serious declines in effective patient service levels.

2.5. Contributions and Conclusions

We have developed a novel optimization model to obtain staffing schedules in medical emergency departments based on data and observation of three metropolitan EDs. Our paper makes contributions to both the management of emergency departments and to personnel staffing modeling and methodology.

For ED administrators, we provide a tool that can be used to generate effective staff schedules that reduce costs while maintaining target service levels for arriving patients. We demonstrate that assigning different service level targets to different patient acuity classes can further reduce costs without large degradation in service for lower-acuity patients. Our model can also be employed by administrators to evaluate alternative hiring and staffing strategies for an ED, so that the appropriate mix of providers is deployed to minimize costs and maintain service. Finally, we have demonstrated the power of provider work teams in reducing costs and improving utilization. Counter-intuitively, we find that the use of teams may increase provider head-count, but can improve provider utilization and significantly reduce total staffing costs.

Unique methodological aspects of our model include the calculation of aggregate demand distributions for various levels of service providers and the inclusion of teams with both attending and supervising providers, each with different service-time demands. The calculation of separate demand distributions for different service provider classifications was necessary since highly skilled providers can attend to patients of any acuity level, while lesser skilled providers can only attend to low acuity patients. Incorporating provider teams was based on our observations of teaming behavior in our reference EDs, where highly skilled providers supervise lesser skilled providers to treat medium-acuity patients. Incorporating teams into our model required significant modification to the usual personnel staffing integer programming formulations, as shown in Section 2. We have also shown that the need for multiple providers to attend to a single patient increases with the variability of service time. We are unaware of previous papers or models that incorporate these features in the examination of personnel staffing decisions.

While our focus in this paper has been on ED staffing, we note that our model can be extended and used in other settings wherever there is a range of provider skill-levels, and customers or tasks with a range of problems and needs. Examples include mechanics working on aircraft, chefs and cooks working in large kitchens, offices of security traders, and call centers, among many others. In each of these settings, lesser skilled providers typically work on easier problems and customers, but can be supervised or aided by higher-skilled providers to work on tougher problems and customers.

This work can be extended in interesting ways. For example, emergency department administrators are investigating the use on-call providers to help out in case of unexpected patient demand. We currently are working on this extension, and others, to the foundational ED staff planning model developed in this paper. Improvements in computational performance is another avenue of enhancing our work, since our model currently takes a long time to converge to optimality, especially when team providers are involved.

Chapter 3 - PROVIDER'S WAIT-PREEMPT DILEMMA

In Ganguly and Samorani (2013), we develop an analytical model to derive the optimal preemption policy for outpatient clinics that operate on an appointment basis.

3.1. Introduction

It has been known for decades that patients tend to arrive early (Swartzman 1970) for their appointment at outpatient clinics. More recently, Cayirli et al. (2006) reported that patients at a primary health care clinic in New York arrive 17 minutes early on average with a standard deviation of 27 minutes. Depending on the length of the appointment slot, these numbers suggest that the order in which patients arrive is often different from the order in which they are scheduled. Out-of-order patients may result in disruptions which, if not dealt with effectively, could lead to a long patient waiting time or a long clinic overtime. In particular, it is not clear whether a provider should see early patients as soon as they arrive or should instead wait for the patient scheduled to be seen next.

Typically, this problem is resolved using a first-come-first-served policy (FCFS), under which the provider sees patients in the order of their arrival and does not stay idle in the presence of waiting patients. Although the provider's idle time (and, thus, clinic overtime) is minimized under the FCFS policy, some patients may have to wait for a very long time, even if they are punctual. For example, if a patient arrives early and out of turn, an idle provider will start seeing the early patient, without waiting for the patient scheduled to be seen next. If the patient who was supposed to be seen next appears soon thereafter, she will have to wait for her turn, even though she may have been punctual.

Thus motivated, we analyze the dilemma that an idle provider faces when, while waiting for the next scheduled patient (target patient T) to show up, finds a later scheduled patient (waiting patient W) in the waiting room. Should the provider keep waiting for patient T, who might not show up at all, or see patient W right away (i.e., preempt)? On one hand, waiting will result in idle time, which may in turn lead to overtime; on the other hand, preempting may result in a long waiting time experienced by T, if T

arrives shortly after the provider starts seeing W. This paper attempts to answer this question, which we call the "wait-preempt" dilemma.

We optimally solve the dilemma in a simplified setting with only two patients; then, we extend our method to an arbitrary number of scheduled patients. Our objective is to find the time intervals in which the provider should wait for the missing patient and those where she should preempt and see the early patient. We find these intervals by analyzing the cost function, which, for each point in time x, computes the expected cost incurred if the provider decides to wait up to x for the missing patient. The cost function is a weighted sum of the patients' waiting time and the clinic overtime. We extend our procedure to the N-patient case by solving a "virtual" dilemma for each pair of consecutively scheduled patients. The cost is high if the delay is likely to propagate to the end of the clinic; it is low if the delay is likely to be absorbed soon, for example, by an empty appointment slot.

Our contributions are summarized as follows:

1. **Analytical method for the 2-patient case**. We optimally solve the wait-preempt dilemma for the 2-patient case. Compared to the FCFS policy, our method leads to a dramatic reduction in waiting time at the cost of a modest increase in overtime.

2. **Managerial insights**. We analyze structural properties of the cost function and identify important insights. The main results are: 1) the FCFS policy is never optimal under any parameter configuration: for example, it is always optimal to wait for a patient when the dilemma arises shortly before his/her scheduled time; 2) a "wait" decision is more desirable for those parameter configurations under which patients are unlikely to arrive more than an appointment slot late; 3) the patient no-show probability does not affect the optimal decision in the 2-patient case. However, in the N-patient case, a "wait" decision is more desirable for lower show probabilities. This counterintuitive result stems from the higher probability that any delay induced by waiting will be likely absorbed by a subsequent slot left empty by a non-showing patient.

3. **Extension to the N-patient case**. We extend our analytical method to the N-patient case and test it using both simulated and empirical arrival data from an outpatient clinic on a university campus. Our results indicate that over a wide range of clinic characteristics, the recommended analytical method reduces expected costs of patient waiting and clinic overtime by 4-20%.

4. **Software**. We provide a software program that clinics can readily use to solve the dilemma.

Our work is not only applicable to outpatient health clinics, but also to any other appointment-based activity, such as, for instance, lawyer offices.

3.1.1. Relevant Literature

Despite its relevance, the dilemma has not been addressed in existing works on appointment scheduling, such as Huang and Zuniga (2012), Cayirli et al. (2012), LaGanga and Lawrence (2007 and 2012), Liu et al. (2010), Robinson and Chen (2010), Zeng et al. (2009), just to name a few. These works have focused on finding an appointment rule which defines how appointment slots should be assigned given that patients might not show up, but they do not suggest what a provider should do when patients arrive out of order during the clinic session. In fact, all these works assume patients to be punctual. While considering unpunctuality is not very critical in the choice of appointment rules (Cayirli et al. 2006), we will show that clinics can dramatically improve their performance by considering unpunctuality information during the clinic sessions.

In line with the discussed literature, Cayirli et al. (2006) do not consider unpunctuality to obtain an appointment rule; however, they consider it during their simulation experiments. While they recognize that deciding which patient to see first "is more complex than it may at first appear", they suggest that the provider should always see waiting patients. If two or more patients are waiting, then following an FCFS policy, the provider sees the patient who arrived first. The literature survey by Cayirli and Veral (2003) reveals that all works that consider unpunctuality use an FCFS policy to decide which patient to see first, because "doctors would not keep idle waiting for the next appointment in the presence of other waiting

patients". Although conventional wisdom would justify this statement, we show that sometimes the provider should in fact remain idle even in the presence of a waiting patient who arrived early.

Under an alternative approach, observed by Gupta and Denton (2008), patients are deemed to be no-shows if they are 10 minutes late. If a patient has not arrived by then, the provider sees another patient. Our method generalizes this idea by finding the optimal amount of time to wait for a missing patient.

Finally, note that although preemption problems are well studied in the traditional scheduling literature (Julien et al. 1997), the problem considered here is different because jobs (patients) may not show up.

3.2. Problem Definition and Analytical Method

In this section, we develop the notation (Table 3-1), the assumptions, and the model for the wait-preempt dilemma involving only two patients. Throughout the paper, we use W to denote the waiting patient, who has arrived early and out of turn, and T to denote the target patient, who is scheduled to be seen next.

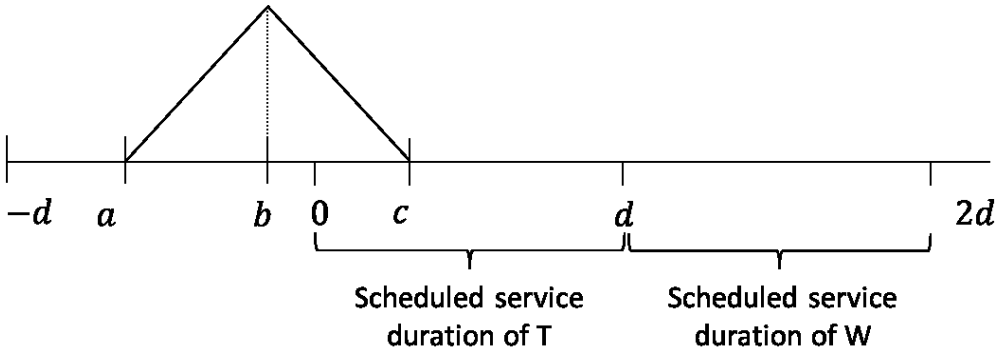
Table 3-1. Notation

d	Service time (duration)
t_0	Current time
a	Earliest possible arrival time of target patient T
b	Expected arrival time of patient T
c	Latest possible arrival time of patient T
q	Unconditional show probability of patient T
$p(t)$	Probability density function for time t at which patient T arrives.
$p(t_0, t)$	Conditional probability density function for time t at which patient T arrives, given that T has not shown up by time t_0.
ω	Cost for each unit of waiting time per patient
τ	Cost for each unit of clinic delay
E_{TC}	Expected total cost
E_T	Expected waiting time of patient T
E_W	Expected waiting time of patient W
E_D	Expected clinic delay

Without loss of generality, let patient T's visit be scheduled from time 0 to time d, where d is the service time (assumed constant in this section), and W's visit be scheduled from d to $2d$, i.e. patients T

and W are scheduled back to back. Suppose that at current time t_0, W has already arrived and T has not shown up yet. This situation is graphically shown in Figure 3-1. For reasons explained below, we approximate the arrival time of T with a triangular distribution with parameters a = earliest arrival time, b = most likely arrival time, and c = latest arrival time. Consistent with the literature (Swartzman 1970, Cayirli et al 2006), we assume $a < b < 0 < c$, i.e., patients may arrive early or late, but they tend to be early. The (unconditional) show probability of T is q, while the probability of T showing up at time t, given that T has not shown by time t_0, follows the density function $p(t_0, t)$. The derivation of $p(t_0, t)$ is reported in Appendix 3.1.

Figure 3-1. Graphical representation of the Wait-Preempt dilemma



The problem consists of finding a wait-up-to time x ($x \geq t_0$) until which the provider should wait for T before seeing W. If T arrives by x, the provider sees T first, and then sees W. Otherwise, the provider starts seeing W at x, and then possibly sees T, if T arrives. Let x^* denote the optimal value for x. The relevant clinic cost is the weighted sum of a "waiting" cost and a "clinic delay" cost. A waiting cost of ω is incurred for each unit of waiting time experienced by patients after their scheduled appointment time or their arrival time, whichever is later. This model is commonly used to measure the performance of an appointment system with unpunctual patients (Cayirli and Veral 2003; Cayirli et al. 2006). The waiting cost includes the patient's loss of earning while waiting and the nonmonetary cost corresponding to the loss of goodwill incurred by the clinic. Finally, a clinic delay cost of τ is incurred for each time unit by which the finish time of the two visits exceeds their nominal end time $2d$. The delay cost measures the negative impact of the delay on the rest of the clinic session: it subsumes the

waiting time of the following patients and the cost of overtime (such as overtime wages, electricity cost, etc…) incurred by the clinic if the delay is propagated until the end of the clinic session. Here, we make a distinction between "clinic delay" and "clinic overtime", since a clinic delay does not necessarily lead to clinic overtime if the providers have enough slack time later in the day. Note that if $t_0 \leq -d$ or $t_0 \geq \min(c, d)$, the dilemma is trivially solved by choosing to preempt ($x^* = t_0$). In the former case, there is enough time to see W before T starts incurring waiting time cost; in the latter case, either there is no chance that T arrives ($t_0 \geq c$) or W should be visited because he/she is already incurring waiting time cost ($t_0 \geq d$). So, if $t_0 \notin [-d, min(c, d)]$, the solution is to preempt. For this reason, we assume $x \in [t_0, min(c, d)]$. Accordingly, we set up expressions for expected waiting times as a function of t_0 and x.

Waiting Time of Patient T: Patient T may incur waiting cost only if (s)he arrives when the provider is busy seeing patient W, i.e., when T arrives at time t such that $x < t < x + d$. In this case, T begins incurring waiting cost at $t^+ = max[0, t]$ and ends waiting at time $x + d$. By conditioning on the arrival time of T and using the law of total probability, we can express T's expected waiting time as:

$$E_T(t_0, x) = \int_x^{x+d} p(t_0, t) max[0, (x + d - t^+)] dt \quad (1)$$

Waiting Time of Patient W: Patient W may incur waiting cost only if patient T arrives at time t such that $t_0 < t \leq x$. In this case, T's visit starts at t and ends at $t + d$. Since W has arrived early, (s)he begins incurring waiting cost at d and ends waiting at $(t + d)$. Thus, W's waiting time is equivalent to T's lateness, t^+. Therefore, expected waiting time of patient W is:

$$E_W(t_0, x) = \int_{t_0}^{x} t^+ \cdot p(t_0, t) \cdot dt \quad (2)$$

Clinic Delay: If T arrives at $t \in [t_0, x]$, then the provider starts seeing T at t and W at $t + d$ and ends at $t + 2d$, causing a delay of at most t. If T arrives at $t \in [x, x + d]$ (i.e., while W is being seen), then the provider starts seeing T at $x + d$ and ends at $x + 2d$, causing a delay of at most x. Finally, if T arrives at $t \in [x + d, 2d]$ (i.e., after W has been seen), then the provider starts seeing T at t and ends at $t + d$, inducing a delay equal to T's lateness, $(t - d)^+$. Thus, the expected clinic delay can be expressed as:

$$E_D(t_0, x) = \int_{t_0}^{x} t^+ p(t_0, t) \cdot dt + \int_{x}^{x+d} x^+ p(t_0, t) \cdot dt + \int_{x+d}^{2\mathrm{d}} (t-d)^+ p(t_0, t) \cdot dt \quad (3)$$

Let $E_{TC}(t_0, x) = \omega E_T(t_0, x) + \omega E_W(t_0, x) + \tau E_D(t_0, x)$ be the total cost of the expected patient waiting times and the expected clinic delay. The basic analytical problem (BAP) consists of finding the value of x that minimizes E_{TC}:

(BAP_{t_0}) $\quad \min_{\mathrm{x}\in[\mathrm{t}_0, \min(c,d)]} \omega E_T(t_0, x) + \omega E_W(t_0, x) + \tau E_D(t_0, x)$

According to empirical evidence presented by Swartzman (1970) and Cayirli et al. (2006), $p(t)$ follows a normal distribution. If that is the case, closed form expressions of the indefinite integrals of $p(t)$ and $t \cdot p(t)$ do not exist. As a consequence, indefinite integrals of $p(t_0, t)$ and $t \cdot p(t_0, t)$ cannot be expressed in closed form as well, which makes it difficult to solve BAP_{t_0} analytically. Hence, for $p(t)$, we use a triangular approximation of normal distribution and solve the problem analytically.

3.2.1. The Analytical Method

Whenever the dilemma arises, the BAP_{t_0} needs to be solved for current time t_0 to find the optimal time x^* until which the provider should wait before seeing W. While this is possible, it will be inconvenient for the clinic to solve the problem for all possible values of t_0, or to solve the problem in real time whenever the dilemma arises. In this section, we develop a method to find the optimal decision for each point in time by solving the BAP only once.

We now show that the local minima and maxima achieved by the cost function $E_{TC}(t_0, x)$ are independent of t_0. This implies that if the optimal decision at time t_0 is to wait up to time x, then the optimal decision computed at any time in (t_0, x) is still to wait up to time x. To establish this important property, we prove that the derivative of the cost function computed at t_0 is proportional to the derivative of the cost function computed at $t_0 + \lambda$.

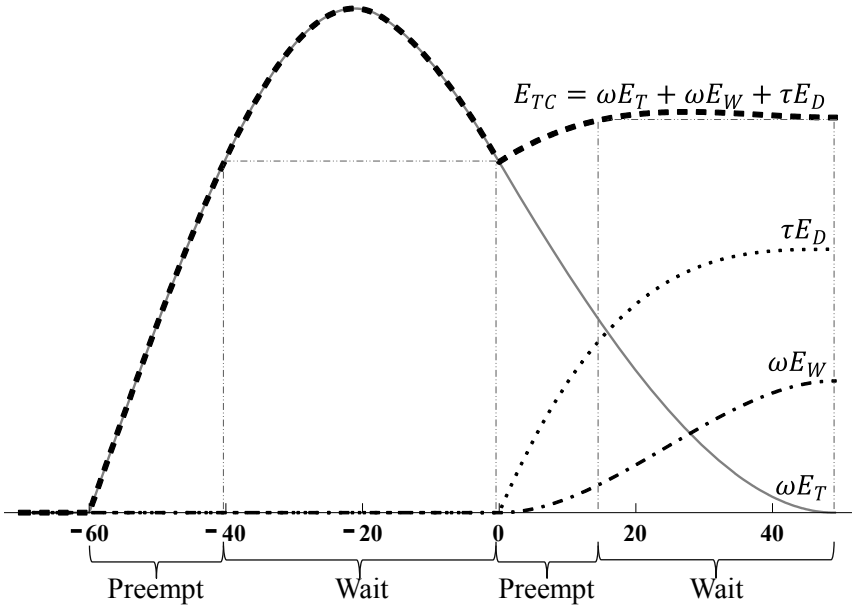
THEOREM 1. *For any* $\lambda > 0$, $\frac{\partial E_T(t_0+\lambda, x)}{\partial x} = k_1 \cdot \frac{\partial E_T(t_0, x)}{\partial x}$, $\frac{\partial E_W(t_0+\lambda, x)}{\partial x} = k_2 \cdot \frac{\partial E_W(t_0, x)}{\partial x}$, *and* $\frac{\partial E_O(t_0+\lambda, x)}{\partial x} = k_3 \cdot \frac{\partial E_O(t_0, x)}{\partial x}$, *where* k_1, k_2, *and* k_3 *are positive quantities independent of x.*

Proof: see Appendix 3.2.

This theorem implies that the minima and maxima attained by the cost function are independent of t_0. Thus, COROLLARY. *Solving $BAP_{t_0+\lambda}$ is equivalent to solving BAP_{t_0} with the additional constraint* $x \geq t_0 + \lambda$

This corollary implies that $BAP_{t_0+\lambda}$ can be solved by plotting the cost function $E_{TC}(t_0, x)$ of the original problem BAP_{t_0} and finding the minimizer in the interval $x \in [t_0 + \lambda, \min(c, d)]$. If the minimizer is $t_0 + \lambda$, then the optimal decision is to preempt; otherwise, it is to wait. So, we arbitrarily fix t_0 to a so that the expression of $p(t_0, t)$ simplifies to the expression of $p(t)$, and compute the value of $E_{TC}(a, x)$ in $x \in [a, \min(c, d)]$. Per the Corollary above, if the dilemma arises at a future time $a + \lambda$, the optimal wait-up-to time will be $x^* = \text{argmin}_{x \in [a+\lambda, \min(c,d)]} E_{TC}(a, x)$. This process is best illustrated through an example. Suppose that $d = 60$ minutes and $a = -80$ minutes, and that at time $t_0 = a$, the total cost curve $E_{TC}(t_0, x)$ is the one depicted in Figure 3-2.

Figure 3-2. Components of the cost function for d = 60 minutes, a = -80 minutes, τ = 2, ω = 1



If the dilemma arises at t_0 = -80 minutes, the optimal solution is to preempt, because there is no time in the future that has a lower total cost. By Theorem 1, if we recomputed the cost function at $t_0 + \lambda$ = -55 minutes, we would obtain a "scaled" function, whose derivative is multiplied by a constant factor.

This "scaling" transformation, though, does not affect the optimal decisions. Therefore, we can use the function in Figure 3-2 to establish that the solution of the dilemma arising at -55 is still to preempt. However, at about $t_0 + \lambda$ = -40 minutes, the optimal solution changes to *wait until* $x^* = 0$, where the total cost is lower. In conclusion, rather than solving the problem every time the dilemma arises, it is possible to solve it just once by analyzing the cost function $E_{TC}(a, x)$ (denoted as $E_{TC}(x)$ from now on).

The optimal solution of a dilemma arising at time t is to wait if there exists a point in time $x \in (t, \min(c, d)]$ such that $E_{TC}(x) < E_{TC}(t)$, and to preempt otherwise. The analytical method consists of finding the intervals where it is optimal to wait (wait intervals) and those where it is optimal to preempt (preempt intervals), both reported in Figure 3-2. When the optimal decision is to wait, the provider should wait until the next preempt interval or until $min(c, d)$, whichever is sooner. So, in our example, if W arrives at -50, the provider should preempt and see him/her right away; however, if W arrives at -30, the provider should wait for T until the scheduled time 0, and then, should T have not shown yet, see W. Appendix 3.3 provides the theoretical support for the analytical method and describes its implementation details.

In the following section, we consider several possible combinations of clinic parameters and uncover interesting managerial insights that will help clinic administrators choose an appropriate wait-preempt policy.

3.3. Analytical Results

Depending on the values of the clinic parameters c, d, ω, and τ, the shape of the cost function E_{TC} may be different from that shown in Figure 3-2. Note that irrespective of the parameter values, if the provider waits up to $x \leq 0$, neither does the waiting patient suffer any waiting time, nor is there any possibility of clinic delay caused as a result of the decision to preempt. Therefore, for $x \leq 0$, $E_{TC} = E_T$. Recall also that no cost is incurred for $x = -d$, i.e., $E_{TC}(-d) = E_T(-d) = 0$.

THEOREM 2. *The interval* $[-d, 0)$ *always contains a wait interval.*

Proof: see Appendix 3.2.

This implies that if the dilemma arises shortly before the scheduled time of T, then irrespective of the values of clinic parameters, the provider should wait. The presence of a wait interval proves that FCFS is never an optimal policy under any parameter configuration.

Having established that one wait interval exists in $[-d, 0)$, we now consider the behavior of the cost function for $x \geq 0$. Assuming $t_0 = a$, the expression of the derivative of E_{TC} in $x \in [0, \min(c, d)]$ is derived as follows:

$\frac{\partial E_T(x)}{\partial x} = -dp(x) + \int_x^{x+d} p(t)dt$ (Case 1 in Appendix 3.4)

$\frac{\partial E_W(x)}{\partial x} = x \cdot p(x)$ (Case 1 in the Appendix 3.4)

$\frac{\partial E_D(x)}{\partial x} = \int_x^{\min(c,x+d)} p(t) \cdot dt$ (Cases 2.1, 2.2, and 2.3 in Appendix 3.4)

$$E'_{TC}(x) = \omega \frac{\partial E_T(x)}{\partial x} + \omega \frac{\partial E_W(x)}{\partial x} + \tau \frac{\partial E_O(x)}{\partial x}$$

$$= \frac{2q}{(c-a)(c-b)} \left[\omega(x-d)(c-x) + (\omega + \tau) \int_x^{\min(c,x+d)} (c-t)dt \right]$$

The behavior of E_{TC} is summarized in Table 3-2 (derived in Appendix 3.5).

Table 3-2. Cost function E_{TC} in $[0, \min(c, d)]$

Case	Conditions	Behavior of $E_{TC}(x)$
1	$\tau > \omega, \frac{2\omega}{\tau+\omega} d \leq c < d$	Increases in $[0, x_2)$, decreases in $(x_2, c]$
2	$\tau > \omega, c < \frac{2\omega}{\tau+\omega} d \leq d$	Decreases in $[0, c]$
3	$\tau < \omega, c < d$	Decreases in $[0, c]$
4	$\tau > \omega, c \geq d$	Increases in $[0, d]$
5	$\tau < \omega, d \leq c \leq \frac{3\omega-\tau}{2\omega} d$	Decreases in $[0, x_2)$, increases in $(x_2, d]$
6	$\tau < \omega, d \leq \frac{3\omega-\tau}{2\omega} d \leq c \leq \frac{\omega+\tau}{2\tau} d$	Decreases in $[0, x_3)$, increases in $(x_3, d]$
7	$\tau < \omega, d \leq \frac{3\omega-\tau}{2\omega} d \leq \frac{\omega+\tau}{2\tau} d \leq c$	Increases in $[0, d]$

Several interesting findings emerge from our analysis of the cost function E_{TC}. First, the wait and preempt intervals are independent of patient T's show probability q, as evident by its absence from the expressions presented in this section. This is counter-intuitive, as conventional wisdom would justify not waiting for a patient who is very unlikely to show.

Second, let us consider the interval $[0, \min(c, d)]$. From Table 3-2, we see that whether E_{TC} increases or decreases depends on the ratios τ/ω and c/d. This is shown graphically in Figure 3-3, where '--' ('++') indicates that E_{TC} decreases (increases) in the whole interval $[0, \min(c, d)]$, whereas '-' ('+') indicates that although E_{TC} is decreasing (increasing) for the most part of the interval $[0, \min(c, d)]$, it is increasing (decreasing) for the rest of it.

Figure 3-3. Increasing (+, ++) and decreasing (-, --) trends of the cost function in $[0, \boldsymbol{min}(\boldsymbol{c}, \boldsymbol{d})]$

Delay-to-waiting cost factor ratio (τ/ω)

Lateness-to-slot ratio (c/d)	**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**	**1.2**	**1.4**	**1.6**	**1.8**	**2.0**
0.0	--	--	--	--	--	--	--	--	--	--	--
0.2	--	--	--	--	--	--	--	--	--	--	--
0.4	--	--	--	--	--	--	--	--	--	--	--
0.6	--	--	--	--	--	--	--	--	--	--	--
0.8	--	--	--	--	--	--	--	--	-	-	+
1.0	-	-	-	-	-	++	++	++	++	++	++
1.2	-	-	-	+	++	++	++	++	++	++	++
1.4	-	+	+	++	++	++	++	++	++	++	++
1.6	+	+	+	++	++	++	++	++	++	++	++
1.8	+	+	++	++	++	++	++	++	++	++	++
2.0	+	+	++	++	++	++	++	++	++	++	++

When waiting time cost factor ω and maximum patient lateness c are constant, the cost function E_{TC} tends to be increasing in $x \geq 0$ for large delay cost factors τ or short service times d; this indicates that in these cases, a "preempt" decision is more likely to be the optimal. For large delay costs,

preemption is favored because it is undesirable to induce any delay; for short service times, preemption is favored because even if patient T arrives soon after W's visit starts, T will wait only for a short while. Therefore, for these parameter configurations, the FCFS policy tends to be optimal after T's scheduled time $t = 0$. Conversely, for small delay costs or long service times, a "wait" decision is more likely to be the optimal, as it can be prudent to wait for patient T even past her scheduled time. The numerical experiments presented in the next section confirm our conjecture.

In our stylized settings, only c, d, τ, and ω affect the critical ratios τ/ω and c/d, and determine whether waiting should be preferred to preempting. However, in a realistic environment, the following factors come into play as well. First, consider the variability of the patients' lateness: lower variability of patient lateness would imply a smaller value for the latest arrival time c, which would enhance the favorability of a "wait" decision. Second, consider the average lateness: if patients arrive earlier on an average, that would also result in a smaller value for the latest arrival time c, and thus have a similar effect. Finally, consider the likelihood of absorbing the delay due to a slack in schedule. If, for example, the next appointment slot is empty (i.e., not booked) or if the show rate is low, then any delay induced by waiting will likely be absorbed soon, without impacting many patients and without resulting in clinic overtime. In such situations, the delay cost τ is low, reflecting a preference towards waiting. The computational experiments in the next sections validate these conjectures.

3.4. Numerical Experiments

3.4.1. Experiments with Two Patients

We tested the efficacy of our analytical method and confirmed the analytical results presented in Section 3 through simulation, by generating situations where the dilemma between two patients W and T arises. To generate an instance of dilemma, both patient W's and patient T's lateness is sampled from a Normal distribution with parameters $\mu \in \{-32, -17, -2\}$ minutes and $\sigma \in \{20, 27, 34\}$ minutes. These values were chosen by building ranges of ± 15 minutes and ± 7 minutes around the values observed by Cayirli et al. (2006), μ = -17 minutes and σ = 27 minutes. If T happens to arrive before W or outside the domain

$[-d, min(c, d)]$, the dilemma does not arise; therefore, the instance is discarded and a new one is generated. If, on the other hand, the dilemma arises, a service time for both T and W is sampled from a Gamma distribution with parameters $\alpha \in \{1, 15, 10^5\}$ and $\beta = d/\alpha$, where the expected service duration $d \in \{10, 20, 30, 60\}$ minutes. Thus, for $\alpha = 1$, service times are exponentially distributed; for $\alpha = 15$, they are somewhat variable; for $\alpha = 10^5$, they are practically constant. Without loss of generality, we fix the waiting time cost factor to $\omega = 1$, and consider the clinic delay cost factor $\tau \in \{0.0, 0.5, 1.0, 2.0\}$. We generated 10,000 dilemmas for each parameter configuration (μ, σ, α, d, τ) indicated above ($3 \times 3 \times 3 \times 4 \times 4 = 432$ combinations).

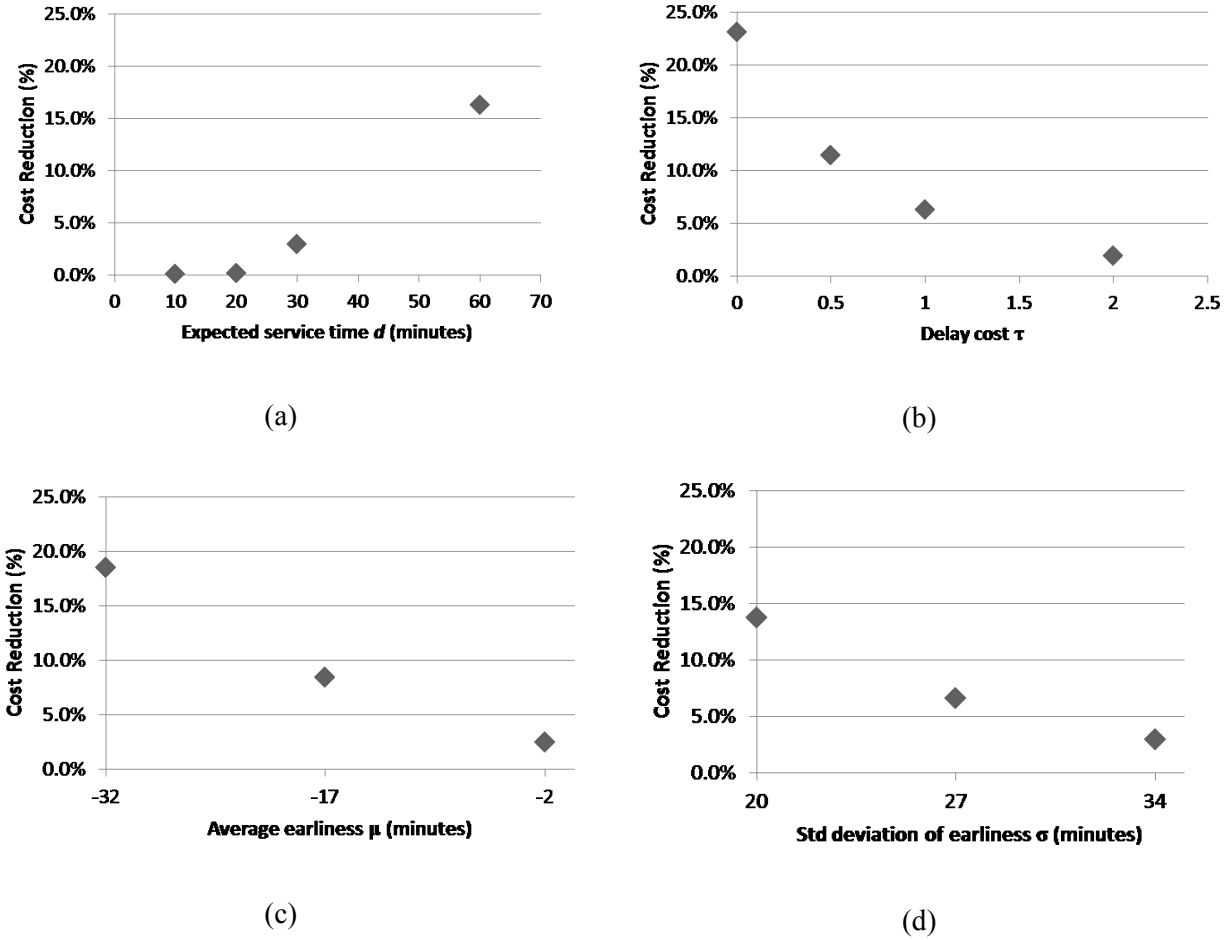
We resolve each instance of dilemma using three different methods: our analytical method, the first-come-first-served (FCFS) policy, and the 10-minute policy observed in practice by Gupta and Denton (2008). To execute the analytical method, we approximate the normal distribution $N(\mu, \sigma)$ with a triangular distribution $T(a, b, c)$ by setting $a = \mu - \sqrt{6}\sigma, b = \mu, c = \mu + \sqrt{6}\sigma$, as suggested by Scherer et al. (2003). Note that this triangular distribution is used only to make the decision; the simulated arrival times are sampled from a normal distribution. Under the FCFS policy, patients are seen in the order of arrival, regardless of their scheduled time. In the 2-patient case, this is equivalent to always choosing to preempt. Under the 10-minute policy, the provider will wait until T is 10 minutes late, at which point she will start seeing W. Table 3-3 reports the average cost, the average longest wait (i.e., the average waiting time experienced by the patient who waits longer), and the average overtime obtained across all simulations.

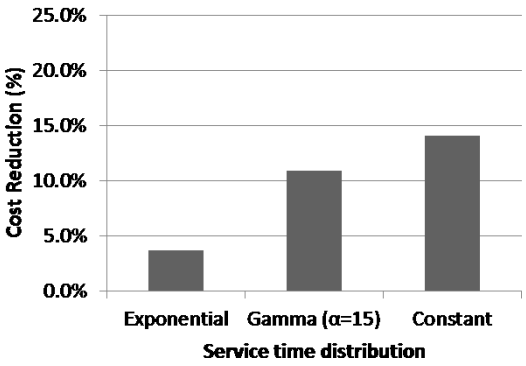
Table 3-3. Results in the 2-Patient Case

	Cost	Longest wait (minutes)	Clinic delay (minutes)
FCFS	0.56	14.60	8.98
10-minute	0.70	13.39	11.64
Analytical method	0.52	10.92	10.72

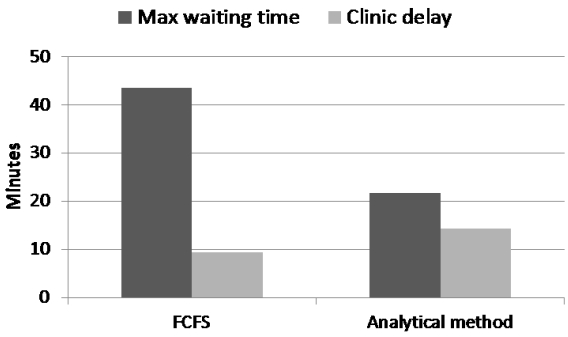
On an average, the analytical method results in the lowest cost among the three policies. It dominates the 10-minute policy because it achieves a lower value for longest wait as well as a lower clinic delay, and therefore we focus on comparing the analytical method to FCFS for the remainder of this section. However, compared to the FCFS policy, it obtains superior result for longest wait time at the expense of a higher clinic delay. This is because FCFS minimizes overtime by avoiding provider's idle time. The advantage of our analytical method over the FCFS policy is most evident for certain parameter configurations (Figure 3-4).

Figure 3-4. Comparison to the first-come-first-served policy in the 2-patient case





(e)



(f)

For short service times, the performance of the analytical method is similar to that of a FCFS policy, whereas the analytical method dominates for longer service times (Figure 3-4a). This observation is in line with our analytical inference that for short service times, FCFS tends to be an optimal policy (Section 3). Similarly, our conjecture that higher weight of clinic delay cost makes FCFS an attractive policy is validated by Figure 3-4b, which shows that the analytical method performs significantly better than FCFS only for small values of clinic delay cost.

Figure 3-4-c suggests that the earlier patients arrive, the larger is the advantage of using the analytical method. This is because early patients imply a smaller value for maximum lateness c, and as we have inferred analytically in Section 3, a lower value of c is associated with an optimal decision of waiting. The same reasoning holds for the standard deviation of lateness (Figure 3-4-d): the smaller the variability of lateness, the smaller is the value of c, which increases the attractiveness of the decision to wait. Also, the advantage of using the analytical method diminishes with increase in service time variability (Figure 3-4-e). This can be explained by the fact that the analytical method assumes constant service times, and its purported benefits decrease with increased variability of service times. Finally, Figure 3-4-f reports the expected clinic delay and the maximum waiting time obtained across some of the most favorable parameter combinations (d = 60 minutes, $\mu \in$ {-17, -32}, $\sigma \in$ {20, 27}, $\alpha \in$ {15, 10^5}, τ = 2). The analytical method results in a 22-minute reduction of waiting time, at the cost of a modest 5-minute increase in clinic delay.

In summary, the FCFS policy performs similar to the analytical policy if appointment slots are short, service times are exponential, patients tend to arrive late, or arrival times are highly variable. In these cases, a clinic may prefer to adopt the FCFS policy due to its simplicity. For all other cases, the analytical policy should be preferred, as it offers a superior policy by balancing the costs of patient wait times and clinic delay.

3.5. Extension to N Patients

In the prior section, we limited ourselves to dilemmas arising out of just two scheduled patients. However, clinic sessions are typically composed of $N > 2$ appointment slots. As in the 2-patient case, a wait-preempt dilemma arises whenever the provider is idle and a patient (W) arrives before an earlier-scheduled patient (T). However, there are two complications which, despite being unlikely, may arise:

1) Whereas in the 2-patient case there was only one patient (T) who could possibly arrive between the current time t_0 and the wait-up-to time x, in the N-patient case there may be others.
2) The dilemma may arise between non-adjacent appointments, i.e., W is scheduled more than a slot after T, but W arrives before T.

To find the optimal solution to the dilemma in the N-patient case, one could generalize equations (1)-(3), but this task would require considering all possible orders of arrival between N patients, which would be impractical even for small values of N. Note that complications 1) and 2) arise frequently if appointment slots are short or if patient lateness is highly variable. As noted in Section 3, in these cases there is little advantage in adopting the analytical method because a FCFS policy will tend to be optimal. In other cases, these complications will arise rarely. So, we first develop a solution method that assumes that these complications do not arise, and then suggest a heuristic to deal with these complications, should they arise.

3.5.1. Analytical Method for the N-patient Case

Here, we assume that the dilemma may arise only between adjacently scheduled patients T and W, and should the provider decide to wait for T, only T may arrive. So, we simply solve the dilemma for all $N-1$ pairs of adjacent patients (T and W) using the analytical method developed for the 2-patient case. The solution to each dilemma (T, W) provides the conditions under which it is optimal to wait for T in case W shows up sooner. However, to do so, we need to compute the appropriate value of delay cost τ such that it accommodates expected clinic overtime and expected waiting times of patients to follow.

Let n be the number of patients scheduled back-to-back after the two patients T and W, and q be the show probability of each patient. Since we still assume deterministic service times, if all these n patients show up, then each of them suffers a waiting time cost equal to ω per time unit of delay. Since the probability of this event is q^n, the total expected waiting time cost is $q^n n\omega$ per time unit of delay. Furthermore, since the delay propagates up to the end of the clinic session, the clinic incurs an expected overtime cost of $q^n \Psi$ per time unit of delay, where Ψ is the cost of clinic overtime per time unit.

However, if only the first k patients ($k < n$) show up, and the $(k+1)$-th patient does not, then only the first k patients suffer the delay, and the clinic does not incur overtime. Since this event occurs with a probability of $q^k(1-q)$, the total expected waiting time cost is $q^k(1-q)k\omega$ per time unit of delay. Note that if any of the subsequent slot is empty (i.e., not booked), the delay introduced will be certainly absorbed by that slot, without affecting either clinic overtime or waiting time of patients scheduled after the open slot. Therefore, the clinic delay cost factor can be computed as:

$$\tau = \theta_n q^n(n\omega + \Psi) + \sum_{k=1}^{n-1} q^k(1-q)\theta_k k\omega \quad (4)$$

where, θ_i = 1 if the i subsequent slots are booked and θ_i = 1 otherwise, and the right-most term (the summation) in (4) applies only when $n \geq 2$. Note that this computation overestimates the value of τ, since it assumes that none of the subsequent arriving patients shall arrive late. In reality, some of the subsequent arriving patients may arrive late, thereby suffering a lower waiting cost.

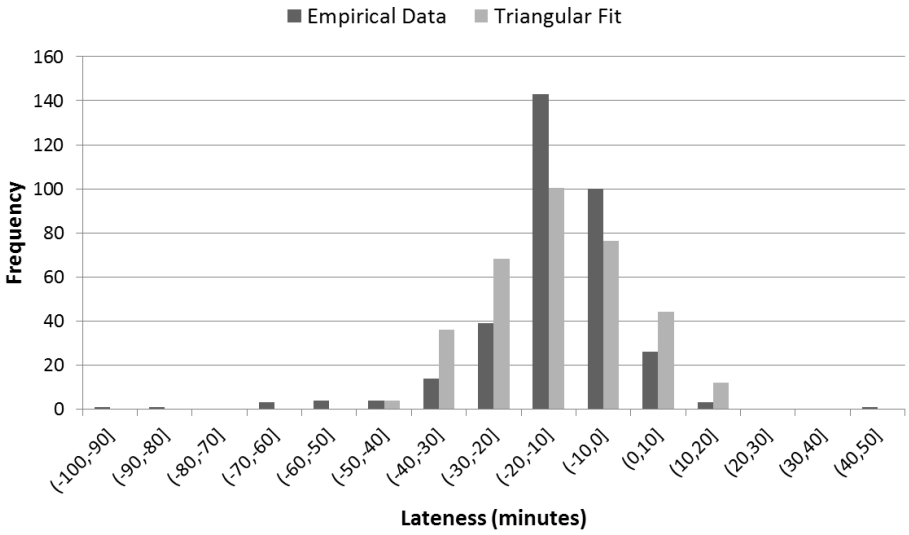
3.5.2. Experiments with N patients

We evaluate the performance of our method in the N-patient case through simulation, where each simulated clinic session consists of N appointment slots (with $N \in \{5, 10, 20\}$) of length $d \in \{30, 60\}$ minutes each. A slot is empty (i.e., unassigned to any patient) with probability $\varphi \in \{0.0, 0.1, 0.2, 0.3\}$ and booked (i.e., allocated to a patient) with probability $1 - \varphi$. Patients show up with probability $q \in \{0.6, 0.8, 1.0\}$ and their lateness is sampled from a Normal distribution with parameters $\mu \in \{-32, -17\}$ and $\sigma \in \{20, 27\}$ minutes. We consider service times that are somewhat variable (gamma distributed with $\alpha = 15$) or effectively constant (gamma distributed with $\alpha = 10^5$). Without loss of generality, we fix the waiting time cost factor to $\omega = 1$, and consider an overtime cost factor $\Psi \in \{1, 2, 5\}$. As in the 2-patient case, we executed 10,000 simulations for each parameter combination and recorded the average clinic cost, the average overtime, and the average maximum waiting time obtained using the analytical method and other policies. In contrast to the 2-patient simulations, here we do not consider certain parameter values ($\mu = -2$, $\sigma = 34$, $d = 10$, $d = 20$, $\alpha = 1$) because, as shown in Section 3, the FCFS policy performs almost optimally in these cases.

Aside from testing our method with "simulated arrival times" as described above, we also consider empirical arrival times, which have been observed at a university campus outpatient clinic. We collected 339 samples of "lateness", as measured by the difference between check-in time and appointment time. The average lateness μ is equal to -13.7 minutes and the standard deviation σ is equal to 13.3 minutes. The low value of μ may be influenced by the clinic's practice of instructing patients to arrive 15 minutes before their appointment times, while the small value of σ may be explained by the proximity of the patients to the clinic (most patients are students, who live on campus). To set up our analytical method, we fit the empirical lateness data with a triangular distribution, obtaining the distribution T(-46.2, -13.7, 18.8). Figure 3-5 shows the empirical frequency in comparison to that obtained by the triangular fit. For the empirical arrival times, we executed 10,000 simulations for several combinations of parameters and compared the performance of the analytical method to that of other

policies. In particular, we executed simulations for the same combinations of parameters as for the simulated data except the arrival time parameters, μ and σ. Instead of simulating arrival times, here we simply sample them from the empirical data.

Figure 3-5. Frequency of Lateness Occurrences



During the experiments with both the "simulated"- and "empirical"-data, the two assumptions stated at the beginning of this section may not hold. First, it is possible that the dilemma arises between non-adjacent appointments. In this case, we solve the dilemma using a heuristic, which we call the *chain rule*. Let M_1, M_2, …, M_z be the sequence of appointments scheduled between the target patient T, represented here by M_0, and and patient W, represented by M_{z+1}. Without loss of generality, we can assume that M_1, M_2, …, M_z have not shown up yet; if one of them had showed, a new dilemma can be defined where this appointment takes the part of W. For each $i = 0, \dots, z$, consider the pair of adjacent appointments (M_i, M_{i+1}). If at the current time a wait-preempt dilemma arose between M_i and M_{i+1} and its optimal solution was to wait, it would mean that even if M_{i+1} showed up, we should wait for M_i. For better reason, then, W should wait, because scheduled after M_{i+1}. So, for each $i = 0, \dots, z$, the chain rule solves an hypothetical dilemma arising between (M_i, M_{i+1}). If the optimal solution of any of these dilemmas is to wait, then W should wait; otherwise, we should preempt. In case of wait decision, the

length of the wait should be greater than the wait lengths of the optimal solutions found throughout the chain. After this time, if no more patients have arrived, we re-solve the dilemma through the chain rule.

Second, whether or not T and W are in adjacent appointment slots, if the provider has chosen to wait for patient T, it is possible for a different patient Z to arrive before T. If Z is scheduled before W, then Z is the new "waiting" patient, and we solve the dilemma between T and Z. If Z is scheduled after W, then Z should obviously wait (because it has already been determined that W should wait). However, the clinic delay factor τ used to solve the dilemma arising between T and W must be recomputed, to account for the fact that Z has shown up, and the dilemma between T and W must be resolved.

We compared the average performance measures obtained by the analytical method to those obtained by the FCFS policy. We excluded the 10-minute policy from our analysis because it is outperformed by the analytical method (see Section 4). Results are summarized in Table 3-4.

Table 3-4. Results in the *N*-Patient Case

Arrival time	Policy	Cost	Longest wait (minutes)	Overtime (minutes)
Simulated	First-come-first-served	1.14	14.63	4.39
	Analytical method	1.10	11.73	4.81
Empirical	First-come-first-served	1.00	17.23	3.71
	Analytical method	0.83	8.22	4.08

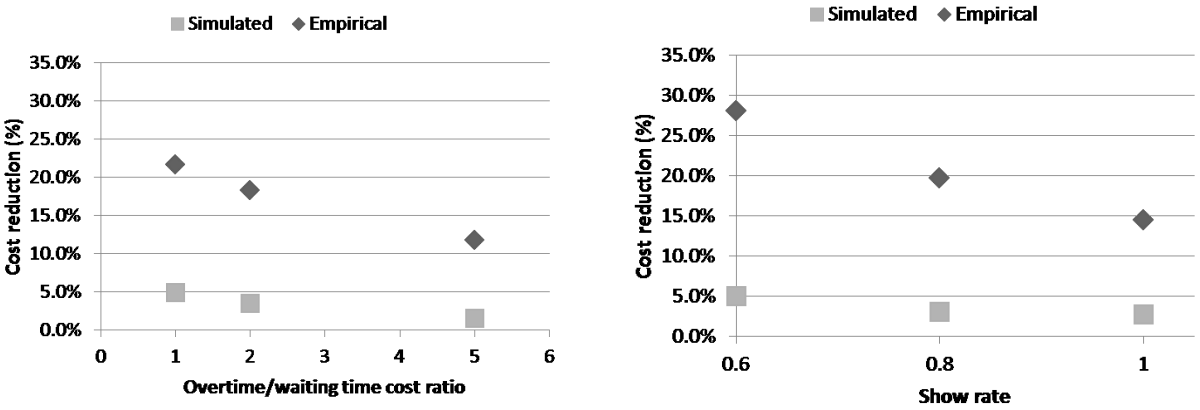
As in the 2-patient case, the adoption of the analytical method results in a great decrease in waiting time and a small increase in overtime compared to using FCFS. However, for simulated arrival times, the magnitude of the improvement is much smaller than that obtained in the 2-patient case. We attribute this to the simplifications necessary to tackle the N-patient case, in particular the relaxation of the two complications described at the beginning of this section. Despite this relatively modest performance, it is possible to show that the magnitude of the improvement is affected by the parameters in

the same way as for the 2-patient case. In particular, it is still true that the greatest advantage of the analytical method is achieved for those parameter combinations that lead to the maximum lateness c being much smaller than the service time d ($c \ll d$). This explains why the analytical method performs much better on the empirical data than on the simulated ones: with the empirical arrival distribution, for $d \in \{30, 60\}$, it is highly unlikely for a patient to arrive more than an appointment slot late (see Figure 3-5).

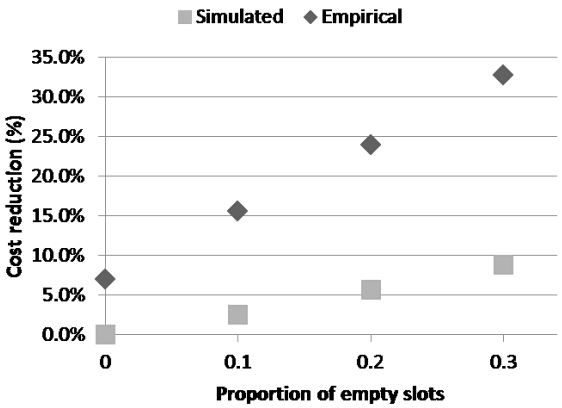
Finally, we consider a variation of the first-come-first-served policy, the first-scheduled-first-served (FSFS) policy, under which the provider sees the patient scheduled earliest among the patients who are waiting. FSFS provides a fairer queue discipline for preemption than FCFS, because if multiple patients are waiting at a clinic, a strict enforcement of FCFS can lead to a deliberate violation of scheduled sequence. The first-scheduled-first-served policy outperforms the first-come-first-served policy: it leads to the same overtime but results in a 0.76% reduction in maximum waiting time and a 0.94% reduction in cost. A paired Student t-test shows that the difference in cost is significant (p value < 0.01).

However, our analytical method outperforms FSFS: Figure 3-6 reports the cost reduction obtained by adopting the analytical method over the first-scheduled-first-served policy under different parameter configurations.

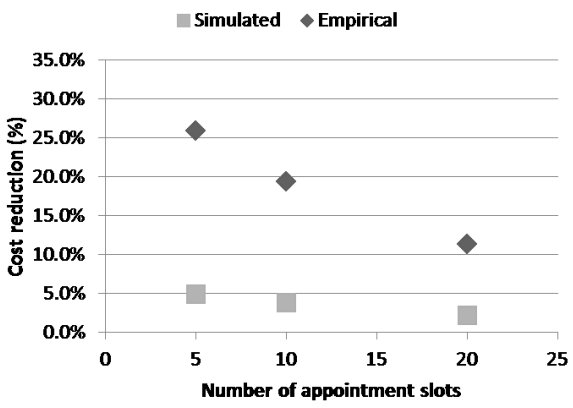
Figure 3-6. Comparison to the first-scheduled-first-served policy in the *N*-patient case



(a)



(b)



(c)

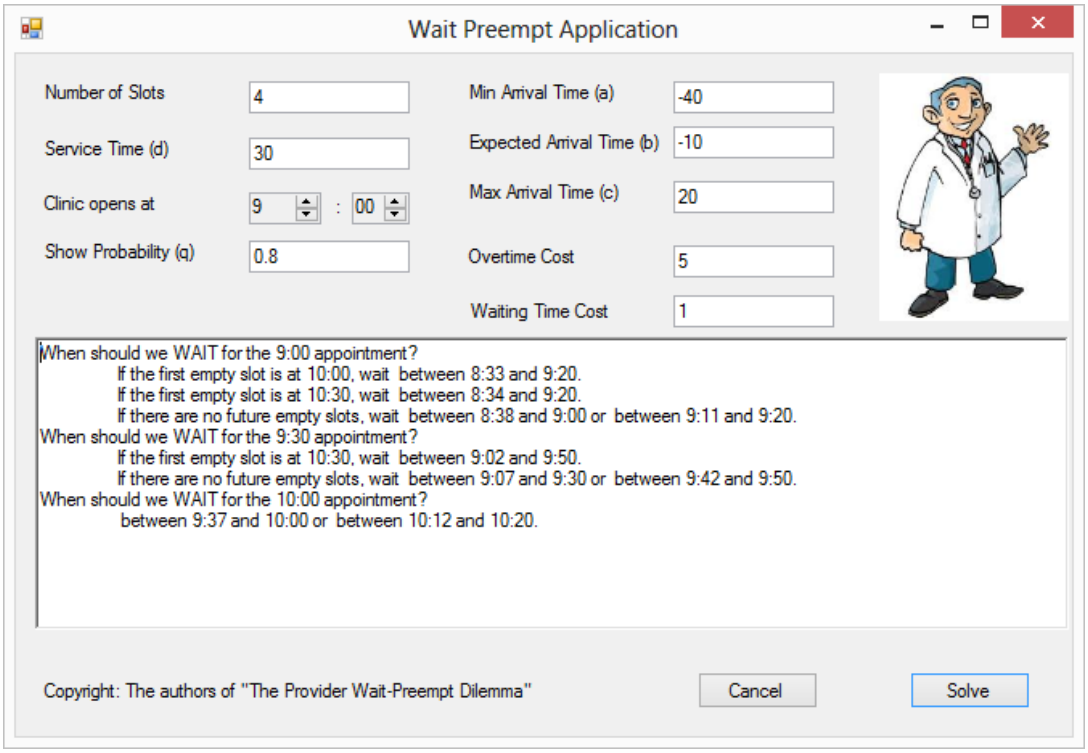
(d)

Figures 6-a, 6-b, and 6-c show that the analytical method outperforms FSFS if the overtime cost is large compared to the waiting cost, if the show rate is low, or if the proportion of empty slots is high. This is explained, once again, through the structural properties of the cost function. In Section 3, we noted that the optimal decision tends to be to wait when the delay cost factor is small compared to the waiting time cost factor ($\tau < \omega$). So, we expect the analytical method to outperform FSFS if $\tau < \omega$ and the two policies to perform similarly if $\tau > \omega$. By considering (4), it is clear that τ is small if the overtime cost Ψ is low (Figure 3-6-a), if the show rate q is low (Figure 3-6-b), or if the proportion of empty slots is high (Figure 3-6-c). Interestingly, FSFS performs poorly for low show rates. In this case, in fact, it may be more desirable to wait than it is to preempt, because any delay induced is likely to be absorbed before propagating to the end of the clinic session. Finally, Figure 3-6-d shows that the larger the number of appointment slots, the smaller the advantage of the analytical method. This phenomenon can be attributed to two distinct factors. First, larger value of patients-to-follow n implies larger weight on clinic delay τ, which reduces the attractiveness of the analytical method. Second, the extension of the analytical method to the N-patient case does not consider the two possible complications defined at the beginning of this section, which are more likely to arise if the number of slots is higher.

In summary, the analytical method outperforms the FCFS and FSFS policies in the N-patient case; for favorable parameter configurations, the advantage is substantial.

3.6. Implementation and Extensions

Figure 3-7. Snapshot of the Software Application



Once clinic parameters are known, implementing the analytical method to make decisions is relatively straightforward. However, to make it easier for the clinic management to adopt our method, we provide a simple software application (provided as an online supplement and shown in Figure 3-7), which may be used to obtain easy-to-follow guidelines for wait-preempt decisions. For a given parameter configuration, the application finds the intervals when the provider should wait for each patient T, should a dilemma arise. In the example of Figure 3-7, we consider a clinic session comprising 4 appointment slots of 30 minutes each (starting at 9:00, 9:30, 10:00, and 10:30). Assuming that there are no future empty slots, one of the rules says that if the 10:00 patient arrives before the 9:30 patient, then the provider should wait in two different intervals: (i) between 9:07 and 9:30 (i.e., shortly before the missing patient's appointment

start time); and (ii) between 9:42 and 9:50 (i.e., when the missing patient is more likely to be a "no show").

Limitations of our method in the N-patient case include the assumption of constant service times and that of future patients' punctuality. To take these factors into account, we implemented a solution method based on simulation optimization. When the dilemma arises, this method generates scenarios that represent possible realizations of stochastic events, such as patient no-show and arrivals. Then, the method compares the average performance across the scenarios for the wait and the preempt decisions, and chooses the one that leads to the lower cost. Because of the higher computation time needed to execute this procedure, we tested it only on a subset of the N-patient simulations of section 5. Our results show that the simulation optimization approach leads to an average of 1.3% cost reduction compared to the analytical approach. We conjecture that this difference is due to the assumptions made to develop the analytical approach for the N-patient case, which are not needed to develop the simulation optimization approach. If on one hand the simulation optimization approach performs slightly better, on the other hand the analytical approach is easier to implement and provides intuitive managerial insight as to why one should wait or preempt, and therefore has a higher potential of being adopted by practitioners.

3.7. Conclusions

How to deal with patients arriving out of order is determined by the answer to a seemingly simple question: to wait or not to wait? That is the dilemma that an idle provider faces when, while waiting for the next scheduled patient (target patient T) to show up, a later scheduled patient arrives (waiting patient W). Should the provider keep waiting for patient T or see patient W right away (i.e., preempt)? Waiting may result in overtime, but preempting may result in a long waiting time experienced by T.

This study introduces an analytical method that informs clinic managers about the decisions they should make when faced with such dilemma. First, we solve the dilemma in a simplified setting with only two patients in order to generate insight; then, we extend our method to an arbitrary number of scheduled patients. Our objective is to find the time intervals in which the provider should wait for the

late patient and those where she should preempt and see the early patient. We find these intervals by analyzing a cost function, which, for each point in time x, computes the expected cost incurred if the provider waits up to time x for the late patient. The cost function is a weighted sum of the patients' waiting time and the clinic overtime.

Although existing literature reports that in practice the dilemma is generally resolved by choosing to "preempt" (often implemented with a first-come-first-served policy), our analytical results show that under certain situations, the provider should decide to wait instead, even in the presence of a waiting patient. Our work provides interesting managerial insights and identifies the configurations of clinic parameters for which our proposed method makes the most impact. In general, if the dilemma arises long before the target patient's scheduled appointment, then the provider should preempt; in that case, the target patient will suffer little or no delay. On the other hand, if the dilemma arises shortly before the target patient's scheduled appointment, then the provider should wait. Interestingly, this is true for any parameter configuration. However, if the dilemma arises after the target patient's scheduled appointment, then the probability of a "preempt" decision being preferable to a "wait" decision is positively correlated to the ratio between maximum lateness and slot length and to the ratio between delay cost factor and waiting time cost factor. Preempting tends to be preferable whenever these ratios are high; for example, for short appointment slots, very variable arrival times, very late patients, and high overtime cost factors. In these situations, the performance of our analytical method is similar to that of the first-come-first-served-policy (FCFS). However, for the other situations, our simulation experiments show that our analytical method outperforms FCFS, reducing the total costs of patient wait and overtime by up to 20%. In order to make it easy for clinic managers to reap the benefits of our work, we provide a free and easy-to-use software application.

Chapter 4 - SOLVING A CYCLIC AND BATCHING SCHEDULING PROBLEM WITH TWO TYPES OF SETUPS

In Ganguly and Laguna (2013), we develop a mixed-integer linear program and two alternative heuristic solution methods for a particular type of job sequencing problem that involves two different types of setup.

4.1. Introduction

Many production systems with closed-loop facilities must deal with the problem of scheduling batches in consecutive loops (e.g. electrolytic painting of automotive parts in closed conveyors and cyclic painting of metallic furniture). We study such a scheduling problem, which arises in a production system at a manufacturing facility for plastic auto parts. The system combines cyclic and batch- scheduling to paint rearview mirrors of varying geometries and colors. The paint line consists of a moving train that forms a continuous loop and contains a fixed number of spaces (positions). A jig is placed in each position in order to hang one or more parts. The sequence in which the parts are hung on the jig determines the number of setups that are incurred. A setup occurs when there is a change in the part geometry and also when there is a change in the color.

Each part geometry requires a special jig but the same jig can be used to paint parts of different colors. Therefore, the jig must be changed every time the geometry of the part to be painted changes. The jigs do not have to be changed when the color changes and the geometry of the parts remains the same. However, a change of color requires the use of a solvent to clean the pipes in which the paint flows. The painting area is located at a fixed position in the line and the problem is to minimize a cost function associated with the setups caused by changing both colors and geometries.

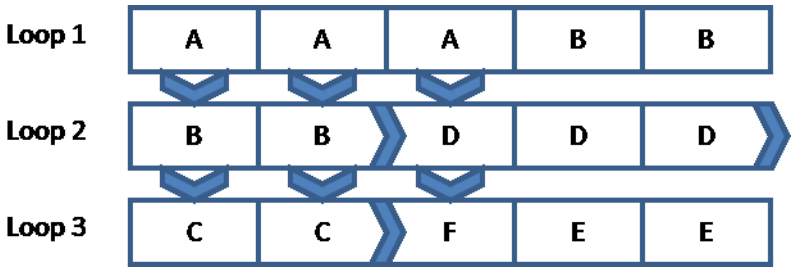
A simple example may be used to illustrate the problem. Suppose that 6 distinct products (labeled A through F) with the characteristics and demand values (in parentheses) shown in Table 4-1 must be painted. For instance, there is a demand of 3 parts for product A, which has color 1 and geometry 1. Also suppose that the paint line is such that there are 5 positions per loop (PPL). Therefore, 3 loops are

necessary to paint all 15 parts (i.e., the sum of all the demand values in Table 4-1). A possible solution to this problem is shown in Figure 4-1, where the product labels are used to indicate the characteristics of the parts assigned to each position.

Table 4-1. Characteristics and demand values (number of parts) of 6 products (A to F)

	Geometry		
Color	**1**	**2**	**3**
1	A(3)	B(4)	C(2)
2		D(3)	
3		E(2)	F(1)

Figure 4-1. Positioning of 15 parts in 3 loops with $PPL = 5$



The thick arrows in Figure 4-1 indicate the setups that occur when placing the products in the specified positions. There are a total of three color setups and six jig setups. Solvent and labor costs are incurred when color setups occur. When the jig changes from one loop to the next, a labor cost is incurred because labor is required to make the change. Following the schematic representation of a solution to the problem presented in Figure 4-1, we will refer to a change of color as a *horizontal* setup while a change of jig as a *vertical* setup.

Scheduling problems with setups have received considerable attention in the operations research literature (Cheng et al. 2000, Zhu and Wilhelm 2006, Allahverdi et al. 1999, Allahverdi et al. 2008); however, the particular problem that we have described above — to the best of our knowledge — has not been discussed, with the exception of Garcia-Sabater et al. (2008). This work, however, has a narrow scope because it is limited to describing the problem, introducing a mathematical formulation and solving a single instance of the problem with commercial software. While some work has been devoted to scheduling problems with sequence dependent setups (Szwarc and Gupta 1987, Laguna 1999, Rajendran

and Ziegler 2003, Ruiz et al. 2005 and Jungwattanakit et al. 2009), the extent of sequence dependence addressed in these works is limited to adjacent jobs. In other words, these works have typically assumed that the setup cost incurred by the current job is dependent only on the current job and its immediate predecessor. In contrast, this paper addresses not only the horizontal setup cost due to the current job and its immediate predecessor, but also the vertical setup cost incurred by the current job and the job that occupied the same position in the prior loop. As we will show here, consideration of these two types of setup costs adds significant complexity to the problem.

The balance of the paper is organized as follows. In the next section, we present a mathematical programming formulation of the problem and show that it is not practical to use this approach to solve industrial scale problems. Thereafter, we analyze bounds and the landscape of the problem, an analysis that helps us choose the appropriate heuristic methods. Then, we present two alternative heuristic approaches: one based on commercially available software and the other a procedure developed specifically for the problem at hand. Finally, we present experimental results and discuss the relative merits of the alternative heuristic approaches.

4.2. Notation and Mathematical Programming Formulation

The problem that we described above may be formulated as a mixed integer linear programming (MILP) model. For this purpose, we employ the following notation that assumes that a product is unique combination of color and geometry:

C	:	set of colors
G	:	set of geometries
S	:	set of products $\subset (C \times G)$
CC	:	set of pairs of dissimilar colors $\subset (C \times C)$
GG	:	set of pairs of dissimilar geometries $\subset (G \times G)$
LB_c	:	lower bound on number of color changeovers, default 0
LB_g	:	lower bound on number of geometry changeovers, default 0
LB	:	Lower-bound on objective-function value (from the best node in the branch-and-bound tree, rounded up to nearest feasible objective value)
$h_{ii'}$	:	(horizontal) setup cost of changing from color i to i'
$v_{jj'}$	:	(vertical) setup cost of changing from geometry j to j'
$d_{(i,j)}$	:	demand for product with color i and geometry j, for $(i,j) \in S$
PPL	:	parts per loop

n : number of positions, where $n = \sum_{(i,j)\in S} d_{ij}$

The MIP model is formulated with the following variables:

Primary Variables:

$$x_{ijk} = \begin{cases} 1 \text{ if product } (i,j) \text{ is placed in position } k \\ 0 \text{ otherwise} \end{cases}$$

Auxiliary Variables:

$$y_{ik} = \begin{cases} 1 \text{ if a part with color } i \text{ is located in position } k \\ 0 \text{ otherwise} \end{cases}$$

$$z_{jk} = \begin{cases} 1 \text{ if a part with geometry } j \text{ is located in position } k \\ 0 \text{ otherwise} \end{cases}$$

$$p_{ii'k} = \begin{cases} 1 \text{ if color changes from } i \text{ to } i' \ (ii' \in CC) \text{ between positions } k \text{ and } k+1 \\ 0 \text{ otherwise} \end{cases}$$

$$q_{jj'k} = \begin{cases} 1 \text{ if geometry changes from } j \text{ to } j' (jj' \in GG) \text{ between positions } k \text{ and } k+PPL \\ 0 \text{ otherwise} \end{cases}$$

The mathematical formulation has the following form:

Minimize $\sum_{k=1}^{N-1} \sum_{(i,i')\in CC} h_{ii'} p_{ii'k} + \sum_{k=1}^{N-PPL} \sum_{(j,j')\in GG} v_{jj'} q_{jj'k}$ (1)

Subject to

$\sum_{(i,j)\in S} x_{ijk} = 1$ $k = 1, \dots, n$ (2)

$\sum_{k=1}^{n} x_{ijk} = d_{ij}$ $\forall (i,j) \in S$ (3)

$y_{ik} = \sum_{j:(i,j)\in S} x_{ijk}$ $\forall i: \exists (i,j) \in S, k = 1, \dots, n$ (4)

$z_{jk} = \sum_{i:(i,j)\in S} x_{ijk}$ $\forall j: \exists (i,j) \in S, k = 1, \dots, n$ (5)

$p_{ii'k} \geq y_{ik} + y_{i'k+1} - 1$ $\forall (i,i') \in CC, k = 1, \dots, n-1$ (6)

$q_{jj'k} \geq z_{jk} + z_{j'k+PPL} - 1$ $\forall (j,j') \in GG, k = 1, \dots, n-PPL$ (7)

$x_{ijk} \in \{0,1\}$ $\forall (i,j) \in S, k = 1, \dots, n$ (8)

The objective is to minimize the total cost associated with the changes in color and jigs. Constraint set (2) enforces the assignment of a single product to each available position; while set (3) guarantees that the demand for each product is satisfied. In the remainder of this article, the demand for product with color i and geometry j is denoted by. Based on the values of the primary variable x, constraint sets (4) and (5) fully determine the values for the auxiliary variables y and z respectively. Constraint sets (6) and (7) capture the color changeovers p and geometry changeovers q respectively. Constraint set (8) enforces the

integrality of the primary binary variables x. Note that the integrality of the auxiliary variables p, q, y, and z are implicitly enforced by the constraint sets (4) through (7) and the minimization of the objective function.

Our MIP formulation is more general than the one introduced by Garcia-Sabater *et al*. (2008) because it allows for setup costs to depend on the products. Also, our modeling of the changeovers, represented by constraints (4) through (7), is done without reference to color IDs, as done in Garcia-Sabater *et al*. (2008). While the MIP model above is a valid formulation of the problem, it has a weak LP-relaxation and experimentation shows that it takes a long time to converge to optimality except for very small instances. In preliminary experimentation, the LP relaxation often found optimal non-integer solutions with values of the y variables such as those shown in Table 4-2, without incurring any color changeover costs.

Table 4-2. Values of y that are feasible for LP relaxation of the MILP model

Variable y	**Color 1**	**Color 2**	**Color 3**
Position k	0.5	0.5	0.0
Position k+1	0.0	0.5	0.5

A structure similar to the one in Table 4-2 was observed in geometry changeovers as well. To excise such non-integer solutions from the formulation, we added a set of cuts (9 and 10 below) that strengthen the LP relaxation and improve performance. Further, we added constraint sets (11) and (12) to provide lower bounds of color and geometry changeovers. The bounds LB_c and LB_g are obtained using the approach described in the next section. The bound on the objective function value (13) is added during the branch and bound process because, without loss of generality, we consider that both h and v are integers. Therefore, anytime that the lower bound (LB) on the total cost becomes fractional, the value is rounded up to the nearest integer value ($\lceil LB \rceil$), and thus helping with the convergence.

$$\sum_{i \neq i'} p_{ii'k} \geq y_{i'(k+1)} + \sum_{i \neq i'} y_{ik} - 1 \qquad \forall i' \in C, k = 1, \dots, n-1 \tag{9}$$

$$\sum_{j \neq j'} q_{jj'k} \geq z_{j'(k+PPL)} + \sum_{j \neq j'} z_{jk} - 1 \qquad \forall j' \in G, k = 1, \dots, n-PPL \tag{10}$$

$$\sum_{k=1}^{n-1} \sum_{(i,i') \in CC} p_{ii'k} \geq LB_c \tag{11}$$

$$\sum_{k=1}^{n-PPL} \sum_{(j,j')\in GG} q_{jj'k} \geq LB_g \quad (12)$$

$$\sum_{k=1}^{n-1} h_{ii'} p_{ii'k} + \sum_{k=1}^{n-PPL} v_{jj'} q_{jj'k} \geq \lceil LB \rceil \quad (13)$$

In the experimental section, we show that despite of our efforts to add cuts and improve bounds, our experience with the MIP formulation has severe limitations as a practical method for solving industrial scale instances of this problem for which $|C| \geq 10$, $|G| \geq 10$, and $|PPL| \geq 50$. We even considered a simplified MIP formulation and found similar practical limitations.

4.3. Bounds and Landscape Analysis

Before embarking on the task of developing an approximate solution method we analyzed the structure of the problem with the goal of gaining additional insights as well as producing bounds for the two types of setups. Understanding the landscape associated with the objective function was part of this analysis.

If we consider that the setup cost is constant and does not depend on the specific colors or geometries involved in the change, then lower bounds on the number of setups of each type (i.e., LB_c and LB_g) may be used to calculate a lower bound on the cost. (For problems where it is absolutely necessary to consider variable setup costs, the lower bound could be calculated using the minimum horizontal and vertical costs.) If vertical setups are ignored, a bound on the number of color setups is given by $LB_c = |C| - 1$. That is, in the absence of vertical setups, the optimal way of arranging the parts is by blocking all parts of the same color and sequentially process all blocks in any order, as long as all parts with the same color are processed together.

Now let us disregard the horizontal costs and focus on the vertical costs. A solution may be conceptualized as a two-dimensional matrix with number of columns equal to PPL and number of rows equal to n/PPL. (See, for instance, Figure 4-1 where a solution to a problem with $n = 15$ and $PPL = 5$ is depicted.) The demand associated with each geometry j is given by $d_{(*,j)} = \sum_i d_{(i,j)}$. Define $m_j = mod\left(\frac{d_{(*,j)}}{n/PPL}\right)$. It is trivial to show that $LB_g = 0$ if $m_j = 0$ for all j. In this case, each part geometry is placed in as many columns as needed to meet the demand; each column includes only one geometry, and the number of geometry setups is zero. However, when $m_j \neq 0$ for at least one $j \in G$, geometry setups

become unavoidable. This means that when horizontal costs are ignored, the problem of minimizing vertical costs can be reduced to a smaller problem in the following manner. For each geometry $j \in G$, $\left\lfloor \frac{d_{(*,j)}}{n/PPL} \right\rfloor$ gives the number of columns that may be fully occupied by these items without incurring any changeover. The reduced problem then consists of placing m_j units for each $j \in G$ in a matrix with the same number of rows as the original problem, i.e., n/PPL, and a reduced number of columns. The reduced number of columns, or reduced parts-per-loop ($RPPL$), is given by

$$RPPL = PPL - \sum_j \left\lfloor \frac{d_{(*,j)}}{n/PPL} \right\rfloor = \frac{\sum_j m_j}{n/PPL}$$

Moreover, since we are ignoring horizontal costs altogether, the variables and constraints related to color changeovers can be omitted from the formulation shown in the previous section. Thus, ignoring horizontal-costs and considering only geometry-costs allows us to reduce the optimization problem both in scale and complexity. The minimum number of geometry changeovers obtained by solving this reduced model is a true lower bound on the geometry-changeovers (LB_g).

While a lower bound on the total cost can be obtained by computing the number of setups separately, simple examples show that a solution method that focuses on each type of setup one at a time results in solutions that are not only suboptimal but also significantly inferior than those found by strategies that address both setups at the same time. That is, we have verified that a procedure that minimizes one type of setups first and then minimizes the second type (considering only solutions that do not increase the number of setups of the first type) is destined to perform poorly.

A natural representation of a solution to the problem consists of a vector π of size n, where $\pi(k)$ is the index of the product in position k. Simplifying our previous notation, let d_p be the demand of product p (that is, we assume that the index of the product is enough to specify its color and geometry). If for a given problem instance, the demand for each product is 1 (i.e., $d_p = 1$ for all p), then $n = |S|$ and the solution space consists of $n!$ solutions. In general, however, the size of the solution space is given by the following product, where C_x^y represents the binomial function $\binom{y}{x}$:

$$\prod_{p=1}^{p=|S|} C_{d_p}^{n-\sum_{p'=0}^{p'=p-1} d_{p'}}$$

Suppose that $h = v = 1$. Then, the total cost associated with a given solution is the same as the total number of setups necessary to implement the solution. An upper bound on the number of horizontal setups (and therefore on the horizontal cost when $h = 1$) is $n - 1$. This occurs in the worst possible scenario, which forces a change of color after every position in the solution. Likewise, $n - PPL$ is an upper bound on the number of vertical setups. This occurs when the parts are arranged in such a way that changing a jig is necessary in every position after each loop. Therefore, when considering unit costs, an upper bound on the total cost is given by $2n - PPL - 1$.

This analysis shows that the natural solution representation provided by the π vector with n positions results in a solution space with a flat landscape, that is, one for which many solutions have the same objective function value. For instance, using the data in Table 4-1 —which corresponds to a problem with 6 products (i.e., $|S| = 6$, a total demand of 15 (i.e., $n = 15$) and a configuration with 5 parts per loop (i.e., $PPL = 5$)— it is easy to verify that, if $h = v = 1$, the number of unique objective function values is only about 6.08×10^{-8} of the size of the solution space. The solution space associated with π consists of 378,378,000 solutions. An upper bound on the total cost is 24 ($2n - PPL - 1 = 2 \times 15 - 5 - 1$) and a lower bound is 2 ($LB_c = |C| - 1 = 3 - 1$), resulting in at most 23 unique objective values for 378,378,000 solutions. Similar results are obtained with problems of larger size.

4.4. Heuristic Solution Methods

In order to tackle instances that are typical in real settings, we approached the problem with metaheuristic technology. Practitioners, whenever possible, tend to prefer the use of commercial software instead of developing specialized procedures for each optimization problem that they face. The effort required to adapt commercial software to produce solutions of acceptable quality, however, depends on the optimization problem of interest. Hence, for certain types of problems, development of a specialized metaheuristic procedure that is tailored to the problem yields better results. In this work, we take both

avenues and compare their merits. In particular, we compare the adaptation of the commercial software OptQuest with a specialized VNS (variable neighborhood search) and discuss the pros and cons of following these approaches. The choice of our particular implementation of VNS over other metaheuristics was driven by the structure of the problem landscape discussed in the last section.

4.4.1. OptQuest Adaptation

OptQuest is a general-purpose optimization system that is built on a platform of metaheuristic technologies (Laguna 2011). The main engine consists of an implementation of scatter search that includes specialized memory structures typically found in tabu search implementations. OptQuest is designed to operate in a black-box structure and therefore it does not take advantage of the specific characteristics of the problem that it is attempting to solve. The main choice that OptQuest users must make relates to the selection of a solution representation. The OptQuest system offers several choices that include continuous, discrete and binary variables. A basic OptQuest implementation is as follows:

1. Define n discrete variables within the range $(1, |S|)$
2. Create an `Evaluate()` method that calculates the objective function value associated with solutions represented by π

The OptQuest Engine calls the `Evaluate()` method every time a solution needs to be evaluated. The optimization problem for OptQuest consists of selecting the value of n discrete variables that are bounded between 1 and $|S|$. Since no information is given to OptQuest about the demand for each product, the system may produce infeasible solutions. Therefore, the `Evaluate()` method must be designed in a way that infeasible solutions are either penalized or are mapped into feasible ones. Note that the absence of this information makes the space where OptQuest searches for a solution much larger than the original. The search space contains $|S|^n$ solutions.

Our mapping from an infeasible solution to a feasible one is straightforward. Given a solution , we first evaluate its objective function value and at the same time we calculate $parts(p)$, the number of parts of product p in the solution. If $parts(p) > d_p$ then product p has a surplus of parts in the solution. If $parts(p) < d_p$ then product has a shortage of parts in the solution. Because the solution has n

positions and n is equal to the total demand for parts, a feasible solution is one where $parts(p) = d_p$ for all p. We identify the best exchange of a product that has a surplus with a product that has a shortage. The best exchange is the one that either decreases the objective function value the most or increases it the least. After the change is made, the $parts(p)$ values are updated and the procedure searches for the next best exchange. The process ends when there are no products with either surplus or shortages.

Figure 4-2. Feasibility mapping within the `Evaluate()` method

1. Calculate the objective function and $parts(p)$

While ($parts(p) \neq d_p$ for at least one p)

2. Find the best exchange of a part corresponding to a product with a surplus and a part corresponding to a product with a shortage
3. Make the change, update the objective function value and $parts(p)$

The mapping within the `Evaluate()` method (shown in Figure 4-2) changes the values of the decision variables for infeasible solutions that are turned into feasible ones. When this occurs, there are two choices: 1) the updated solution along with the objective function value are returned to the OptQuest engine or 2) only the objective function value is returned to the OptQuest search engine. Previous experience with OptQuest and similar mapping strategies have shown that performance is enhanced when the second option is used. The first option leads to a highly focused search that lacks diversity and this is why it is not preferred.

In addition to evaluating the objective function of trial solutions built by OptQuest and performing feasibility mapping, the `Evaluate()` method may be employed to execute neighborhood explorations. For instance, Ugray et al. (2007) combine OptQuest with a gradient-based local solver for nonlinear (NLP) programming problems. In this implementation, the NLP solver not only turns infeasible solutions (constructed by the OptQuest engine) into feasible ones but also searches for improved outcomes. When solutions are modified by the NLP solver, the updated variable values are returned to OptQuest. In order to avoid premature convergence of the search, the NLP solver is called selectively. The procedure uses both a distance and a merit filter. As stated by the authors, "the distance filter helps insure that these starting points are diverse, in the sense that they are not too close to any previously found local solution." Likewise, "the merit filter helps insure that the starting points have high quality, by not

starting from candidate points whose exact penalty function value is greater than a threshold." In the Ugray, et al. (2007) design, OptQuest is treated as an engine whose goal is to provide a set of high quality and diverse starting points for a local optimizer.

We follow a similar approach and in addition to mapping infeasible solutions into feasible ones, we add an improvement procedure, which is run immediately after a feasible solution is identified within the call to the `Evaluate()` method. The improvement method is a simple short-term memory tabu search (see Figure 4-3). An iteration involves evaluating all possible swaps of different parts. The non-tabu swap that has the best move value (i.e., that decreases the current objective function the most or increases it the least) is executed. A $\text{swap}(k_1, k_2)$ is tabu if the parts in positions k_1 and k_2 have been swapped in the last $|S|$ iterations. A tabu swap is executed if it leads to a new incumbent solution. After a swap of the parts in positions k_1 and k_2, the pair (k_1, k_2) is made tabu-active for $|S|$ iterations. The search stops after $|S|$ iterations have been performed without finding a new incumbent solution. Note that the incumbent solution is the best solution found during the current tabu search, that is, during the search that starts in the current call to the `Evaluate()` method. The overall best solution is the best incumbent found after all the OptQuest iterations have been performed.

Figure 4-3. Short-term tabu search within the `Evaluate()` method

1. Make current solution the incumbent

While (new incumbent solution found in the last $|S|$ iterations)

2. Find the best non-tabu swap (waving the tabu status if the swap leads to a new incumbent)
3. Make the swap, update the tabu memory and the incumbent solution if appropriate

The procedures in Figures 2 and 3 require a very modest implementation effort. In fact, our implementation in C# consists of fewer than 100 lines of code that include both the feasibility mapper and the improvement method. The rest of our code is standard to formulating optimization problems with OptQuest and was in fact adapted from one of the examples that are distributed with the OptQuest engine. We mention this because a discussion that is often omitted, but that we believe is important, relates to the total amount of effort involved in developing solution methods for practical problems. We now turn our

attention to the description of a specialized search heuristic that we have developed for the dual-setup problem.

4.4.2. Variable Neighborhood Search (VNS) Approach

As an alternative to the method described in the previous section, we implemented a multi-start VNS procedure. VNS is based on the notion of creating multiple neighborhoods of varied complexity (Talbi 2009). The main elements of our implementation are:

1. Pseudo-greedy constructions
2. Swap neighborhood
3. Shift neighborhood
4. Shake

The search starts from multiple points that are constructed as follows. The starting point alternates between a completely random solution and a greedy random solution. Our experiments showed that this mix of starting points helped obtain solutions of higher quality by combining the diversity created by totally random solutions with the starting points created by the greedy random solutions. The greedy random construction begins by assigning a random part to the first position. Then, the subsequent positions are filled sequentially by the remaining parts using the following greedy rule. The most preferred part for a position is one which will not result in any changeover, e.g. for position j, the color should match with position $j-1$(when $j>1$), and the geometry should match with position $j-PPL$(when $j>PPL$). The next preferred parts are those which avoid one type of changeover, either color or geometry. When no preferred parts are available, a part is chosen at random from those with unfulfilled demand. The primary neighborhood search in our implementation consists of swaps of parts that are not of the same product. We again represent a solution with a vector π of size n, resulting in an upper bound of at most $(n^2-n)/2$ moves. This only occurs when $d_p=1$ for all p. In fact, the total number of swaps —in terms of the demand for each product— is given by the following formula:

$$\sum_{p=1}^{|S|-1} d_p \left(\sum_{p'=p+1}^{|S|} d_{p'} \right)$$

The value of a swap is easy to compute. The calculation consists of at the most of sixteen simple operations, eight to calculate the "savings" of removing the parts from the current positions and eight to calculate the "cost" of inserting the parts in their new positions. Note that each part has at most four immediate neighboring parts (left, right, up and down, if we picture a solution as given in Figure 4-1).

In addition to swap, our VNS uses shifts. A shift may be thought of as a partial swap in the sense that a single part is taken from its current position and is transferred to another position in the solution. Here again, we don't have to examine all shifts because some of them do not produce a change of the current solution. For instance, consider the solution depicted in Figure 4-1 and its representation as a vector of product IDs:

$$\pi = (A, A, A, B, B, B, B, D, D, D, C, C, F, E, E)$$

Clearly, relocating $\pi(1) = A$ between positions 2 and 3 does not change the solution because $\pi(2) = \pi(2) = A$. In fact, the solution does not change either if $\pi(1)$ is relocated between positions 3 4. This shows that the set of "valid" shifts depends on the configuration of the current solution. The number of shifts however is in the same order of magnitude as the number of swaps. The VNS operates by applying the primary neighborhood to the initial solution. The procedure keeps iterating as long as the incumbent solution improves. (The definition of incumbent is the same as in the OptQuest adaptation and refers to the best solution found from the current starting point.) When no more improvement is possible, the neighborhood is switched to shifts. Again, this neighborhood remains operational until either a new incumbent solution is found or the entire shift neighborhood is explored and no improvement of the incumbent solution is achieved. If a new incumbent solution is found, then the neighborhood search reverts to the primary mechanism (that is, swaps). However, if the exploration of the shift neighborhood ends without improving the incumbent solution then the shake mechanism is invoked. Our "shaking" strategy simply consists of applying random shifts to the incumbent solution. The number of items involved in the random shifts increases from 1 up to 4 as the number of shakings progresses. The entire process ends when there no new incumbents are found after a predetermined number of shakes, which in our case we have set to 40.

Compared to the OptQuest adaptation, the VNS implementation requires more effort. However, it is still based on fairly straightforward mechanisms and the tuning involved choosing the right mix of starting points (alternating between greedy random and totally random), the number of starts (500), the number of shakes (40), and the number of random shifts to be made during the shakes (1 to 4). The merit of the two approaches is analyzed in the following section.

4.5. Computational Experiments

Performances of the multi-start VNS procedure and the Optquest application with the tabu search `Evaluate()` method were compared employing randomly generated instances of various complexities listed in Table 4-3. For the small and medium size instances (total demand of 30 and 40 parts) we have been able to obtain optimal solutions by solving the MILP formulation (augmented with the (9)-(13) cuts) presented in Section 2 with CPLEX. Thus, for these instances, we have been able to verify the heuristic procedures' ability to reach optimality. The average computational time to find these optimal solutions with a CPLEX 11.2.1 solver running on a Windows Enterprise Server with 3.16 GHz dual processors and 32 GB RAM is given in the last column of Table 4-3.

Table 4-3. Characteristics of Problem Instances

Type	Demand	Colors	Geometries	Products	Cplex Time
A	30	3	3	5	3 min
B	40	3	3	8	30 min
C	40	5	5	8	10 hours
D	60	5	5	15	—
E	300	10	10	50	—

We generated ten instances of each type; each instance was solved for two different cost structures – one considering unit cost for each setup, i.e. $h = v = 1$, and another considering vertical and horizontal costs to be sequence dependent and asymmetric while obeying the triangle inequality. Problem instances of types A through D were solved with $PPL = 5$ and $PPL = 10$. The results found by setting $h = v = 1$ are summarized in Table 4-4 (for $PPL = 5$) and Table 4-5 (for $PPL = 10$); results using the asymmetric sequence dependent costs are summarized in Table 4-6 (for $PPL = 5$) and Table 4-7 (for $PPL = 10$). OptQuest was set to perform 1000 iterations while VNS was set to execute 500 starts.

The tables report, for each problem type and for each procedure, 1) the average number of starts to reach the best solution for the Multistart VNS (Starts to Best) or the average number of iterations to the best solution for OptQuest (Iterations to Best), 2) the average time to reach the best solution (Time to Best) and 3) the percentage of optimal solutions found (Optimal Solutions), when those solutions are known. The last column of these tables reports the average percentage gap between the OptQuest/TS solutions and the Multistart VNS solutions (OQ-VNS Gap).

Table 4-4. Results with Uniform Setup Costs, $PPL = 5$

	Multistart VNS			OptQuest/TS			
Problem Type	Starts to Best	Time to Best	Optimal Solutions	Iterations to Best	Time to Best	Optimal Solutions	OQ-VNS Gap
A	1.5	0.43	100%	30.8	0.38	100%	0.00%
B	13.4	8.65	100%	225.5	1.82	90%	2.00%
C	1.3	0.81	100%	25.5	0.30	100%	0.00%
D	33.5	100.00	—	445.4	12.67	—	8.21%

Table 4-5. Results with Uniform Setup Costs, $PPL = 10$

	Multistart VNS			OptQuest/TS			
Problem Type	Starts to Best	Time to Best	Optimal Solutions	Iterations to Best	Time to Best	Optimal Solutions	OQ-VNS Gap
A	1.0	0.19	100%	23.7	0.33	100%	0.00%
B	2.0	1.12	100%	225.9	1.58	100%	0.00%
C	3.7	2.59	100%	68.8	0.67	80%	1.68%
D	83.4	242.83	—	272.7	7.17	—	8.72%

Table 4-6. Results with Non-Uniform, Asymmetric Setup Costs, $PPL = 5$

	Multistart VNS			OptQuest/TS			
Problem Type	Starts to Best	Time to Best	Optimal Solutions	Iterations to Best	Time to Best	Optimal Solutions	OQ-VNS Gap
A	2.2	0.79	100%	29.5	0.27	90%	11.42%
B	12.6	9.83	100%	225.5	5.41	80%	1.57%
C	10.7	9.96	100%	175.5	3.87	90%	0.15%
D	246.2	957.78	—	563.7	94.08	—	11.40%

Table 4-7. Results with Non-Uniform, Asymmetric Setup Costs, $PPL = 10$

	Multistart VNS			OptQuest/TS			
Problem Type	Starts to Best	Time to Best	Optimal Solutions	Iterations to Best	Time to Best	Optimal Solutions	OQ-VNS Gap
A	3.4	0.81	100%	65.4	0.42	90%	10.75%
B	13.9	8.19	100%	208.2	4.23	50%	4.79%
C	21.9	18.13	100%	298.5	5.56	70%	1.50%
D	220.2	783.87	—	465.7	71.17	—	14.02%

Tables 4 through 7 show that the Multistart VNS is able to find all known optima. OptQuest fails to match the optimal solution in a few occasions, and it tends to perform worse when setup costs are non-uniform and sequence dependent. The specialized VNS procedure finds better solutions than the OptQuest adaptation for most instances in the D set, and all instances in the E set. Our conjecture is that the flat nature of the landscape is the primary reason behind the superior performance of VNS; in fact, VNS performed better than other metaheuristics that we have attempted on this problem. Navigating different neighborhoods, which is the essence of VNS, helps explore diverse areas of the flat landscape. For the same reason, i.e. flatness of the landscape, the multiple-start nature of the procedure makes it more effective. The size and complexity of the problem instance are important factors to be considered when choosing the number of starts to be used. The **Starts/Iterations to Best** columns in the tables above refer to the earliest start or iteration where the best solution was found. For smaller and simpler instances (types A, B, and C), the best solutions (all optimal for the Multistart VNS) were found within 50 iterations, while for the larger type-D instances, better solutions were found even after 300 starts/iterations.

In terms of computational times, both procedures performed at a similar level for problem types A, B and C. Computational times for the VNS in problem instances of types D and E are significantly larger than the OptQuest times. This is due to best solutions found late in the search and to the computational demands of a VNS start compared to an OptQuest iteration. This may seem like an unfair comparison but we have verified that additional OptQuest iterations do not improved the results reported in Tables 4 to 7.

For the large E-type problems, ten instances were created —each instance was solved for uniform setup costs, as well as for non-uniform asymmetric setup costs. For these instances, PPL was set to 50; the results are summarized in Table 4-8. OptQuest was set to perform 100 iterations while VNS was set to execute 50 starts.

Table 4-8. Results for Type E (Large) instances, $PPL = 50$

Setup Cost Structure	Multistart VNS			OptQuest/TS			
	Starts to Best	Time to Best	Optimal Solutions	Iterations to Best	Time to Best	Optimal Solutions	OQ-VNS Gap
Uniform	21.1	9466.17	--	40.3	873.88	--	7.00%
Non-Uniform	19.8	10194.80	--	46.7	7162.69	--	15.33%

Table 4-8 shows the ability of the Multistart VNS to find better solutions than OptQuest/TS. Consistent with the results obtained for the smaller problem instances, the gap tends to be wider for a non-uniform cost structure. For these problem instances, VNS required more computational time than OptQuest but both procedures are capable of delivering solutions within the timeframe needed in the application (which does not require an online real-time response).

4.6. Conclusions

We have described a sequencing problem with two types of setups. A MILP formulation of the problem is presented and used to find optimal solutions to a set of problem instances. We also describe the adaptation of OptQuest and the development of a Multistart VNS as two alternative heuristic approaches to the problem. While the MILP formulation works well for very small size instances, heuristic approaches are essential to obtain reasonably good solutions for industrial scale problem instances. The OptQuest adaptation required less effort to develop compared to the Multistart VNS, which is tailored to the problem of interest. We show that the problem inherently has a very flat landscape, which motivated the choices we made to create a broader exploration of diverse regions of the landscape; navigating multiple neighborhoods through VNS, and using multiple starts proved to be effective strategies.

In our implementation, both the OptQuest adaptation and the Multistart VNS performed reasonably well for small size instances. However, for larger scale problems, the Multistart VNS performed better, indicating that for this type of problem, it might be worthwhile to invest in a heuristic custom-made for the problem. Experiments show that for large problems, savings of at least 5% are very likely when employing the specialized VNS procedure instead of the OptQuest adaptation. The benefits

of using the Multistart VNS are expected to be even larger when setup costs are non-uniform. These predicted savings must be thought of as the benefits that help amortize the difference in development cost between the specialized method and the adaptation of the commercial software.

An interesting extension of this work would be to adapt the solution framework to problems that involve three or more types of sequence-dependent setups.

Chapter 5 - CONCLUSION AND FURTHER WORK

In the prior chapters, I have presented three papers that address real life scheduling problems spanning across two different industries – health care and manufacturing. Each of the three papers helps balance supply or capacity with demand, something that is fundamental to improving operational efficiency. Chapter 2 addresses a problem of personnel capacity planning to serve stochastic demand of patient care at medical EDs; Chapter 3 helps managing the demand side at outpatient clinics by presenting optimal policies to handle wait-preempt decisions when unpunctual patients arrive out of turn; and Chapter 4 adds to the literature of manufacturing job sequencing that involves sequence dependent setup costs. The three essays employ a variety of methodologies ranging from mathematical programming to analytical methods and metaheuristics; they provide useful solution methods and interesting managerial insights for three real-life problems, thereby making worthwhile contributions to operations management in general, and to health care and manufacturing scheduling literatures in particular. Contribution of these papers and possible interesting extensions of these works are discussed in the following paragraphs.

The ED staff planning paper helps make effective and efficient provider schedule in a medical ED. We have used a mixed integer programming model to arrive at the optimal staffing schedule that minimizes staffing costs while fulfilling a certain service level. The model accommodates patients of different acuity levels ranging from trivial to life threatning; it also accommodates providers of different skill levels, ranging from physicians' assistants qualified to treat only low acuity patients, to MDs in emergency medicine qualified to treat even the most acute patients. Furthermore, in addition to solo providers treating individual patients, the model allows providers to work in *ad hoc* teams, where less qualified providers, when supervised by more skilled providers, can attend to patients of higher acuity. Using the ED staff planning model, we evaluate the impact of different parameters such as: availability of different classes of providers; target service levels; use of provider teams; and variability of service times. For example, we show that using provider teams improves provider utilization and reduces staffing costs, even though it increases overall person hours scheduled. We also see that the service levels attained are

quite robust to variability in service times; however, higher service times necessitate that the care of a few patients be divided among multiple providers. Thus, our research provides interesting insights regarding personnel capacity planning for ED administrators and opens up avenues of further research as mentioned below.

While the current ED staff planning paper (Chapter 2) provides a useful model to schedule the right mix of providers in an ED, it considers only regularly scheduled providers and excludes providers who may be on-call. In practice, there is an interest among ED administrators to use on-call providers, who can be summoned to handle unexpectedly high patient demand. Given availability of on-call providers, it will be interesting to find the optimal mix of regularly scheduled and on-call providers, and deciding on the threshold demand scenario that would trigger summoning an on-call provider. One approach would be to model it as a two-stage stochastic program, where the first stage decision involves coming up with the regular staffing schedule, and the second stage involves summoning the on call provider(s) upon realization of a demand scenario. An alternative approach would be to formulate it as a newsvendor problem with expensive reactive capacity that can be invoked based on better, but not perfect, forecast. Our conjecture is that the value of engaging on-call providers will be increasing in autocorrelation of demand and decreasing in the lead time involved in summoning an on-call provider.

In Chapter 3, we deal with demand side management at outpatient clinics through real time updates to patient appointments to accommodate situations when the provider is idle and patients arrive out of turn. It is well known that patient no-shows and patient unpunctuality are problems that affect the efficiency of clinics operating on an appointment basis. While several papers have addressed the issue of patient no-shows, the problems arising out of patient unpunctuality has not received enough attention in the extant literature. In Chapter 3, we formally define a provider's wait-preempt dilemma, the situation that an idle provider faces when the patient scheduled to be seen next has not shown up yet, but a patient scheduled later in the day has already arrived. In this predicament, the provider essentially has two choices: either keep waiting for the patient scheduled next, or preempt the appointment scheduled next and start seeing the patient who has already arrived. Taking into account the cost of clinic overtime and

the cost of patient waiting times, we show that for reasonable assumptions regarding the arrival pattern of patients, it is possible to analytically determine that in order to maximize the expected utility of the clinic, how long should the provider remain idle while waiting for the next scheduled patient, when a patient scheduled later has already arrived. The results provide insights into real-time schedule management at a clinic; for example, by identifying properties of the cost function, we prove that a first-come-first-served policy is never optimal, and that depending on clinic specific parameters, clinic managers can arrive at optimal preemption policies. We solve the wait-preempt dilemma optimally for the two-patient case, and we propose a heuristic for cases when the dilemma arises between patients with non-adjacent appointments. We also provide a software program that clinics can readily use to arrive at optimal policies of dealing with the dilemma. Our work is not only applicable to outpatient health clinics, but also to any other appointment-based activity, such as lawyer offices. The research can be extended in ways outlined below.

In our current research, we have assumed the lateness distribution to be same for all patients. As an extension, it will be interesting to consider different lateness distributions for individual patients, and evaluate its impact on the optimal sequence of patient appointments and the policy of preemption. Considering more realistic patient lateness distribution (rather than a triangular approximation that we have used) and service time variability are other avenues of further research that can be pursued. Last but not the least, recommended optimal wait-preempt policies could be implemented at real clinics, data gathered before and after the implementation and compared to measure improvements; such empirical validation will be immensely valuable to our research.

Chapter 4 discusses the demand side management of a closed-loop manufacturing problem: sequencing jobs for a given production order where there are two types of sequence dependent setup costs. There is a "horizontal" cost that depends on adjacent jobs, and there is a "vertical" cost that depends on the jobs that are separated by the length of the production loop. This particular problem is different from other sequencing problems with setups that have been addressed before; we introduce the problem and present alternative solution methods. The first approach is a mixed integer program that

finds the optimal solution within a reasonable time for small problem instances, and its results serve as a benchmark for the alternative heuristic procedures. Among the two heuristic approaches, one employs a commercial optimization software built on metaheuristic technology, and hence required less development effort. The other heuristic method was built from scratch, was tailored to the problem at hand, took advantage of the structural properties of the problem, and generally performed better. Apart from providing good solution methods for a new scheduling problem, this paper provides a nice example of how the benefits and limitations of using commercial software compare with custom solution developed for a particular problem. This work can be extended in the following ways.

The current paper (Chapter 4) deals with two types of sequence-dependent setups. An interesting extension of this work would be to adapt the solution framework to problems that involve three or more types of sequence-dependent setups. Another direction would be to use the unified modeling framework proposed by Kochenberger *et. al.* (2004), where the problem is modeled as an unconstrained quadtratic program (UQP) such that all the constraints in the original linear integer program are eliminated by recasting them as penalizing quadratic terms in the objective. The UQP can then be solved using appropriate, possibly heuristic, methods.

In this dissertation, I have addressed specific problems pertaining to both sides of the capacity-demand balance. In ED staff planning (Chapter 2), we dealt with the supply side, where we investigated the issue of personnel capacity planning at a medical emergency department that faces stochastic demand. The remaining papers dealt with managing the demand side; in Chapter 3, we dealt with patient appointments in an out-patient clinic to balance clinic overtime costs and patient dissatisfaction caused by long wait times; and in Chapter 4, we tried to arrive at the best sequence of manufacturing jobs that has two different types of sequence dependent setup costs. I hope that my research makes a humble contribution to the body of knowledge of operations management that deals with balancing supply and demand, through planning of personnel capacity and managing schedules of customers and jobs.

References

Allahverdi, A., J. N. D. Gupta, T. Aldowaisan (1999) "A Review of Scheduling Research involving Setup Considerations," *OMEGA The International Journal of Management Sciences, 27*(2), 219-239.

Allahverdi, A., C. T. Ng, T. C. E. Cheng, and M. Y. Kovalyov (2008) "A Survey of Scheduling Problems with Setup Times or Costs," *European Journal of Operational Research, 187*(3), pp. 985-1032.

Bard J.F. (2004), "Staff scheduling in high volume service facilities with downgrading," *IIE Transactions, 36*(10), 985-997.

Beaulieu H, J.A. Ferland, B. Gendron, and P. Michelon (2000) "A mathematical programming approach for scheduling physicians in the emergency room," *Health Care Management Science, 3*, 193-200.

Ben-Gal I., M. Wangenheim, and A. Shtub (2010) "A new standardization model for physician staffing at hospitals," *International Journal of Productivity and Performance Management, 59*(8), 769-791

Birge, J.R. and F. Louveaux. (2010) *Introduction to Stochastic Programming (second edition)*. New York: Springer.

Brunner J.O, J.F. Bard, and R. Kolisch (2009) "Flexible shift scheduling of physicians," *Health Care Management Science, 12*(3), 285-305.

Brunner J.O, J.F. Bard, and R. Kolisch (2010) "Midterm scheduling of physicians with flexible shifts using branch and price," *IIE Transactions, 43*(2), 84-109.

Brunner J.O, and G.M. Edenharter (2011) "Long term staff scheduling of physicians with different experience levels in hospitals using column generation," *Health Care Management Science, 14*(2), 189-202.

Brusco M.J., and Johns T.R. (1998) "Staffing a multiskilled workforce with varying levels of productivity: an analysis of cross-training policies," *Decision Sciences, 29*(2), 499-515.

Campbell G.M. (2011) "A two-stage stochastic program for scheduling and allocating cross-trained workers," *Journal of the Operational Research Society, 62*, 1038-1047.

Campbell G.M. (1999) "Cross-utilization of workers whose capabilities differ," *Management Science, 45*(5), 722-732.

Carter M.W, and S.D. Lapierre (2001) "Scheduling emergency room physicians," *Health Care Management Science, 4*, 347-360.

Cayirli, T., E. Veral (2003) "Outpatient Scheduling in Health Care: A Review of Literature," *Production and Operations Management*. *12*(4) 519-549.

Cayirli, T., E. Veral, H. Rosen (2006) "Designing appointment scheduling systems for ambulatory care services," *Health Care Management Science*. *9*(1) 47-58.

Cayirli, T., K.K. Yang, S.A. Quek (2012) "A Universal Appointment Rule in the Presence of No-Shows and Walk-Ins," *Production and Operations Management*. *21*(4) 681-697.

Cezik M.T. and P. L'Ecuyer (2008) "Staffing multiskill call centers via linear programming and Simulation," *Management Science*, *54*(2), 310-323.

Charnes A. and W.W.Cooper (1959) "Chance-constrained programming," *Management Science*, *6*(1), 73-79.

Cheang B, H. Li, A. Lim, and B. Rodrigues (2003) "Nurse rostering problems – a bibliographic survey," *European Journal of Operational Research, 151*(3), 447-460.

Cheng, T. C. E., J. N. D. Gupta and G. Wang (2000) "A Review of Flowshop Scheduling Research with Setup Times," *Production and Operations Management, 9*(3), 262-282.

Easton F.F (2011), Cross-training performance in flexible labor scheduling environments, *IIE Transactions, 43*(8), 589-603.

Ernst A.T, H. Jiang, M. Krishnamoorthy, and D. Sier (2004) "Staff scheduling and rostering: a review of applications, methods and models," *European Journal of Operational Research 153*(1), 3-27.

Ferrand Y., M. Magazine, U.S. Rao, and T.F. Glass (2011) "Building cyclic schedules for emergency department physicians," *Interfaces, 41*(6), 521-533.

Ganguly S, S.R. Lawrence, and M. Prather (2013) "Emergency Department Staff Scheduling to Improve Patient Care and Reduce Costs, " *Working Paper*

Ganguly S. and M. Laguna (2013) "Modeling and solving a cyclic and batching scheduling problem with two types of setups," *Working Paper*

Ganguly S. and M. Samorani (2013) "Provider's Wait-Preempt Dilemma," *Working Paper*

Garcia-Sabater, J. P., C. Andres, C. Miralles, and J. J. Garcia-Sabater (2008) "An Application Oriented Approach for Scheduling a Production Line with Two Dimensional Setups, in *Proceedings of the 11th International Workshop on Project Management and Scheduling*, F. Şerifoğlu and Ü. Bilge (eds.), Istanbul, Turkey, 90-93.

Garmel, G.M. (2005) "Approach to the Emergency Patient," in *Principles of Emergency Medicine*, S.V. Mahadevan and G.M Garmel (editors), Cambridge University Press, Cambridge UK, 3-18.

Ginde A.A, J.A. Espinola, A.F. Sullivan, F.C. Blum, and C.A. Camargo (2010) "Use of midlevel providers in US EDs, 1993 to 2005: implications for the workforce," *American Journal of Emergency Medicine, 28*(1), 90-94.

Graff, L.G. and M.J. Radford (1990), Formula for emergency physician staffing, *American Journal of Emergency Medicine, 8*(3), 194-199.

Green, L.V., J. Soares, J.F. Giglio, and R.A. Green (2006) "Using queueing theory to increase the effectiveness of emergency department provider staffing," *Academic Emergency Medicine 13*(1), 61-68.

Green, L.V., P.J. Kelesar, and W. Whitt (2007) "Coping with time-varying demand when setting staffing requirements for a service system," *Production and Operations Management 16*(2), 13-39.

Gupta, D., B. Denton (2008) "Appointment Scheduling in Health Care: Challenges and Opportunities," *IIE Transactions 40*, 800-819

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten (2009) *The WEKA Data Mining Software: An Update*; SIGKDD Explorations, Volume 11, Issue 1.

Henrion, Rene (2004) "Introduction to Chance-Constrained Programming." Accessed on April 15, 2013. http://stoprog.org/index.html?SPIntro/intro2ccp.html.

Hillier, F.S. and G.J. Lieberman (2001) *Introduction to Operations Research.* Boston, MA: McGraw Hill.

Huang, Y. and P. Zuniga (2012) "Dynamic overbooking scheduling system to improve patient access," *Journal of the Operational Research Society 63,* 810–820.

Jordan, W.C. and S.C. Graves (1995) "Principles on the benefits of manufacturing process flexibility," *Management Science, 41*(4), 577-594.

Julien, F. M., M. J. Magazine, and N. G. Hall (1997) "Generalized preemption models for single-machine dynamic scheduling problems," *IIE Transactions*, *29*, 359-372

Jungwattanakit, J., M. Reodecha, P. Chaovalitwongse and F. Werner (2009) "A Comparison of Scheduling Algorithms for Flexible Flow Shop Problems with Unrelated Parallel Machines, Setup Times and Dual Criteria," *Computers and Operations Research, 36*(2), 358-378.

Jun, J.B, S.H. Jacobson, and J.R. Swisher (1999) "Application of discrete-event simulation in health care clinics: a survey," *Journal of the Operational Research Society, 50*(2), 109-123.

Kachalia, A, T.K. Gandhi, A.L. Puopolo, C. Yoon, E.J. Thomas, R. Griffey, T.A. Brennan, and D.M. Studdert (2007) "Missed and delayed diagnoses in the emergency department: a study of closed malpractice claims from 4 liability insurers," *Annals of Emergency Medicine*, *49*(2), 196-205.

Klassen, K. J., T. R. Rohleder (1996) "Scheduling outpatient appointments in a dynamic environment," *Journal of Operations Management, 14*(2), 83–101.

Klassen, K, R. Yoogalingam (2009) "Improving Performance in Outpatient Appointment Services with a Simulation Optimization Approach," *Production and Operations Management*, *18*(4), 447-458.

Kochenberger, G.A., F. Glover, B. Alidaee, and C. Rego (2004) "A unified modeling and solution framework for combinatorial optimization problems," *OR Spectrum, 26*(2), 237-250

LaGanga, L.R., S.R. Lawrence (2007) "Clinic overbooking to improve patient access and increase provider productivity," *Decision Sciences, 38*(2), 251-276.

LaGanga, L.R., S.R. Lawrence (2012) "Appointment Overbooking in Health Care Clinics to Improve Patient Service and Clinic Performance," *Production and Operations Management. 21*(5) 874–888.

Laguna, M. (1999) "A Heuristic for Production Scheduling and Inventory Control in the Presence of Sequence-dependent Setup Times," *IIE Transactions, 31*(2), 125-134.

Laguna, M. (2011) "OptQuest: Optimization of Complex Systems," White Paper, OptTek Systems Inc., http://www.opttek.com/white-papers.

Lipscomb J. (1991), *Physician Staffing for the VA: Volume I. Committee to Develop Methods Useful to the Department of Veterans Affairs in Estimating its Physician Requirements*, *Institute of Medicine*, National Academy Press, Washington, DC.

Liu, N., S. Ziya, and V.G. Kulkarni (2010) "Dynamic scheduling of outpatient appointments under patient no-Shows and cancellations," *Manufacturing & Service Operations Management 12*(2) 347–364.

Mula J, R. Poler, J.P. García-Sabater, and F.C. Lario (2006) "Models for production planning under uncertainty: a review," *International Journal of Production Economics, 103*(1), 271-285

Oddone E., S. Guarisco, and D. Simel (1993) "Comparison of housestaff's estimates of their workday activities with results of a random work-sampling study," *Academic Medicine: Journal of the Association of American Medical Colleges*, *68*(11), 859-861

Priore, P., D. de la Fuente, A. Gomez, and J. Puente (2001) "A review of machine learning in dynamic scheduling of flexible manufacturing systems," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, *15*, 251-263.

Rajendran, C. and H. Ziegler (2003) "Scheduling to Minimize the sum of Weighted Flowtime and Weighted Tardiness of Jobs in a Flowshop with Sequence-dependent Setup Times," *European Journal of Operational Research, 149*(3), 513-522.

Robbins T. (2010) "Tour scheduling and rostering," In J.J. Cochran (Ed.), *Wiley Encyclopedia of Operations Research and Management Science.*

Robinson, L., R. Chen (2010) "A Comparison of Traditional and Open-Access Policies for Appointment Scheduling," *Manufacturing & Service Operations Management*, ***12***(2), 330-346.

Rotstein Z, R. Wilf-Miron, B. Lavi, A. Shahar, U. Gabbay, and S. Noy (1997) "The dynamics of patient visits to a public hospital ED: a statistical model," *American Journal of Emergency Medicine*, *15*(6), 596-599.

Ruiz, R., C. Maroto, and J. Alcaraz (2005) "Solving the Flowshop Scheduling Problem with Sequence Dependent Setup Times using Advanced Metaheuristics," *European Journal of Operational Research, 165*(1), 34-54.

Scherer, W.T., T.A. Pomroy, D.N. Fuller (2003) "The triangular density to approximate the normal density: decision rules-of-thumb," *Reliability Engineering & System Safety*, *82*(3), 331-341.

Sox C.M., H.R. Burstin, E.J. Orav, A. Conn, G. Setnik, D.W. Rucker, P. Dasse, and T.A. Brennan (1998) "The effect of supervision of residents on quality of care in five university-affiliated emergency departments," *Journal of the Association of American Medical Colleges*, *73*, 776-782

Swartzman, G (1970) "The Patient Arrival Process in Hospitals: Statistical Analysis," *Health Serv Res*, *5*(4), 320-329.

Szwarc, W. and J. N. D. Gupta (1987) "A Flow-Shop Problem with Sequence-Dependent Additive Setup Times," *Naval Research Logistics, 34*(5), 619-627.

Talbi, G. (2009) *Metaheuristics – From Design to Implementation*, John Wiley & Sons: Hoboken, 150-154.

Ugray, Z., L. Lasdon, J. Plummer, F. Glover, J. Kelly and R. Martí (2007) "Scatter Search and Local NLP Solvers: A Multistart Framework for Global Optimization," *INFORMS Journal on Computing*, *19*(3), 328–340.

White, M.J.B. and M.C. Pike (1964) "Appointment Systems in Out-Patients' Clinics and the Effect of Patients' Unpunctuality," *Medical Care*, *2*(3), 133-141, 144-145

Witten, I. H. and E. Frank (2005) *Data mining : practical machine learning tools and techniques, Morgan Kaufmann series in data management systems*. Amsterdam; Boston, MA: Morgan Kaufman.

Zeng, B., A. Turkcan, J. Lin, and M. Lawley (2010) "Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities," *Annals of Operations Research 178*(1) 121–144.

Zhu, X. and W. E. Wilhelm (2006) "Scheduling and Lot Sizing with Sequence-Dependent Setup: A Literature Review," *IIE Transactions, 38*(11), 987-1007.

Appendix 2.1 – The Staff Assignment Algorithm

Given an ED staff schedule determined by our ED planning model, this follow-on staff assignment algorithm determines an allocation of providers that meets realized demand for each hour of the planning horizon. Demand arriving during each hour becomes apparent chronologically, and any unmet demand during an hour is added to the demand of the following hour. When capacity is insufficient to meet demand, priority is accorded to demand of the highest acuities.

The algorithm invokes different integer programs in succession to check if demand can be fulfilled without deploying team providers or without having multiple providers treat any patient. First, it checks the most restrictive problem: fulfill demand by assigning only one individual provider to each patient (individual provider, indivisible patient). If there is a feasible solution to this problem, the other combinations are not tested. Otherwise, it considers two options that are less restrictive. First it tries to fulfill demand by assigning only one team or one individual provider to each patient (team provider, indivisible patient). Then, it tries to fulfill demand by assigning one or more individual providers to each patient (individual provider, divisible patients).

Finally, if all the above attempts fail to obtain any feasible solution that fulfills demand, the algorithm invokes a linear program that minimizes weighted sum of unmet demand by assigning one or more individual or team providers to each patient (team provider, divisible patient). Following is a pseudo code representation of the algorithm.

Input Files:

1. Staffing for each hour (by training-level)
2. Demand for each hourly period (by acuity)

For each hourly period

```
{
    Obtain staffing levels during current period;
        Obtain demand-arrival-in-current-period by acuity;
    Compute total-demand-on-hand
                        = demand-arrival-in-current-period + unmet-demand-from-prior-period;
    Check if total-demand can be met without employing teams and without dividing patients;
        IF no feasible solution exists
        {
            Check if total-demand can be met by employing teams (but indivisible patients);
            Check if total-demand can be met with divisible patients (but no teams);
            IF no feasible solution exists for either
            {
                Minimize weighted unmet-demand (allow teams & patient division);
            }
        }
}
```

Note: *Underlined and italicized lines involve a call to a commercial solver* (CPLEX)

Appendix 3.1 – Expressions of Conditional Show Probabilities

Recall that the arrival time t follows a triangular distribution $T_r(a, b, c)$. For any time t following the current time t_0, let $p(t_0, t)$ be the probability density of patient T arriving at t given that she has not arrived by t_0. Let q be the unconditional show probability, $p(t)$ the unconditional probability density of T arriving at t, and $P(N_{t_0})$ the unconditional probability of patient T not showing up by t_0.

$$p(t_0, t) = p(t)/P(N_{t_0})$$

$$P(N_{t_0}) = \begin{cases} 1, for\ t_0 \le a \\ 1 - (q(t_o - a)^2/(c-a)(b-a)), for\ a \le t_0 < b \\ 1 - q + (q(c - t_o)^2/(c-a)(c-b)), for\ b \le t_0 \le c \\ 1 - q, for\ t_0 > c \end{cases}$$

$$p(t) = \begin{cases} \dfrac{2q(t-a)}{(c-a)(b-a)}, for\ a \le t < b \\ \dfrac{2q(c-t)}{(c-a)(c-b)}, for\ b \le t \le c \\ 0, for\ t \le a \text{ or } t \ge c \end{cases}$$

Depending on the values of t_0 and t, we obtain the following expressions for $p(t_0, t)$:

If $t \le a$: $p(t_0, t) = 0$

If $a < t < b$ and $t_0 \le a$: $p(t_0, t) = \frac{2q(t-a)}{(c-a)(b-a)}$

If $a < t < b$ and $a < t_0 < b$: $p(t_0, t) = \frac{2q(t-a)}{[(c-a)(b-a) - q(t_0-a)^2]}$

If $b < t < c$ and $t_0 \le a$: $p(t_0, t) = \frac{2q(c-t)}{(c-a)(c-b)}$

If $b < t < c$ and $a < t_0 < b$: $p(t_0, t) = \frac{2q(c-t)(b-a)}{(c-b).[(c-a)(b-a) - q(t_0-a)^2]}$

If $b < t < c$ and $b < t_0 < c$: $p(t_0, t) = \frac{2q(c-t)}{[(1-q)(c-a)(c-b) + q(c-t_0)^2]}$

If $t \ge c$: $p(t_0, t) = 0$

Appendix 3.2 – Proofs

THEOREM S2-1. *For any* $\lambda > 0$, $\frac{\partial E_T(t_0+\lambda,x)}{\partial x} = k_1 \cdot \frac{\partial E_T(t_0,x)}{\partial x}$, $\frac{\partial E_W(t_0+\lambda,x)}{\partial x} = k_2 \cdot \frac{\partial E_W(t_0,x)}{\partial x}$, *and* $\frac{\partial E_O(t_0+\lambda,x)}{\partial x} = k_3 \cdot \frac{\partial E_O(t_0,x)}{\partial x}$, *where* k_1, k_2, *and* k_3 *are positive quantities independent of* x.

Proof: Appendix 3.4 reports the derivatives of all components of the cost function. Let us consider $\frac{\partial E_T(t_0,x)}{\partial x}$ and restrict our proof to the case $x \geq 0$; the other cases for E_T and all cases for E_W and E_O can be derived similarly.

$$\frac{\partial E_T(t_0,x)}{\partial x} = -p(t_0,x)d + \int_x^{x+d} p(t_0,t)dt$$

First, let us assume that λ is small enough that the formula used for $p(t_0,x)$ is the same as that for $p(t_0+\lambda,x)$. Since for any formula of $p(t_0,t)$ t_0 appears only in the denominator and t only the numerator (Appendix 3.1), we can write $p(t,t_0) = \frac{f(t)}{g(t_0)}$.

$$\frac{\partial E_T(t_0,x)}{\partial x} = -d\frac{f(t)}{g(t_0)} + \int_x^{x+d} \frac{f(t)}{g(t_0)}dt = \frac{-f(t)d + \int_x^{x+d} f(t)dt}{g(t_0)}$$

$$\frac{\partial E_T(t_0+\lambda,x)}{\partial x} = \frac{-f(t)d + \int_x^{x+d} f(t)dt}{g(t_0+\lambda)}$$

Therefore,

$\frac{\partial ET(t_0+\lambda,x)}{\partial x} = k_1 \cdot \frac{\partial ET(t_0,x)}{\partial x}$, where $k_1 = \frac{g(t_0)}{g(t_0+\lambda)}$ (S2-1)

If λ is not small enough, then let $t_0+\varepsilon$ be the point where the formula changes from $p(t,t_0) = \frac{f(t)}{g(t_0)}$ to $p(t,t_0) = \frac{\hat{f}(t)}{\hat{g}(t_0)}$ (our proof can be easily generalized to the case where λ is so large that the formula changes more than once). By using (S2-1) twice, we obtain $\frac{\partial ET(t_0+\lambda,x)}{\partial x} = \frac{\hat{g}(t_0)}{\hat{g}(t_0+\lambda)} \cdot \frac{g(t_0)}{g(t_0+\lambda)} \cdot \frac{\partial ET(t_0,x)}{\partial x}$. ■

THEOREM S2-2. *The interval* $[-d, 0)$ *always contains a wait interval.*

Proof: Since E_D and E_W are 0 in $[-d, 0)$, we only need to prove that E_T is decreasing at $x = 0$. The expressions for ${E_T}'(x \leq 0)$ can be taken from Appendix 3.3 (case $x + d \geq 0$ and $x \leq 0$).

$$ {E_T}'(x) = -p(t_0, x)(x + d) + \int_x^{\min(x+d,c)} p(t_0, t)dt $$

From Theorem S2-1, we can assume $t_0 < a$. Thus, we use $p(t)$, the unconditional expression for $p(t_0, t)$.

$$ E_T'(0^-) = -p(0^-)d + \int_0^{\min(d,c)} p(t)dt = -p(0^-)d + \int_0^{\min(d,c)} p(t)dt $$

$$ = -\frac{2qcd}{(c-a)(c-b)} + \frac{2q}{(c-a)(c-b)} \int_0^{\min(d,c)} (c - t)dt $$

$$ = \frac{2q}{(c-a)(c-b)} \left[-cd + \int_0^{\min(d,c)} (c - t)dt \right] $$

If $c < d$

$$ E_T'(0^-) = \frac{2q}{(c-a)(c-b)} \left[-cd + c^2 - \frac{1}{2}c^2 \right] < 0 $$

If $c \geq d$

$$ E_T'(0^-) = \frac{2q}{(c-a)(c-b)} \left[-cd + cd - \frac{1}{2}d^2 \right] < 0 \blacksquare $$

Appendix 3.3 – Theoretical Support for the Analytical Method

Here we prove properties of the cost function, which allow us to develop the analytical method.

Note that the cost function E_{TC} is continuous and piece-wise differentiable on $[k_1, k_2]$ where $k_1 = -d$ and $k_2 = min(c, d)$. Given two points in time u and t such that $u \geq t$, if $E_{TC}(u) < E_{TC}(t)$, then it is better to wait up to time u rather than preempting at time t. Let us define the set of "wait-up-to deadlines" $w(t)$ of time t as the set of times until which it is optimal to wait if this decision is made at time t:

$$w(t) = \{u \in [t, k_2] \; such \; that \; \forall v \in [t, k_2]: E_{TC}(u) \leq E_{TC}(v)\}$$

Note that if $w(t) = \{t\}$, the best decision is to preempt. Without loss of generality, we can further assume that the derivatives at the extreme points are non-zero, i.e., ${E_{TC}}'(k_2^-) \neq 0$ and ${E_{TC}}'(k_1^+) \neq 0$. If the derivative is zero at one extreme, we can shorten the interval $[k_1, k_2]$ so that the derivative at the new extreme is not zero.

LEMMA S3-1. *If* $w(t) = \{u\}$, *with* $t < u < k_2$, *then* $\exists \varepsilon > 0 \; such \; that \; \forall z \in (u, u + \varepsilon]: w(z) = z$.

Interpretation: if it is optimal to wait until u, then right after u it is optimal to preempt.

Proof: Since $w(t) = \{u\}$, $\forall p > u, E_{TC}(p) > E_{TC}(u)$. Let $p^* > u$ be the point where E_{TC} achieves the minimum value after u. Then, $E_{TC}(u) < E_{TC}(p^*)$. Since E_{TC} is continuous, $\exists \varepsilon > 0 \; such \; that \; {E_{TC}}' \geq 0 \; in \; (u, u + \varepsilon)$. For ε small enough, $EC(u) < EC(\varepsilon) < EC(p^*)$. Therefore, $\forall z \in (u, u + \varepsilon]: w(z) = z$.

Let us introduce the two important concepts of Wait-Preempt (WP) point and Preempt-Wait (PW) points. A WP point is such that it is optimal to wait before it and preempt after it; a PW point is such that it is optimal to preempt before it and wait after it.

Definition: WP point

A point t is a WP point if $\exists \; \varepsilon_1, \varepsilon_2 > 0: \forall u \in (t - \varepsilon_1, t], w(u) = \{t\}$ and $\forall u \in [t, t + \varepsilon_2], w(u) = \{u\}$.

The *left range* of a WP point t is defined as $(t - \varepsilon_1^*, t]$, where ε_1^* is the maximum value of ε_1 for which

$\forall u \in (t-\varepsilon_1, t], w(u) = \{t\}$. The *right range* of a WP point t is defined as $[t, t+\varepsilon_2^*]$, where ε_2^* is the maximum value of ε_2 for which $\forall u \in [t, t+\varepsilon_2]\ w(u) = \{u\}$.

Definition: PW point

A point t is a PW point if $\exists\ \varepsilon_1, \varepsilon_2 > 0: \forall u \in [t-\varepsilon_1, t), w(u) = \{u\},\ w(t) = \{t, t+\varepsilon_2\}$, and $\forall u \in (t, t+\varepsilon_2]: w(u)=\{t+\varepsilon_2\}$. The left range of a PW point t is defined as $[t-\varepsilon_1^*, t)$, where ε_1^* is the maximum value of ε_1 for which $\forall u \in [t-\varepsilon_1, t), w(u) = \{u\}$. The right range of a PW point t is defined as $(t, t+\varepsilon_2^*]$, where ε_2^* is the maximum value of ε_2 for which $\forall u \in (t, t+\varepsilon_2], w(u)=\{t+\varepsilon_2\}$. Clearly, if t is a PW point, the optimal decision in its right range is to wait up to $t+\varepsilon_2^*$.

Now, we turn our attention to defining characteristics of WP and PW points.

PROPOSITION S3-1. *A point t is a WP point if and only if it is a local minimum and $w(t) = \{t\}$. In practice, a WP point is a local minimum at which the best decision is to preempt.*

Proof:

PART 1:

If t is a WP point, then by definition, $\exists \varepsilon_2 > 0$ such that $\forall u \in [t, t+\varepsilon_2], w(u) = \{u\}$. Therefore, $w(t) = \{t\}$. By taking $\varepsilon < min(\varepsilon_1, \varepsilon_2)$, it is easy to show that $E_{TC}(t) < E_{TC}(v)$ when $|t-v| \leq \varepsilon$. Therefore, t is a local minimum.

PART 2:

Assume $w(t) = \{t\}$ and t is a local minimum. Because E_{TC} is piecewise differentiable, $\exists \gamma > 0$ such that $E_{TC}'(u) \leq 0\ \forall u \in [t-\gamma, t]$. So, $\forall u \in [t-\gamma, t], w(u)$ contains t or, alternatively, another point $v > t$ such that $E_{TC}(v) < E_{TC}(t)$. But since $w(t) = \{t\}$, the latter scenario is impossible, and it must be $w(u) = \{t\}$. By setting $\varepsilon_1 = \gamma$, the first condition for a WP point is satisfied. Analogously, $\exists \gamma > 0$ such that $E_{TC}'(u) \geq 0\ \forall u \in (t, t+\gamma]$. Furthermore, if γ is small enough, it is also true that $\forall z >$

$t, E_{TC}(t+\gamma) < E_{TC}(z)$, because for γ that tends to 0, $E_{TC}(t+\gamma)$ tends to $E_{TC}(t) < E_{TC}(z)$. Thus, $\varepsilon_2 = \gamma$ satisfies the second condition for a WP point. ■

PROPOSITION S3-2. *A point t is a PW point if and only if $E_{TC}'(t) > 0$ and $w(t) = \{t, v\}$, where $v > t$. In practice, the cost function at a PW point is positive and that there is a future point in time with the same cost.*

Proof:

PART 1:

Assume that t is a PW point. By definition, $\exists \varepsilon_2 > 0$ such that $w(t) = \{t, v\}$ with $v = t + \varepsilon_2$. Since $\exists\ \varepsilon_1 > 0: \forall u \in [t-\varepsilon_1, t), w(u) = \{u\}$, then $\forall z$ such that $u < z < t: E_{TC}(u) < E_{TC}(z)$. Therefore, $E_{TC}'(t^-) > 0$. It can be shown analogously that $E_{TC}'(t^+) > 0$.

PART 2:

Assume $w(t) = \{t, v\}$, $v > t$, and $E_{TC}'(t) > 0$. We want to show that t is a PW point. Since $w(t) = \{t, v\}$, then $E_{TC}(t) = E_{TC}(v)$ and $\forall z \neq v \; and \; z > t, E_{TC}(z) > E_{TC}(v)$. Because $E_{TC}'(t) > 0$, $\exists\ \varepsilon_1 > 0: \forall u \in [t-\varepsilon_1, t), E_{TC}(u) < E_{TC}(t)$. So, $t \notin w(u)$. Since there is no point greater than t with a lower value than $E_{TC}(t)$, the elements of $w(u)$ must be in $[u, t)$. But since $E_{TC}' > 0 \; in \; [t-\varepsilon_1, t)$, it follows $w(u) = \{u\}$, which verifies the first condition for a PW point. By setting $v = t + \varepsilon_2$, the second condition for a PW point is trivially verified. We want to show $\forall u \in (t, v], w(u)$=$\{v\}$. Since Since $w(t) = \{t, v\}$, then $\forall z > t \; and \; z \neq v, E_{TC}(z) > E_{TC}(v)$. Therefore, $w(u) = \{v\}$. ■

The next theorems show that a PW (WP) point is followed either by a WP (PW) point, or by the point k_2. They will lay out the foundations for the algorithm that finds the intervals where it is optimal to wait (or preempt).

THEOREM S3-2. *If a point t is a PW point with left range $[t-\varepsilon_1^*, t)$, then $t-\varepsilon_1^*$ is either a WP point or k_1.*

Proof: $\forall u \in [t-\varepsilon_1^*, t): w(u) = \{u\}$, which implies $E_{TC}'(u) > 0$, otherwise $w(u)$ would contain at least another point beside u. By definition of ε_1^*, if $t - \varepsilon_1^* \neq k_1$, then $\exists \mu > 0: w(t-\varepsilon_1^* - \mu) \neq \{t - \varepsilon_1^* - \mu\}$. Therefore, $w(t-\varepsilon_1^* - \mu)$ must contain a point in the interval $(t-\varepsilon_1^* - \mu, t-\varepsilon_1^*]$. For μ that tends to 0 this implies $E_{TC}'(t-\varepsilon_1^*)^- \leq 0$ and $t-\varepsilon_1^*$ is a local minimum. Combining this result with $w(t-\varepsilon_1^*) = t-\varepsilon_1^*$, Proposition S3-1 implies that $t-\varepsilon_1^*$ is a WP point. ■

COROLLARY OF THEOREM S3-2. *If a point t is a PW point, then the largest $p \in (k_1, t)$ satisfying $E_{TC}'(p^-) \leq 0$ and $E_{TC}'(p^+) > 0$ is a WP point or, in case such p does not exist, the optimal decision in $[k_1, t]$ is to preempt.*

THEOREM S3-3. *If $E_{TC}'(k_2^-) > 0$, then either the optimal decision is to preempt in $[k_1, k_2]$ or $\exists \varepsilon^* > 0$ such that $k_2 - \varepsilon^*$ is a WP point with right range $[k_2 - \varepsilon^*, k_2]$.*

Proof: Because $E_{TC}'(k_2^-) > 0$, $\exists \varepsilon > 0\ \forall u \in [k_2 - \varepsilon, k_2], w(u) = u$. Let ε^* be the maximum value of ε that satisfies this property. If $k_2 - \varepsilon^* = k_1$, then $\forall u \in [k_1, k_2], w(u) = u$. Otherwise, $\exists \mu > 0$ such that $w(k_2 - \varepsilon^* - \mu) \neq \{k_2 - \varepsilon^* - \mu\}$. By following the same proof as in Theorem S3-2, we conclude that $k_2 - \varepsilon^*$ is a WP point.

THEOREM S3-4. *If a point t is a WP point with left range $(t-\varepsilon_1^*, t]$, then $t-\varepsilon_1^*$ is either a PW point or k_1.*

Proof: $\forall z \in (t-\varepsilon_1^*, t]\ w(z) = \{t\}$. If $t - \varepsilon_1^* = k_1$, the proof terminates; otherwise, $t - \varepsilon_1^* > k_1$. Then, $\exists \mu > 0$ such that $\forall z \in [t-\varepsilon_1^* - \mu, t-\varepsilon_1^*)\ w(z) = \{z\}$. So, for $z \to (t-\varepsilon_1^*)^-$, $w(z) = \{z\}$, while for $z \to (t-\varepsilon_1^*)^+$, $w(z) = \{t\}$. By considering that E_{TC} is continuous, it is possible to prove $E_{TC}(t-\varepsilon_1^*) =$

$E_{TC}(t)$, which, combined with $w(t) = \{t\}$ (there is no point $u > t$ such that $E_{TC}(u) \leq E_{TC}(t)$), implies $w(t - \varepsilon_1^*) = \{t - \varepsilon_1^*, t\}$. By Proposition S3-2, $t - \varepsilon_1^*$ is a PW point. ■

COROLLARY OF THEOREM S3-4. *If a point t is a WP point, then the largest $p \in (k_1, t)$ satisfying $E_{TC}'(p) > 0$ and $E_{TC}(p) = E_{TC}(t)$ is a PW point or, in case such p does not exist, the optimal decision in $[k_1, t]$ is to wait until t.*

THEOREM S3-5. *If $E_{TC}'(k_2^-) < 0$, then either the optimal decision in $[k_1, k_2]$ is to wait up to k_2 or $\exists \varepsilon^* > 0$ such that $k_2 - \varepsilon^*$ is a PW point with right range $(k_2 - \varepsilon^*, k_2)$.*

Proof: Because $E_{TC}'(k_2^-) < 0$, $\exists \varepsilon > 0\ \forall u \in [k_2 - \varepsilon, k_2], w(u) = k_2$. Let ε^* be the maximum value of ε that satisfies this property. Assume that $k_2 - \varepsilon^* > k_1$. Then, $\exists \mu > 0$ such that $\forall z \in [k_2 - \varepsilon^* - \mu, k_2 - \varepsilon^*)$ $w(z) = \{z\}$. By following the same proof as in Theorem S3-3, we conclude that $k_2 - \varepsilon^*$ is a PW point. If, on the other hand, $k_2 - \varepsilon^* = k_1$, then $\forall u \in [k_1, k_2], w(u) = k_2$.

Theorems S3-2 to S3-5 and their corollaries suggest the algorithm to detect the wait-preempt decisions (Figure S3-1).

Figure S3-1. The Analytical Method

1. Initialize the set of wait-preempt decisions $D = \{\}$ and the current point $p = t = k_2$. If $E_{TC}'(k_2^-) < 0$, then tag p as WP; if $E_{TC}'(k_2^-) > 0$, then tag p as PW.
2. While t exists
    3. If p is tagged as WP,
        4. find the largest $t < p$ such that: $\begin{cases} E_{TC}'(t) > 0 \\ E_{TC}(t) = E_{TC}(p) \end{cases}$. If such t exists, add to D the decision "*in the interval $(t, p]$: wait until p*", set $p \coloneqq t$, and tag p as PW; otherwise, add to D the decision "*in the interval $[k_1, p]$: wait until p*".
    5. Else (p is tagged as PW),
        6. find the largest $t < p$ such that: : $\begin{cases} E_{TC}'(t^-) \leq 0 \\ E_{TC}'(t^+) > 0 \end{cases}$. If such t exists, add to D the decision "*in the interval $(t, p]$: preempt*", set $p \coloneqq t$, and tag p as WP; otherwise, add to D the decision "*in the interval $[k_1, p]$: preempt*"

7. End if
8. End while

The algorithm, which we refer to as the "analytical method", scans the interval $[k_1, k_2]$ starting from k_2 and detects the WP and PW points and their ranges. Theorem S3-2 (Theorem S3-4) ensure that a PW (WP) point p is preceded by a WP (PW) point t. Initially (step 1), p is set to k_2, which, if $E_{TC}{}'(k_2^-) > 0$ (< 0), is preceded by a WP (PW) point (Theorems S3-3 and S3-5), and therefore it is "tagged" as a PW (WP) because it "plays the role" of a PW (WP) point in the preceding mechanism. Given a point p tagged as PW (WP), the preceding WP (PW) point t is found by using Proposition S3-1 (Proposition S3-2) and the Corollary of Theorem S3-2 (Theorem S3-4) in steps 5-6 (3-4). At step 6 (4), the decision relative to the right side of the newly found point t is added to the decision set D. When p is not preceded by any WP or PW point, the decision relative to the interval $[k_1, p]$ is added according to the Corollaries of Theorems S3-2 and S3-4 and the algorithm terminates.

Appendix 3.4 - Derivatives of E_T, E_W, and E_D

E_T

$$E_T(t_0, x) = \int_x^{x+d} p(t_0, t) max[0, (x + d - max[0, t])] dt$$

Case 1: $x \geq 0$

$$\frac{\partial E_T(t_0, x)}{\partial x} = \frac{\partial \int_x^{x+d} p(t_0, t)(x + d - t) dt}{\partial x}$$

$$= -dp(t_0, x) + \int_x^{x+d} p(t_0, t) dt$$

Case 2: $-d \leq x \leq 0$

$$E_T(t_0, x) = \int_x^{x+d} p(t_0, t) max[0, (x + d - max[0, t])] dt$$

$$= \int_x^0 p(t_0, t)(x + d) dt + \int_0^{x+d} p(t_0, t)(x + d - t) dt$$

$$\frac{\partial E_T(t_0, x)}{\partial x} = -p(t_0, x)(x + d) + \int_x^{x+d} p(t_0, t) dt$$

E_W

$$E_W(t_0, x) = \int_{t_0}^{x} max[t, 0] \cdot p(t_0, t) \cdot dt$$

Case 1: $t_0 \leq 0$ *and* $x \geq 0$

$$E_W(t_0, x) = \int_{t_0}^{x} \max[t, 0] \cdot p(t_0, t) \cdot dt = \int_0^{x} t \cdot p(t_0, t) \cdot dt$$

$$\frac{\partial E_W(t_0, x)}{\partial x} = \frac{\partial \int_0^{x} t \cdot p(t_0, t) \cdot dt}{\partial x} = x \cdot p(t_0, x)$$

Case 2: $t_0 \geq 0$ *and* $x \geq 0$

$$\frac{\partial E_W(t_0, x)}{\partial x} = \frac{\partial \int_{t_0}^{x} t \cdot p(t, t_0) \cdot dt}{\partial x} = x \cdot p(t_0, x)$$

Case 3: $x \leq 0$

$$\frac{\partial E_W(t_0, x)}{\partial x} = 0$$

$\boldsymbol{E_D}$

$$E_D(t_0, x) = \int_{t_0}^{x} max[0, t]\,.\,p(t_0, t) \cdot dt + \int_{x}^{\min(x+d,c)} max[0, x]\,p(t_0, t) \cdot dt$$

$$+ \int_{\min(x+d,c)}^{\min(2d,c)} max[0, t-d]\,p(t_0, t) \cdot dt =$$

Case 1: $x \leq 0$

$$E_D(t_0, x) = \int_{\min(x+d,c)}^{\min(2d,c)} max[0, t-d] p(t_0, t) \cdot dt = \int_{\min(x+d,c)}^{\min(2d,c)} (t-d) p(t, t_0) \cdot dt = 0$$

Case 2: $x \geq 0$ and $t_0 \leq 0$,

$$E_D(t_0, x) = \int_{0}^{x} t \cdot p(t_0, t) \cdot dt + x \int_{x}^{\min(x+d,c)} p(t_0, t) \cdot dt + \int_{\min(x+d,c)}^{\min(2d,c)} (t-d) p(t_0, t) \cdot dt$$

Case 2.1: $x \geq 0$, $t_0 \leq 0$, $c \leq x + d \leq 2d$

$$E_D(t_0, x) = \int_{0}^{x} t \cdot p(t_0, t) \cdot dt + x \int_{x}^{c} p(t_0, t) \cdot dt$$

$$\frac{\partial E_D(t_0, x)}{\partial x} = x \cdot p(t_0, x) + \int_{x}^{c} p(t_0, t) \cdot dt - x p(t_0, x) = \int_{x}^{c} p(t_0, t) \cdot dt$$

Case 2.2: $x \geq 0$, $t_0 \leq 0$, $x + d \leq c \leq 2d$

$$E_D(t_0, x) = \int_{0}^{t_0+ux} t \cdot p(t_0, t) \cdot dt + x \int_{x}^{x+d} p(t_0, t) \cdot dt + \int_{x+d}^{c} [t-d] p(t_0, t) \cdot dt$$

$$\frac{\partial E_D(t_0, x)}{\partial x} = x \cdot p(t_0, x) + \int_{x}^{x+d} p(t_0, t) \cdot dt + x[p(t_0, x+d) - p(t_0, x)] - x \cdot p(t_0, x+d)$$

$$= \int_{x}^{x+d} p(t_0, t) \cdot dt$$

Case 2.3: $x \geq 0, t_0 \leq 0, x + d \leq 2d \leq c$

$$E_D(t_0, x) = \int_0^x t \cdot p(t_0, t) \cdot dt + x \int_x^{x+d} p(t_0, t) \cdot dt + \int_{x+d}^{2\mathrm{d}} [t - d] p(t_0, t) \cdot dt$$

$$\frac{\partial E_D(t_0, x)}{\partial x} = x \cdot p(t_0, x) + \int_x^{x+d} p(t_0, t) \cdot dt + x[p(t_0, x + d) - p(t_0, x)] - xp(t_0, x + d)$$

$$= \int_x^{x+d} p(t_0, t) \cdot dt$$

Case 3: $x \geq 0$ and $t_0 > 0$,

$$E_D(t_0, x) = \int_{t_0}^x max[0, t] \cdot p(t_0, t) \cdot dt + \int_x^{\min(x+d,c)} max[0, x]\, p(t_0, t) \cdot dt$$

$$+ \int_{\min(x+d,c)}^{\min(2d,c)} \max[0, t - d]\, p(t_0, t) \cdot dt =$$

$$\int_{t_0}^x t \cdot p(t_0, t) \cdot dt + x \int_x^{\min(x+d,c)} p(t_0, t) \cdot dt + \int_{\min(x+d,c)}^{\min(2d,c)} (t - d) p(t_0, t) \cdot dt$$

Case 3.1: $x \geq 0, t_0 > 0, c \leq x + d \leq 2d$

$$E_D(t_0, x) = \int_{t_0}^x t \cdot p(t_0, t) \cdot dt + x \int_x^{\mathrm{c}} p(t_0, t) \cdot dt$$

$$\frac{\partial E_D(t_0, x)}{\partial x} = x \cdot p(t_0, x) + \int_x^{\mathrm{c}} p(t_0, t) \cdot dt - xp(t_0, x) = \int_x^{\mathrm{c}} p(t_0, t) \cdot dt$$

Case 3.2: $x \geq 0, t_0 > 0, x + d \leq c \leq 2d$

$$E_D(t_0, x) = \int_{t_0}^x t \cdot p(t_0, t) \cdot dt + x \int_x^{x+d} p(t_0, t) \cdot dt + \int_{x+d}^{\mathrm{c}} [t - d] p(t_0, t) \cdot dt$$

$$\frac{\partial E_D(t_0, x)}{\partial x} = x \cdot p(t_0, x) + \int_x^{x+d} p(t_0, t) \cdot dt + x[p(t_0, x + d) - p(t_0, x)] - xp(t_0, x + d)$$

$$= \int_x^{x+d} p(t_0, t) \cdot dt$$

Case 3.3: $x \geq 0, t_0 > 0, x + d \leq 2d \leq c$

$$E_D(t_0, x) = \int_{t_0}^{x} t \cdot p(t_0, t) \cdot dt + x \int_{x}^{x+d} p(t_0, t) \cdot dt + \int_{x+d}^{2\mathrm{d}} [t - d] p(t_0, t) \cdot dt$$

$$\frac{\partial E_D(t_0, x)}{\partial x} = x \cdot p(t_0, x) + \int_{x}^{x+d} p(t_0, t) \cdot dt + x[p(t_0, x + d) - p(t_0, x)] - xp(t_0, x + d)$$

$$= \int_{x}^{x+d} p(t_0, t) \cdot dt$$

Appendix 3.5 – Derivation of Table 3-2

In deriving the cases of Table 3-2, let us first consider the case $c < d$. Since $min(c, x + d) = c$, the derivative is:

$$E'_{TC}(x) = \frac{q}{(c-a)(c-b)}[x^2(-\omega + \tau) + x(2\omega d - 2\tau c) - 2\omega dc + \omega c^2 + \tau c^2] \qquad \text{(S5-1)}$$

The two roots of $E'_{TC}(x)$ are

$$x_1 = \frac{2\tau c - 2\omega c}{-2\omega + 2\tau} = c$$

$$x_2 = \frac{-4\omega d + 2\tau c + 2\omega c}{-2\omega + 2\tau}$$

If $\tau > \omega$, $E'_{TC}(x) \geq 0$ if and only if $x \leq x_2$ or $x \geq x_1$. Note that $x_2 \leq x_1$ and that $x_2 \geq 0$ if $c \geq \frac{2\omega}{\tau+\omega} d$. So, if $\frac{2\omega}{\tau+\omega} d \leq c < d$, the cost function increases in $[0, x_2)$ and then decreases in $(x_2, c]$ (Table 3-2, case 1). Otherwise, $x_2 \leq 0$ and the cost function decreases in $[0, x_1] = [0, c]$ (Table 3-2, case 2). If, on the other hand, $\tau < \omega$, then $x_1 \leq x_2$ and $E'_{TC}(x) \geq 0$ if and only if $c = x_1 \leq x \leq x_2$. (Table 3-2, case 3).

If $c \geq d$, the formula of the derivative depends on whether $x \in [0, c - d)$ or $x \in [c - d, d]$. Let us analyze the case $x \in [c - d, d]$ first, where the derivative is given by (S5-1). If $\tau > \omega$, then $x_1 \leq x_2$ and $E'_{TC}(x) \geq 0$ if and only if $x \leq x_1 = c$ or $x \geq x_2$. In that case, the cost function is increasing in $x \in [0, c]$ (case 4). If, on the other hand, $\tau < \omega$, then $x_2 \leq x_1$ and $E'_{TC}(x) \geq 0$ if and only if $x_2 \leq x \leq x_1$. Since we are considering the case $x \in [c - d, d]$, it is important to distinguish between the following two subcases. If $d \leq c \leq \frac{3\omega-\tau}{2\omega} d$, then $x_2 \in [c - d, d]$ In this subcase, the cost function decreases in $[c - d, x_2)$ and then increases in $(x_2, d]$ (case 5). Otherwise, when $d \leq \frac{3\omega-\tau}{2\omega} d \leq c$, it increases in $[c - d, d]$ (cases 6 and 7).

Let us complete the analysis of cases 4, 5, 6, and 7 by considering the interval $[0, c - d)$, where the derivative can be expressed as:

$$E'_{TC}(x) = \frac{2q}{(c-a)(c-b)}\left[-\omega x^2 + \omega c x - \frac{1}{2}\omega d^2 + \tau c d - \frac{1}{2}\tau d^2 - \tau x d\right] \quad \text{(S5-2)}$$

$E'_{TC}(x) \geq 0$ if the roots are real and x is included between them, i.e., $E'_{TC}(x) \geq 0$ if and only if:

1. $\omega^2 c^2 + \tau^2 d^2 + 2\omega\tau c d - 2\omega^2 d^2 - 2\omega\tau d^2 \geq 0$ and
2. $x_3 \leq x \leq x_4$

where

$x_3 = \frac{\omega c - \tau d - \sqrt{\omega^2 c^2 + \tau^2 d^2 + 2\omega\tau c d - 2\omega^2 d^2 - 2\omega\tau d^2}}{2\omega}$ and $x_4 = \frac{\omega c - \tau d + \sqrt{\omega^2 c^2 + \tau^2 d^2 + 2\omega\tau c d - 2\omega^2 d^2 - 2\omega\tau d^2}}{2\omega}$. Also, it is straightforward to verify that the second derivative decreases in $[0, c-d]$. It is also useful to determine the sign of the derivative right after *T*'s scheduled time, i.e., at $t = 0^+$.

PROPOSITION S5-1: $E'_{TC}(0^+) \geq 0$ if and only if $\frac{2\omega}{\omega+\tau} d \leq c < d$ or ($\frac{\omega+\tau}{2\tau} d \leq c$ and $c \geq d$)

Proof:

Case $c < d$:

If $c < d$, the derivative at 0^+ can be written as (S5-1):

$$E'_{TC}(x) = \frac{q}{(c-a)(c-b)}[x^2(-\omega+\tau) + x(2\omega d - 2\tau c) - 2\omega d c + \omega c^2 + \tau c^2]$$

$$E'_{TC}(0^+) = \frac{q}{(c-a)(c-b)}[-2\omega d c + \omega c^2 + \tau c^2]$$

$E'_{TC}(0^+) \geq 0$ if $-2\omega d + \omega c + \tau c \geq 0$, that is, $c \geq \frac{2\omega}{\omega+\tau} d$

So, case 1 can be summarized as $\frac{2\omega}{\omega+\tau} d \leq c < d$

Case $c \geq d$:

If $c \geq d$, the derivative at 0^+ can be written as (6):

$$E'_{TC}(x) = \frac{2q}{(c-a)(c-b)}\left[-\omega x^2 + \omega c x - \frac{1}{2}\omega d^2 + \tau c d - \frac{1}{2}\tau d^2 - \tau x d\right]$$

$$E'_{TC}(0^+) = \frac{2q}{(c-a)(c-b)}\left[-\frac{1}{2}\omega d^2 + \tau c d - \frac{1}{2}\tau d^2\right]$$

$E'_{TC}(0^+) \geq 0$ if $-\frac{1}{2}\omega d + \tau c - \frac{1}{2}\tau d \geq 0$. That is, $c \geq \frac{\omega+\tau}{2\tau} d$

So, case 2 can be summarized as $\frac{\omega+\tau}{2\tau} d \leq c$ and $c \geq d$. ■

In case 4, $\tau > \omega$ and $c \geq d$. So, $\frac{\omega+\tau}{2\tau} d \leq d \leq c$, and by Proposition S5-1, $E'_{TC}(0^+) \geq 0$. Considering also that $E'_{TC}(c-d) \geq 0$ and that the second derivative decreases, it follows that $E'_{TC}(x) \geq 0$ in $[0, c-d]$. Therefore, in case 4 $E'_{TC}(x) \geq 0$ in $[0, d]$.

In case 5, $\tau < \omega$ and $d \leq c \leq \frac{3\omega-\tau}{2\omega} d$. Note that $\frac{3\omega-\tau}{2\omega} d \leq \frac{\omega+\tau}{2\tau} d$. By Proposition S5-1, $E'_{TC}(0^+) \leq 0$. Considering also that the second derivative decreases, we conclude that in case 5, $E'_{TC}(x) \leq 0$ in $[0, x_2]$ and $E'_{TC}(x) \geq 0$ in $[x_2, d]$.

In case 6, $\tau < \omega$ and $\leq \frac{3\omega-\tau}{2\omega} d \leq c$ By the definition of case 6 in Table 3-2, $c \leq \frac{\omega+\tau}{2\tau} d$. Therefore, $E'_{TC}(0^+) \leq 0$. However, it is easy to verify that $E'_{TC}(c-d) \geq 0$. Considering also that the second derivative decreases, E'_{TC} must change sign at x_3. It follows that $E'_{TC}(x) \leq 0$ in $[0, x_3)$ and $E'_{TC}(x) \geq 0$ in $(x_3, d]$.

In case 7, $\tau < \omega$ and $d \leq \frac{3\omega-\tau}{2\omega} d \leq c$. By the definition of case 7 in Table 3-2, $\frac{\omega+\tau}{2\tau} d \leq c$. Therefore, $E'_{TC}(0^+) \geq 0$. Since $E'_{TC}(c-d) \geq 0$ and the second derivative decreases, $E'_{TC}(x) \geq 0$ in $[0, c-d]$. Therefore, in case 7, $E'_{TC}(x) \geq 0$ in $[0, d]$.