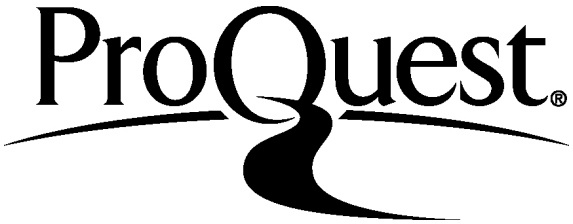
BEHAVIOR, FLEXIBILITY, AND QUALITY IN SERVICE OPERATIONS MANAGEMENT

David D. Cho

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements for the degree
Doctor of Philosophy
in the Kelley School of Business
Indiana University

August 2015



Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____
Kurt M. Bretthauer, Ph.D., Chairman

_____
Kyle D. Cattani, Ph.D.

_____
Alex F. Mills, Ph.D.

_____
Jonathan E. Helm, Ph.D.

August 6, 2015

David D. Cho

BEHAVIOR, FLEXIBILITY, AND QUALITY IN SERVICE OPERATIONS MANAGEMENT

This dissertation bridges two areas in service operations management: (1) capacity planning with the presence of flexible capacity and (2) behavioral and quality phenomena. In service processes, the quality of the work and the behavior of the workers are greatly influenced by the level of work assigned to each worker, which is determined by a firm's capacity planning decisions and utilization of the flexibility. Therefore, we incorporate the behavioral and quality impacts into capacity planning with the presence of flexible capacity. In Chapter 2, we present nurse staffing models that incorporate patient outcomes, nurse burnout, length of stay, and costs when a hospital is setting patient-to-nurse ratios. By incorporating patient and nurse outcomes, we show that lower patient-to-nurse ratios can potentially provide financial benefits in addition to improving the quality of care that hospitals provide. In Chapter 3, we model speedup and slowdown of workers in a very general way to represent many possible joint effects of these behavioral phenomena. We use this model to study the impact of speedup and slowdown on a multi-period workforce staffing problem with recourse. Our results show that the slowdown effect can be strong enough in most settings to cause the firm to aggressively utilize expensive on-call workers to avoid future system congestion. In Chapter 4, we compare traditional and open-access scheduling policies for outpatient medical practices in terms of the number of patients served and financial performance. In contrast to earlier works, we consider the optimal average number of patients served and find that while the

traditional policy may be more profitable by providing doctors more control over their schedule and ability to limit overtime, the open-access policy may lead doctors to serve a greater number of patients. Overall, this dissertation shows that the firm can enjoy the benefits of improved service and better financial performance by taking behavior, quality, and flexibility into consideration for its capacity planning decisions.

_____

Kurt M. Bretthauer, Ph.D., Chairman

_____

Kyle D. Cattani, Ph.D.

_____

Alex F. Mills, Ph.D.

_____

Jonathan E. Helm, Ph.D.

TABLE OF CONTENTS

1. INTRODUCTION 1

2. NURSE STAFFING RATIOS: A CASE FOR HIGHER QUALITY OF CARE 4

2.1. Introduction 5

2.2. Literature Review 8

2.2.1. Hospital Capacity Planning and Nurse Staffing Literature 8

2.2.2. Patient Outcomes, Nurse Satisfaction, and Hospital Outcomes Literature 10

2.3. Patient- and Nurse-Oriented Staffing Ratio Decisions 14

2.3.1. Incorporating Patient Outcomes into the Nurse Staffing Ratio Decision . 14

2.3.2. Patient Length of Stay and Nurse Turnover 19

2.4. Model with Average Patient Demand 21

2.4.1. Base Model 21

2.4.2. Patient Length of Stay and Nurse Turnover 22

2.4.3. Results 23

2.5. Sensitivity Analysis 28

2.6. Conclusions and Future Research 33

3. BEHAVIOR-AWARE WORKFORCE STAFFING 35

3.1. INTRODUCTION 35

3.1.1. Incorporation of Speedup and Slowdown 38

3.1.2. Example: Hospital Nurse Staffing 40

3.2. LITERATURE REVIEW 43

3.2.1. Speedup and Slowdown 43

3.2.2. Staffing with Recourse 47

3.2.3. Patient-to-Nurse Ratios 48

3.3. MODEL 50

3.3.1. Speedup and Slowdown 50

3.3.2. Markov Decision Process Model 54
3.4. ANALYSIS 58
3.4.1. "Speedup Dominates" and "U-Shape" Cases 59
3.4.2. "Slowdown Dominates" and "Inverse U-Shape" Cases 59
3.4.3. Shift-by-Shift Staffing with No Capacity Constraint 60
3.5. NUMERICAL EXPERIMENTS 63
3.5.1. Optimal Recourse Actions 64
3.5.1.1. Sensitivity Analysis 67
3.5.1.2. Cyclic Arrivals ($\boldsymbol{k > 1}$) 71
3.5.2. Comparison with Different Heuristics 72
3.5.3. Sensitivity to a Suboptimal Advance Schedule 76
3.6. CONCLUSION AND FUTURE RESEARCH 79

4. THE PATIENT PATIENT: THE PERFORMANCE OF TRADITIONAL VERSUS OPEN-ACCESS SCHEDULING POLICIES 81

4.1. Introduction 81
4.2. Literature Review 84
4.3. Model 86
4.3.1. Traditional Scheduling Policy 88
4.3.2. Open-Access Scheduling Policy 91
4.3.3. Comparison of the Two Policies: Profitability and Number of Patients Served 93
4.4. A Patient-Friendly Doctor ($\boldsymbol{\gamma = 0}$) and Uniform Demand under the Open-Access Policy 94
4.5. Numerical Comparison of Traditional and Open-Access Policies with Poisson Arrivals 98
4.6. Conclusion 106

5. CONCLUSION 109

Appendix A Proofs for Chapter 2 111

Appendix B More Details and Proofs for Chapter 3 113

Appendix B.1. Notations 113
Appendix B.2. Finite-Horizon Model 114
Appendix B.3. Proofs 116

Appendix C Proofs for Chapter 4 122

References 128

Curriculum Vitae .....

CHAPTER 1

INTRODUCTION

This dissertation bridges two areas in service operations management: (1) capacity planning with the presence of flexible capacity and (2) behavioral and quality phenomena. In service processes, the quality of the work and the behavior of the workers are greatly influenced by the level of work assigned to each worker, which is determined by a firm's capacity planning decisions and its utilization of the flexibility. Consequently, the importance of behavioral and quality impacts and the role of flexibility in capacity planning need to be incorporated in the decision-making process. Recent behavioral research has shown that several assumptions made by traditional operations management decision models do not always hold. For example, in healthcare operations, failure to consider how the patient-to-nurse ratio impacts patient quality outcomes and nurse behavior ignores potential cost savings resulting from higher quality of care and improved nurse satisfaction. Moreover, while service staffing models commonly assume that the service rate is fixed, recent empirical studies have shown that servers employ a changing service rate (either deliberately or as an unintended consequence of system congestion) depending on the assigned workload. In particular, of the three essays in this dissertation, two study worker staffing decisions while taking into account the impacts of those decisions on quality of service and workforce behavior. The third essay examines the effects of flexibility in service systems by investigating a healthcare provider's strategy on utilizing appointment slots to meet patient demand.

Chapter 2, "Nurse Staffing Ratios: A Case for Higher Quality of Care," considers the problem of setting appropriate nurse staffing ratios in a hospital, an issue that is both

complex and widely debated. While this staffing problem has received considerable attention in the operations management literature, traditional operations management models have failed to take advantage of extensive behavioral and quality results on patient and nurse outcomes. For example, empirical studies have shown that each additional patient assigned per nurse in a hospital is associated with a 7% increase in mortality rates and a 23% increase in nurse burnout (Aiken et al. 2002). In Chapter 2, we study the patient-to-nurse ratio decision using stochastic programming methods while incorporating the impacts of the decision on various outcomes such as patient length-of-stay and nurse turnover. Our results show that lower patient-to-nurse ratios can be more cost-effective than higher ratios by allowing hospitals to provide better quality of care and decrease the costs incurred by adverse patient and nurse outcomes, overcoming the higher wage costs incurred by staffing more nurses.

Chapter 3, "Behavior-Aware Workforce Staffing," builds on the first essay and studies how to incorporate the behavioral issues of speedup and slowdown into workforce staffing decisions. Service staffing models commonly assume that the rate at which servers work on each customer is fixed. However, human behavior can be more complex. Several recent empirical studies show that workers may have a tendency to increase their service rate, called *speedup*, to account for an increase in workload. On the other hand, the service rate may also decrease, called *slowdown*, if the system is too congested. We first model speedup and slowdown separately in a very general way and use a convex combination of the two functions to represent many possible joint effects of these behavioral phenomena. We then incorporate this model and study the decision of worker staffing under the joint effects of speedup and slowdown using a stochastic dynamic

program in which the worker productivity depends on the workload, or ratio of customers to servers. Our results show that, when the joint effects of speedup and slowdown are taken into account, a dynamic recourse policy for which the optimal workload varies in the number of customer requests in the system is optimal in most settings. We also identify conditions under which the optimal workload is independent of the number of customer requests. Furthermore, using numerical studies, we propose that a one-step look-ahead policy is a viable alternative to the optimal policy.

While the first two chapters consider capacity planning from the supply side, Chapter 4, "The Patient Patient: The Performance of Traditional versus Open-Access Scheduling Policies," studies the demand side of healthcare providers by comparing the performance of two appointment scheduling policies with different levels of flexibility. Under a traditional scheduling policy, a patient schedules an appointment in advance and thus there is a possibility of patient no-shows. In response, doctors overbook patients to prevent idle time created by no-shows. Under an open-access scheduling policy, a patient is only allowed to schedule a same-day appointment, thereby eliminating patient no-shows but creating more randomness in the daily number of scheduled appointments. In contrast to earlier works, our results show that while the traditional policy may be more profitable by providing doctors more control over their schedule and ability to limit overtime, the open-access policy may lead to doctors to serve a greater number of patients. Our results provide insights that can help policy makers to better incentivize the doctors to implement the open-access policy, which is socially optimal.

CHAPTER 2

NURSE STAFFING RATIOS: A CASE FOR HIGHER QUALITY OF CARE

Abstract

We consider the problem of setting appropriate nurse staffing ratios in a hospital, an issue that is both complex and widely debated. Despite the considerable attention given to healthcare in operations research and operations management, there has been only limited effort to take advantage of the extensive results available from the medical, nursing, and healthcare services literature to make the research more patient- and nurse-oriented, both of which are critical concerns when deciding patient-to-nurse ratios. For example, empirical studies have shown that each additional patient assigned per nurse in a hospital is associated with a 7% increase in mortality rates and a 23% increase in nurse burnout. Failure to consider impacts such as these not only limits the relevance of many healthcare decision models, but also ignores potential cost savings resulting from providing higher quality of care and improved nurse satisfaction. Thus, we present nurse staffing models that incorporate patient outcomes, nurse burnout, length of stay, and costs when a hospital is setting patient-to-nurse ratios. We present results based on data collected from three hospitals. By incorporating patient and nurse outcomes, we show that lower patient-to-nurse ratios can potentially provide financial benefits in addition to improving the quality of care that hospitals provide.

2.1. Introduction

Today's healthcare industry faces a wide range of challenges. Healthcare costs and healthcare expenditures remain high (Martin et al. 2012), and problems such as emergency department overcrowding (U.S. General Accounting Office (GAO) 2009, Pitts et al. 2012), lack of access to care (Gulliford and Morgan 2003, Bodenheimer and Pham 2010), and a shortage of nurses in the United States and Europe persists (Juraschek et al. 2012, OECD/European Union 2014). In addition, many U.S. states have enacted, or are considering enacting, legislation mandating minimum patient-to-nurse ratios in hospitals (Aiken et al. 2010). For example, when California was initially considering legislation setting required patient-to-nurse ratios for typical medical/surgical inpatient wards, hospital management suggested ratios as high as 10:1, and nursing unions suggested as low as 3:1 (Spetz 2004). Lower patient-to-nurse ratios result in the need for more nurses working at a given time and thus higher staffing costs, but also are associated with higher quality of patient care, less nurse turnover, and shorter patient length of stay in the hospital, as shown in studies of hospitals in the United States and Europe (Aiken et al. 2002, Kane et al. 2007b, Aiken et al. 2012, Aiken et al. 2014). These tradeoffs between nurse capacity, quality of care, nurse satisfaction, and costs are complex and at the heart of the debate regarding optimal patient-to-nurse ratios. Therefore, in this chapter we address the following research questions: (1) How can these tradeoffs be taken into account when setting patient-to-nurse ratios? (2) Is there a business case for higher quality of patient care?

Further motivating the importance of issues regarding appropriate capacity levels, the Institute of Medicine (2001) set forth six "aims for improvement" in healthcare

delivery, stating that the healthcare industry should strive to become *safe*, *effective*, *patient-centered*, *timely*, *efficient*, and *equitable*. The goals to be timely and efficient fit in well with traditional cost driven operations management research, and there are already significant contributions along these lines. However, as discussed above, healthcare capacity planning decisions also significantly affect other important patient- and nurse-oriented performance measures. Most operations management literature fails to take into account these considerations. For example, based on data collected from 168 hospitals, Aiken et al. (2002) show that each additional patient assigned per nurse is associated with a 7% increase in the likelihood of dying within 30 days of admission and a 23% increase in nurse burnout. Kane et al. (2007b) provide estimates of the impact of nurse staffing levels on a variety of patient outcomes based on an overview of the literature. Phibbs et al. (2007) show that the size of inpatient hospital units can affect patient mortality rates as well. As will be discussed in more detail in the next section, there is a large body of this type of medical, nursing, and healthcare services literature that reports empirical results on the relationship between several types of operations management decision variables, such as nurse staffing and hospital volume, and a variety of patient and nurse outcomes. Failure to consider how patient-to-nurse ratios may impact patient outcomes and nurse turnover not only limits the relevance of healthcare operations research decision models, but also ignores potential cost savings resulting from higher quality of care and improved nurse satisfaction. Therefore, in this chapter we consider the question of under what circumstances providing higher nursing capacity levels with the associated higher wage costs can actually lead to lower total costs due to better patient care and less nurse turnover.

To address this question, we present nurse staffing models that incorporate the impact of patient-to-nurse ratios on patient outcomes and nurse satisfaction. We consider both stochastic and deterministic patient demand cases. The decisions addressed include patient-to-nurse ratios, usage of unit nurses, and usage of external agency nurses. The models incorporate empirical findings on the relationship between nurse staffing ratios and patient outcomes as well as nurse satisfaction. We report numerical results based on data collected from three participating hospitals. Our results show that lower patient-to-nurse ratios can be more cost-effective than the higher ratios by allowing hospitals to provide better quality of care and decrease adverse patient and nurse outcomes. That is, minimizing cost and maximizing quality are not necessarily at odds with each other. Our goal is to present a methodology to aid hospitals when setting nurse staffing levels.

The remainder of the chapter is organized as follows. Section 2.2 discusses two streams of literature relevant to this study: (i) operations research and operations management literature focused on hospital capacity planning and nurse staffing and (ii) medical, nursing, and healthcare services literature reporting empirical results on how various operations management decision impact patient outcomes, nurse outcomes, and hospital outcomes. Based on this literature, Section 2.3 presents nurse staffing models that incorporate the impact of patient-to-nurse ratios on patient outcomes and nurse turnover. Section 2.4 presents nurse staffing insights based on simplified models that treat patient demand as deterministic instead of stochastic. Section 2.5 presents sensitivity analysis on some of the key input parameters. Section 2.6 provides some concluding remarks and opportunities for future research.

2.2. Literature Review

First we present a review of recent research in hospital capacity planning from the operations research and operations management literature, illustrating the lack of work with regards to incorporating patient outcomes and nurse satisfaction into the models. Then we review the large body of literature in the medical, nursing, and healthcare services fields that study the relationship between a variety of operations management-related decision variables and patient outcomes, nurse outcomes, and hospital outcomes.

2.2.1. Hospital Capacity Planning and Nurse Staffing Literature

Healthcare operations research has been studied for many years, with hospital capacity planning being an important topic within this broad field (Green 2005). Hospital capacity planning involves a wide variety of decisions, including facility size and location (Daskin and Dean 2005), nurse staffing (Cheang et al. 2003, Burke et al. 2004), number of beds and patient flow (Thompson et al. 2009, Dobson et al. 2010, Bretthauer et al. 2011, Mandelbaum et al. 2012), major equipment acquisition (e.g., MRI), surgical scheduling (Olivares et al. 2008, Denton et al. 2010, May et al. 2011), screening and biopsy schedule decisions (Chhatwal et al. 2010, Rauner et al. 2010), and patient appointments (Green and Savin 2008, Gupta and Denton 2008, Lee and Zenios 2009, Dobson et al. 2011, Wang and Gupta 2011). In the remainder of this subsection, we will focus on the literature most relevant to this study, nurse staffing.

Traditionally, the nurse staffing literature focuses on developing efficient work schedules that minimize wage costs. More recently, researchers have incorporated different types of resources and studied the interdependencies among different capacity

decisions. White et al. (2011) examine the benefits of integrating capacity, patient flow, and scheduling in outpatient clinics. Gnanlet and Gilland (2009) and Dobson et al. (2009) address decisions regarding resource allocation and scheduling of labor in healthcare services. California Bill AB 394, which mandates fixed patient-to-nurse staffing ratios for hospitals, inspired debate within the healthcare operations management community, and motivated numerous researchers and their studies. Wright et al. (2006) analyze the impact of mandatory patient-to-nurse ratios on nurse schedule costs, and find that nurse wage costs can be highly nonlinear with respect to changes in patient-to-nurse ratios. De Véricourt and Jennings (2011) model the nurse staffing decision as a closed M/M/s//n queueing system. They determine that it is more effective to deviate from threshold-specific patient-to-nurse ratios, diverging from the recent trend of implementing mandatory ratios. Yankovic and Green (2011) develop a queueing model to represent interaction between the nurse and bed systems. They discuss problems in using rigid patient-to-nurse ratios across a broad range of hospital units. Green et al. (2013) present the optimal staffing levels incorporating the effect of absenteeism rate that is a function of the number of nurses scheduled. We also consider the issue of mandatory patient-to-nurse ratios and present methods for establishing the optimal ratios for hospitals. Our essay differs from other work by taking patient outcomes as well as nurse outcomes into account when determining nurse staffing policies.

Another branch of research that is related to this essay is the utilization of flexible and cross-trained workers. Campbell (1999) explores the benefits of cross-utilization by developing a model for allocating cross-trained workers at the beginning of a shift in a multidepartment service environment. Pinker and Shumsky (2000) suggest that benefits

gained in efficiency may be lost in quality. Jordan et al. (2004) and Hopp et al. (2004) investigate the performance and robustness of chaining, where a few workers are strategically cross-trained. Simchi-Levi and Wei (2012) show mathematically why long chain, in which plants are endowed with the capacity to produce exactly two different products and every product is produced by exactly two plants, is nearly as efficient as full flexibility. Easton (2011) uses a two-stage stochastic model to analyze the relationship between cross-training and scheduling flexibility. Wright and Bretthauer (2010) develop a model that coordinates nurse scheduling, short term adjustments to the schedule, and float and travel nurse decisions to evaluate the benefits of using a flexible workforce. Our models include external agency nurses, while also incorporating patient and nurse outcomes, something not done in the previously mentioned studies.

2.2.2. Patient Outcomes, Nurse Satisfaction, and Hospital Outcomes Literature

The hospital capacity planning literature discussed above typically addresses the impact of capacity planning decisions on efficiency and costs related to nurse wages, beds, and facilities. However, those decisions also have a significant impact on patient outcomes, nurse satisfaction, and a variety of other hospital performance measures. By modeling the impact of nurse staffing levels on patient outcomes and nurse satisfaction, our capacity planning models help to fill this gap in the literature. Fortunately, there is a large body of empirical studies related to capacity planning decisions in the medical, nursing, and healthcare services literature.

Table 2.1 categorizes the medical, nursing, and healthcare services literature according to the type of hospital capacity planning decisions studied and the outcomes

that the decisions impact. As can be seen below, many of the decisions addressed in this literature are conventional types of operations management decisions such as the volume of hospitals and physicians, number of hospital beds, hospital occupancy, and the care environment. Occupancy rate and patient volume are shown to have considerable impacts on mortality rates and process quality (Bond et al. 1999, Phibbs et al. 2007, Theokary and Ren 2011). Care environments of hospitals also have a significant influence on medical and nursing outcomes. According to Aiken et al. (2008), patients have significantly lower mortality rates and failure to rescue in hospitals with better care environments. Spence Laschinger and Leiter (2006) conclude that both patient safety outcomes and nurse burnout are related to the quality of the nursing practice work environment.

Patient-to-nurse ratios and nurse staffing levels are the decisions in the medical literature that are most relevant to this chapter. Patient-to-nurse ratios are one of the capacity planning decisions that have received much attention from nursing and medical scholars for the past several years. Aiken et al. (2002) analyze the association between patient-to-nurse ratios and patient mortality, failure-to-rescue, and factors related to nurse retention. They find that each additional patient per nurse is associated with a significant increase in patient mortality and a significant decrease in the odds of job satisfaction. Rothberg et al. (2005) evaluate the impact of various nurse staffing ratios and study the tradeoff between costs and mortality rates. Kane et al. (2007b) examine the association between nurse staffing and patient outcomes in acute care hospitals and find increased staffing to be associated with lower mortality, lower length of stay, and decreased odds ratios of various adverse medical outcomes. McCue et al. (2003) find that an increase in nurse staffing levels is associated with an increase in operating costs, but find no

statistically significant decrease in profits. Also, Kane et al. (2007b) and Shamliyan et al. (2009) discuss the testing of causality between nurse staffing levels and outcome measures. Lankshear et al. (2005) state that the weight of evidence in their study is strongly suggestive of a causal relationship. Lin (2014) identifies a causal relationship between nurse staffing and quality of care in nursing homes and finds that registered nurse staffing has a large and significant impact on quality of care.

There have been numerous additional studies on the effect of nurse staffing levels and utilization of a flexible workforce on medical outcomes. For example, Cho et al. (2003) observe that a 10% increase in registered nurse proportion of the nursing personnel is associated with a 9.5% decrease in the odds of pneumonia. According to Newhouse et al. (2005), a 10% increase in agency nurse use is related to a significant decrease in the estimated odds of death. Aiken et al. (2007) study the relationship between supplemental nurse staffing and quality of care and find that each 10% increase in the proportion of nonpermanent nurses results in a 9% decrease in permanent nurse burnout, a 28% increase in the likelihood of leaving within one year, and improved patient outcomes. Needleman et al. (2006) show that an increase in hospital costs may be justified by a reduction in adverse outcomes and patient deaths depending on the value patients and payers assign to avoided deaths and complications. Hugonnet et al. (2007) find that a higher staffing level is associated with a greater than 30% infection risk reduction. Furthermore, the results in Tourangeau et al. (2007) show that the proportion of registered nurses in the staff mix and the level of education for the nurses have a significant effect on the 30-day mortality rate.

Healthcare Operations Management Decision	Patient Outcomes and Satisfaction			Nurse Outcomes and Satisfaction		Hospital Outcomes
	Mortality, Failure to Rescue	*Length of stay*	*Other Adverse Patient Events*	*Burnout*	*Job Satisfaction*	*Financial*
Patient-to-Nurse Ratio	Aiken et al. (2002), Rothberg et al. (2005), Kane et al. (2007b), Aiken et al. (2010), Needleman et al. (2011)	Rothberg et al. (2005), Kane et al. (2007b)	Hugonnet et al. (2007), Kane et al. (2007b), Aiken et al. (2010)	Aiken et al. (2002), Gurses et al. (2009), Aiken et al. (2010)	Aiken et al. (2002), Aiken et al. (2010)	Rothberg et al. (2005)
Agency/Float Nurse Mix	Newhouse et al. (2005)	Newhouse et al. (2005), Phibbs et al. (2009)	Newhouse et al. (2005), Aiken et al. (2007), Bae et al. (2010)	Aiken et al. (2007)	Aiken et al. (2007)	
Staffing Levels (Nurses and/or others)	Bond et al. (1999), Needleman et al. (2002), Cho et al. (2003), Stanton and Rutherford (2004), Tourangeau et al. (2007)	Needleman et al. (2002), Cho et al. (2003)	Pronovost et al. (2001), Cho et al. (2003), Stanton and Rutherford (2004), Needleman et al. (2006)	Gurses et al. (2009)	Stanton and Rutherford (2004)	Cho et al. (2003), McCue et al. (2003), Stanton and Rutherford (2004), Needleman et al. (2006)
Hospital/Physician/Surgeon Volume	Halm et al. (2002), Peelen et al. (2007), Iwashyna et al. (2009)	Kahn et al. (2006)	Kahn et al. (2006), Iwashyna et al. (2009)			
Number of Hospital Beds	Phibbs et al. (2007)		Phibbs et al. (2007)			
Hospital Occupancy	Bond et al. (1999)					
Care Environment	Aiken et al. (2008)		Spence Laschinger and Leiter (2006)	Spence Laschinger and Leiter (2006), Aiken et al. (2008)	Aiken et al. (2008)	

Table 2.1. Medical, Nursing, and Healthcare Services Literature Involving Healthcare Operations Management Decisions

Despite the availability of this large body of empirical studies from the nursing and medical literature, there has been very little effort to utilize them in healthcare operations research decision models. A contribution of this essay is to begin bridging that gap by incorporating various nurse and patient outcomes in an operations research model for determining patient-to-nurse ratios and staffing levels.

2.3. Patient- and Nurse-Oriented Staffing Ratio Decisions

We begin with two models that capture the key element of the essay: to put a dollar figure on quality of patient care and nurse turnover and analyze how it impacts the nurse staffing ratio decision. The first model illustrates how to incorporate patient quality of care and adverse outcomes into the staffing ratio decision. The second model focuses on the effect of patient-to-nurse ratios on patient length of stay and nurse turnover.

2.3.1. Incorporating Patient Outcomes into the Nurse Staffing Ratio Decision

We begin by presenting a base model that incorporates the impact of patient-to-nurse ratios on patient outcomes when making nurse staffing decisions. With this model, we can consider such outcomes as the relationship between patient-to-nurse ratios and bloodstream infections, hospital-acquired pneumonia, patient mortality, unplanned extubation, urinary tract infections, patient falls, and other patient outcomes. For example, Kane et al. (2007b) report that each additional patient assigned per nurse is associated with a 16% increase in nosocomial bloodstream infections. The cost per patient with a bloodstream infection has been reported in the range of $20,000 to $30,000 (Anderson et al. 2007, Scott 2009) and the infection rate has been reported at 0.36 per 1,000 patient

days (Anderson et al. 2007). With this data and our model, we can attach a cost to bloodstream infections as a function of the patient to nurse ratio.

Decision Variables – Problems (P1) and (P2)	
r	Patient-to-nurse ratio for the unit (stage 1 decision)
n	Weekly nurse staffing level measured as "nurse shifts" in the unit (stage 1)
n_A	Weekly external agency nurse shifts needed (stage 2 decision, i.e., recourse)
Decision Variable – Problems (P3) and (P4)	
r	Patient-to-nurse ratio for the unit
Parameters	
λ	Number of patients admitted per week in the unit
g	Weekly demand measured as "patient shifts" in the unit ("patient days" × number of shifts per day)
$g(\omega)$	Weekly demand ("patient shifts") in the unit with c.d.f. Φ resulting from outcome ω of a random event
δ	Adjustment factor to account for uncertainty in staffing needs
γ_i	Rate of occurrence for adverse outcome i
β_i	Odds ratio for adverse outcome i with respect to a one unit increase in the patient-to-nurse ratio
β_{LOS}	Odds ratio for patient length of stay with respect to a one unit increase in the patient-to-nurse ratio
β_{TO}	Odds ratio for nurse burnout with respect to a one unit increase in the patient-to-nurse ratio
ρ	Weekly nurse turnover rate
μ	Nurse burnout to turnover conversion factor
σ	Average number of shifts worked in a week by each nurse
r_b	Base level patient-to-nurse ratio in the unit
$\underline{r}, \overline{r}$	Lower and upper bounds on patient-to-nurse ratio r
c_i	Cost per occurrence of adverse outcome i
c_s	Wage per shift per nurse
c_A	Wage per shift per agency nurse
c_{bed}	Cost per shift of an occupied bed (excluding nurse wages) at base patient-to-nurse ratio
c_{TO}	Cost per nurse turnover

Table 2.2. Notation

Consider a typical medical/surgical unit in a hospital. We present a two-stage stochastic programming model where the stage one decisions of interest include setting the patient-to-nurse ratio in the unit and deciding the number of nurses in the unit

available to work a shift at the beginning of time horizon. As the patient-to-nurse ratio increases, fewer nurses are needed to care for the patients and staffing costs go down, but quality of care declines and the corresponding quality costs go up. Patient demand is random. If patient demand exceeds the nurse capacity, then the stage two (recourse) decision involves obtaining additional nurses from an external agency, typically at a higher cost than the hospital unit nurses. Agency nurses give hospitals the short term ability to adhere to the patient-to-nurse ratio. We assume infinite agency nurse pool. Demand is represented by the number of "patient shifts" during which the nurse needs to provide the care. "Patient shifts" is simply a patient day multiplied by the number of shifts in a day. The nurse staffing level is represented by the number of "nurse shifts", which is the sum of the number of shifts worked by the nurses. Model notation is presented in Table 2.2.

We formulate this two-stage stochastic nonlinear programming problem as follows:

$$\text{(P1)} \quad \min F_1 = c_s n + \sum_{i=1}^{m} \left(\beta_i^{r-r_b}\right) \gamma_i \lambda c_i + Q(n, r) \tag{2.1}$$

$$\text{s.t. } \underline{r} \leq r \leq \overline{r} \tag{2.2}$$

$$n \geq 0 \tag{2.3}$$

The function $Q(n, r)$ is defined as:

$$Q(n, r) = E\left[Q\left(n, r, g(\omega)\right)\right] \tag{2.4}$$

The second stage (recourse) problem is:

$$\text{(P1-R)} \quad Q\left(n, r, g(\omega)\right) = \min c_A n_A \tag{2.5}$$

$$\text{s.t.} \quad n_A \geq \max\left[\frac{g(\omega)}{r} - n, 0\right] \tag{2.6}$$

The first term of the objective function (2.1) represents the wage costs corresponding to employing n nurse shifts in the unit. The second term in (2.1) measures the sum of the costs incurred by various adverse outcomes as a function of the patient-to-nurse ratio and uses odds ratios. Subscript i represents adverse outcome i (e.g., infection, patient fall, hospital-acquired pneumonia, et cetera) with odds ratio β_i, rate of occurrence γ_i, and cost parameter c_i. Assume the base level of the patient-to-nurse ratio is r_b = 6 and that from Kane et al. (2007b) the bloodstream infection odds ratio is β_i = 1.16. Then, for example, a patient-to-nurse ratio of r = 8 implies that the infection rate at r = 8 is $1.16^2 = 1.346$ times higher than the infection rate at the base level r = 6. The third term in (2.1) is the expected value of the second stage agency nursing costs as defined in (4) and Problem (P1-R). Constraint (2.2) sets upper and lower bounds on the decision variable r and constraint (2.3) enforces nonnegativity of n. The stage two recourse Problem (P1-R) determines the number of agency nurse shifts needed. It minimizes agency nurse costs (2.5) while ensuring in constraint (2.6) that enough agency nurses are used to achieve the target patient-to-nurse ratio.

The optimal solution to the recourse Problem (P1-R) can be written in closed form as follows:

$$n_A^* = \max\left[\frac{g(\omega)}{r} - n, 0\right] \tag{2.7}$$

$$Q\big(n, r, g(\omega)\big) = c_A \times \max\left[\frac{g(\omega)}{r} - n, 0\right] \tag{2.8}$$

By analyzing the Hessian of (2.1) with a general patient demand distribution Φ, we establish convexity properties of the objective function of Problem (P1), as stated in Proposition 2.1 below. Please refer to the Appendix A for proofs of Propositions.

PROPOSITION 2.1. *For $\beta_i \geq 1$ for all i, the objective function of Problem (P1) is convex with respect to n and r for a general distribution Φ of patient demand.*

As an illustrative example, assume that patient demand in the unit is uniformly distributed between $\underline{g}$ and $\bar{g}$. Then the objective function and first order conditions of Problem (P1) become:

$$F_1 = c_s n + \sum_{i=1}^{m} (\beta_i^{r-r_b}) \gamma_i \lambda c_i + \frac{c_A}{2r} \left(\frac{1}{\bar{g} - \underline{g}} \right) (\bar{g}^2 - n^2 r^2) - c_A n \left(\frac{1}{\bar{g} - \underline{g}} \right) (\bar{g} - nr)$$

$$\frac{\partial F_1}{\partial n} = c_s + \frac{c_A nr}{\bar{g} - \underline{g}} - \frac{c_A \bar{g}}{\bar{g} - \underline{g}} = c_s + \frac{c_A (nr - \bar{g})}{\bar{g} - \underline{g}} = 0$$

$$\frac{\partial F_1}{\partial r} = \frac{c_A (nr - \bar{g})(nr + \bar{g})}{2(\bar{g} - \underline{g}) r^2} + \sum_{i=1}^{m} (\beta_i^{r-r_b}) \gamma_i \lambda c_i \ln \beta_i = 0$$

Therefore, the optimal number of nurse shifts is:

$$n^* = \frac{c_A \bar{g} - c_s (\bar{g} - \underline{g})}{c_A r}$$

Note that agency nurses typically have higher wages than unit nurses (i.e., $c_A \geq c_s$), implying $n^* \geq 0$. Substituting n^* into F_1 and the first order condition $\frac{\partial F_1}{\partial r} = 0$ yields:

$$F_1 = \frac{2 c_s c_A \bar{g} - c_s^2 (\bar{g} - \underline{g})}{2 c_A r} + \sum_{i=1}^{m} (\beta_i^{r-r_b}) \gamma_i \lambda c_i$$

$$\frac{\partial F_1}{\partial r} = \frac{-2c_s c_A \bar{g} + c_s^2(\bar{g} - \underline{g})}{2c_A r^2} + \sum_{i=1}^{m} (\beta_i^{r-r_b}) \gamma_i \lambda c_i \ln \beta_i = 0$$

Although we cannot solve for the optimal r^* in closed form and must therefore use a numerical search method, we can determine an effective upper bound on r^* that would be helpful for hospitals in determining the optimal patient-to-nurse ratios.

PROPOSITION 2.2. *Assume that patient demand is uniformly distributed between* $\underline{g}$ *and* $\bar{g}$ *and that we require* $r \geq r_b$. *Then the optimal patient-to-nurse ratio* r^* *has an upper bound of* $\min\left(\max\left(\sqrt{\frac{2c_s c_A \bar{g} - c_s^2(\bar{g}-\underline{g})}{2c_A \sum_{i=1}^{m} \gamma_i \lambda c_i \ln \beta_i}}, r_b\right), \bar{r}\right)$.

2.3.2. Patient Length of Stay and Nurse Turnover

Problem (P1) can be used to analyze one or more adverse patient outcomes when setting patient-to-nurse ratios. Next we consider the effect of the patient-to-nurse ratio on two specific measures: patient length of stay and nurse turnover. The cumulative effect of the adverse outcomes will impact patient morbidity and patient mortality. Increased morbidity will lead to longer patient length of stay in the hospital and thus higher nursing and bed costs. Therefore, in this subsection we modify the previous analysis in three ways: (1) we use patient length of stay as the one quality outcome to capture the combined effect of individual adverse outcomes, (2) we incorporate the impact of patient-to-nurse ratios on length of stay and thus nurse wages and occupied bed costs, and (3) we incorporate the impact of patient-to-nurse ratios on nurse turnover. Once again, refer to Table 2.2 for model notation.

The two-stage stochastic nonlinear programming problem now becomes:

$$\text{(P2)} \quad \min F_2 = c_s n + (\beta_{LOS}^{r-r_b}) g c_{bed} + \left(\beta_{TO}^{r-r_b}\right) \rho \mu \left(\frac{n}{\sigma}\right) c_{TO} + \mathcal{Q}(n, r) \tag{2.9}$$

$$\text{s.t. } \underline{r} \leq r \leq \bar{r} \tag{2.10}$$

$$n \geq 0 \tag{2.11}$$

The function $\mathcal{Q}(n, r)$ is defined as:

$$\mathcal{Q}(n, r) = E\left[Q\left(n, r, g(\omega)\right)\right] \tag{2.12}$$

The second stage (recourse) problem is:

$$\text{(P2-R)} \qquad Q\left(n, r, g(\omega)\right) = \min c_A n_A \tag{2.13}$$

$$\text{s.t.} \qquad n_A \geq \max\left[\frac{\left(\beta_{LOS}^{r-r_b}\right) g(\omega)}{r} - n, 0\right]. \tag{2.14}$$

Length of stay impacts the total number of shifts during which a bed is occupied by a patient and thus the non-wage portion of occupied bed costs, as illustrated in the second term of (2.9). Patient-to-nurse ratios impact nurse turnover (Aiken et al. 2002), as shown in the third term of (2.9). Note that patient demand is impacted by the relationship between the patient-to-nurse ratio and patient length of stay, as can be seen in constraint (2.14). The optimal solution to the recourse Problem (P2-R) can be written in closed form as follows:

$$n_A^* = \max\left[\frac{(\beta_{LOS}^{r-r_b}) g(\omega)}{r} - n, 0\right] \tag{2.15}$$

$$Q\left(n, r, g(\omega)\right) = c_A \times \max\left[\frac{(\beta_{LOS}^{r-r_b}) g(\omega)}{r} - n, 0\right] \tag{2.16}$$

If patient demand is uniformly distributed between $(\beta_{LOS}^{r-r_b})\underline{g}$ and $(\beta_{LOS}^{r-r_b})\bar{g}$, we again solve for the number of nurses using first order conditions as follows. While

numerical experiments seem to suggest the objective function (2.9) of (P2) is convex, we have not been able to prove that this is true.

$$n^* = \frac{(\beta_{LOS}^{r-r_b})(\beta_{TO}^{-r_b})\left[\left(c_A\bar{g} - c_s(\bar{g}-\underline{g})\right)\sigma\beta_{TO}^{r_b} - c_{TO}(\bar{g}-\underline{g})\beta_{TO}^{r}\rho\mu\right]}{c_A r\sigma} \geq 0$$

2.4. Model with Average Patient Demand

In this section we consider simplified versions of Problems (P1) and (P2) where we use average patient demand rather than treating demand as a random variable. Then we compare the performance of the average demand models versus the random demand models. Similar to the presentation in previous section, we first consider only adverse patient outcomes, and then more specifically study the impact of patient to nurse ratio decisions on patient length of stay and nurse turnover.

2.4.1. Base Model

Once again, consider a typical medical/surgical unit in a hospital. The following model determines the optimal patient-to-nurse ratio r while incorporating the trade-off between nurse staffing costs and quality of care costs. As the patient-to-nurse ratio increases, fewer nurses are needed to care for the patients and staffing costs go down, but quality of care declines and the corresponding quality costs go up. Problem (P3) uses average patient demand g rather than treating demand as a random variable. Model notation is presented in Table 2.2.

$$\text{(P3)} \quad \min F_3 = \delta\left(\frac{g}{r}\right)c_s + \sum_{i=1}^{m}\left(\beta_i^{r-r_b}\right)\gamma_i\lambda c_i \tag{2.17}$$

$$\text{s.t. } \underline{r} \leq r \leq \bar{r} \tag{2.18}$$

The first term of objective function (2.17) represents the wage costs corresponding to the number of nurse shifts required to achieve a specific patient-to-nurse ratio. The parameter δ is an adjustment factor to account for uncertainty in staffing needs. The second term of (2.17) measures the sum of the costs incurred by various patient outcomes (subscript i) as a function of the patient-to-nurse ratio and makes use of odds ratios β_i. Constraint (2.18) sets lower and upper bounds on the decision variable r. While we cannot solve for the optimal r^* in closed form, Proposition 2.3 below verifies that the objective function of (P3) is convex with respect to r and thus first-order conditions can be used to identify the optimal patient-to-nurse ratio. Numerical results will be reported later in this section.

PROPOSITION 2.3. *Objective function (2.17) of Problem (P3) is convex with respect to* r.

2.4.2. Patient Length of Stay and Nurse Turnover

Similar to Section 2.3, we next analyze the patient-to-nurse ratio decision incorporating the impact of r on patient length of stay and nurse turnover. Problem (P4) uses average patient demand instead of treating demand as a random variable as in Problem (P2). Refer to Table 2.2 for notation.

$$\text{(P4)} \quad \min F_4 = \delta\left(\beta_{LOS}^{r-r_b}\right)\left(\frac{g}{r}\right)c_s + \left(\beta_{LOS}^{r-r_b}\right)gc_{bed} + \left(\beta_{TO}^{r-r_b}\right)\rho\mu\left(\frac{g}{r}\times\frac{1}{\sigma}\right)c_{TO} \quad (2.19)$$

$$\text{s.t. } \underline{r} \le r \le \overline{r} \quad (2.20)$$

Changes in the patient-to-nurse ratio impact patient length of stay (Kane et al. 2007b), which affects the number of nurses needed and thus nurse wage costs. This relationship is captured in the first term of objective function (2.19). Similar to the

models presented in the previous section, the second and third terms in (2.19) measure the non-wage occupied bed cost and nurse turnover cost as a function of the patient-to-nurse ratio. Proposition 2.4 verifies that the objective function of (P4) is convex with respect to r and thus first-order conditions can be used to identify the optimal patient-to-nurse ratio.

PROPOSITION 2.4. *Objective function (2.19) of Problem (P4) is convex with respect to* r.

2.4.3. Results

Here we report computational results with Problem (P4) to better understand how incorporating quality of patient care and nurse turnover will impact patient-to-nurse ratio decisions. As the results show, it is not always cost-efficient to implement higher patient-to-nurse ratios and use fewer nurses, thereby suggesting that providing higher quality of care can be a sound business decision. Also, at the end of this subsection, we compare the performance of Problem (P4), which uses average patient demand, with the more complex Problem (P2), which handles stochastic demand, and show that (P4) performs well with δ = 1. Therefore, we use $\delta = 1$ in the remainder of the chapter.

We use common representative numbers as reported in the literature for the following parameters: an increase of one patient per nurse is associated with a 23% increase in nurse burnout ($\beta_{TO} = 1.23$) (Aiken et al. 2002) and a 13% increase in patients' length of stay ($\beta_{LOS} = 1.13$) (Kane et al. 2007b). The value $\beta_{LOS} = 1.13$ is a rough approximation obtained from Kane et al. (2007b) as follows. They report that an increase of one registered nurse full time equivalent leads to a 31% reduction in patient length of

stay. Then as in Appendix F of the Kane et al. (2007a) AHRQ report, if we assume that one RN FTE/patient day = 8 RN hours/patient day, this implies patient/RN ratio = 24 hours/8 hours = 3 patients per nurse. Thus, if we assume an increase of one RN full-time equivalent per patient day is equivalent to a decrease of 3 in the patient-to-nurse ratio, then the odds ratio corresponding to an increase of one patient per nurse turns out to be 1.13. This obviously is a rough estimate since the precise conversion depends on the staffing level before the increase or decrease. The annual nurse turnover rate is assumed to be 20% ($\rho = 0.2$) (Kosel and Olivo 2002), and we use a conservative estimate of $30,000 for the cost of replacing a nurse ($c_{TO} = 30{,}000$) (Rothberg et al. 2005). We report our results in weekly total costs, and thus use the weekly nurse turnover rate converted from the annual rate. We assume each nurse works five shifts per week ($\sigma = 5$). Based on Rothberg et al. (2005), we use a nurse burnout to turnover conversion factor μ of 1. We use a non-wage patient length of stay cost c_{bed} of $100 per day. Because turnover costs, the burnout to turnover conversion factor, and non-wage length of stay costs are difficult to estimate, we also perform sensitivity analysis on these three parameters in a later section. We assume an average length of stay in the unit of 3 days, and the base level patient-to-nurse ratio is 6:1. For wage values, we use data collected from three hospitals for this study: one is located on the west coast of the U.S. and two are located in the Midwest. They are typical U.S. hospitals and range in size from 350-550 beds. We obtained information on nurse wages, shift types, staff size and mix, shift preferences and availability, demand levels, and patient-to-nurse ratios. All the figures in this chapter are based on results for one particular hospital of the three from which we gathered data. The results for the other two hospitals are almost identical in the shape of

the curve, with the only difference being the cost scale on the y-axis. Thus, the conclusions from the analyses below apply qualitatively to all three hospitals.

Figure 2.1 illustrates how total costs are impacted when patient-to-nurse ratio effects on patient length of stay (LOS) and nurse turnover are included as compared to many traditional operations management scheduling models that ignore these effects.

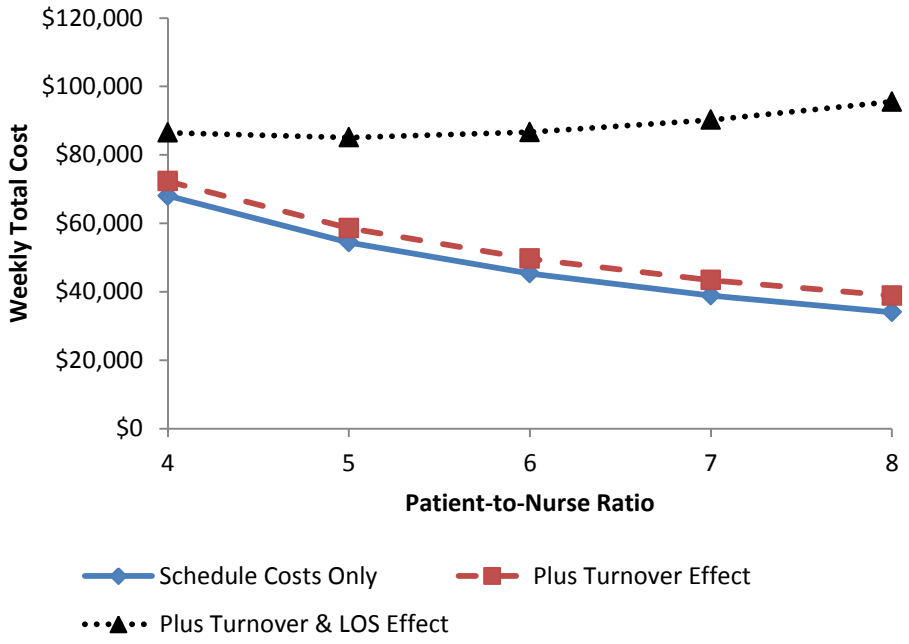


Figure 2.1. Total Cost When Including Patient-to-Nurse Ratio Effects on Nurse Turnover and Patient Length of Stay (LOS)

The "Schedule Costs Only" curve is measured by $\delta\left(\frac{g}{r}\right)c_s$ and ignores the patient-to-nurse ratio impact on nurse turnover and patient length of stay. This myopic approach yields an optimal ratio at its upper bound of 8. This result explains the desire of some hospital administrators to implement a high patient-to-nurse ratio when legislation requiring a certain ratio was first developed in California. Surprisingly, the result does not change when we add the effect of change in turnovers caused by different patient-to-

nurse ratios (i.e., add $\left(\beta_{TO}^{r-r_b}\right)\rho\mu\left(\frac{g}{r}\times\frac{1}{\sigma}\right)c_{TO}$ to the "Schedule Costs Only" curve to get the "Plus Turnover Effect" curve). Although the turnover rate would increase as the patient-to-nurse ratio becomes larger, the number of nurses scheduled per shift decreases, thereby essentially offsetting the impact of higher turnover. Finally, we add the effects of patient-to-nurse ratios on length of stay, represented by the "Plus Turnover & LOS Effects" curve. This curve is the total objective function value of Problem (P4). Figure 2.1 shows that the optimal ratio that minimizes the total costs for this surgical unit is 5:1. As the patient-to-nurse ratio becomes larger, the marginal benefit of employing fewer nurses decreases while the marginal cost of increasing length of stay rises. Different from when only turnover is considered, it is now possible that a lower patient-to-nurse ratio may not always require more nurses to be scheduled per shift, because better quality of care, which is represented by each nurse having to care for fewer patients, can decrease patients' length of stay.

Although one might argue that the cost improvement is not large enough to warrant a change in policy, we speculate that where the money is spent can make a big difference in how well the hospital can compete in the market. Figure 2.2 shows the breakdown of hospital spending at different patient-to-nurse ratios. Higher patient-to-nurse ratios allow hospitals to enjoy savings in nurse wages, but require them to spend those savings for dealing with adverse outcomes. In addition to providing fairly small yet meaningful benefit in total costs, lower patient-to-nurse ratios also allow hospitals to enjoy further intangible benefits including improvements in patient experiences, nurse satisfaction, and reputation that results in competitive advantages both in the market and in recruiting employees.

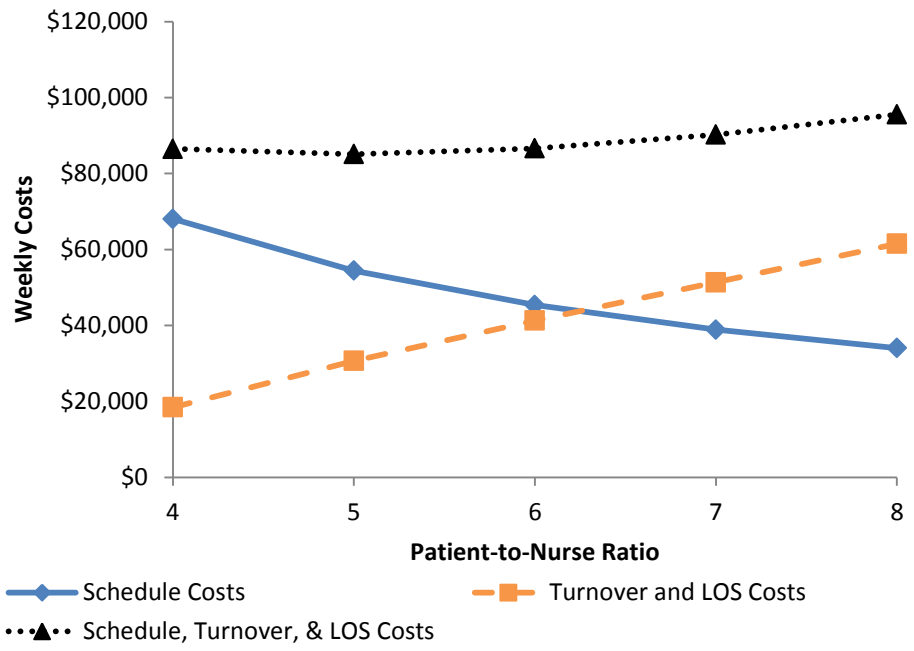


Figure 2.2. Breakdown of Total Costs into Schedule Costs and Turnover plus LOS Costs

Next we compare the performance of Problem (P4) with Problem (P2) for the case of uniformly distributed demand. Figure 2.3 reports the optimal objective values of (P2) relative to (P4). We also test different δ values in Problem (P4), which approximately accounts for uncertainty in staffing needs, to determine which gives the most similar results to Problem (P2). As illustrated in Figure 2.3, Problem (P4) performs very closely to Problem (P2) with δ = 1, always staying below a 5% difference in cost for reasonable ranges of the patient-to-nurse ratio. Because Problem (P4) is much simpler to use while providing results very similar to those from Problem (P2) with stochastic demand, it can be a very valuable tool for hospital management in making staffing and policy decisions. Thus, the remainder of the chapter presents results from Problem (P4) using $\delta = 1$.

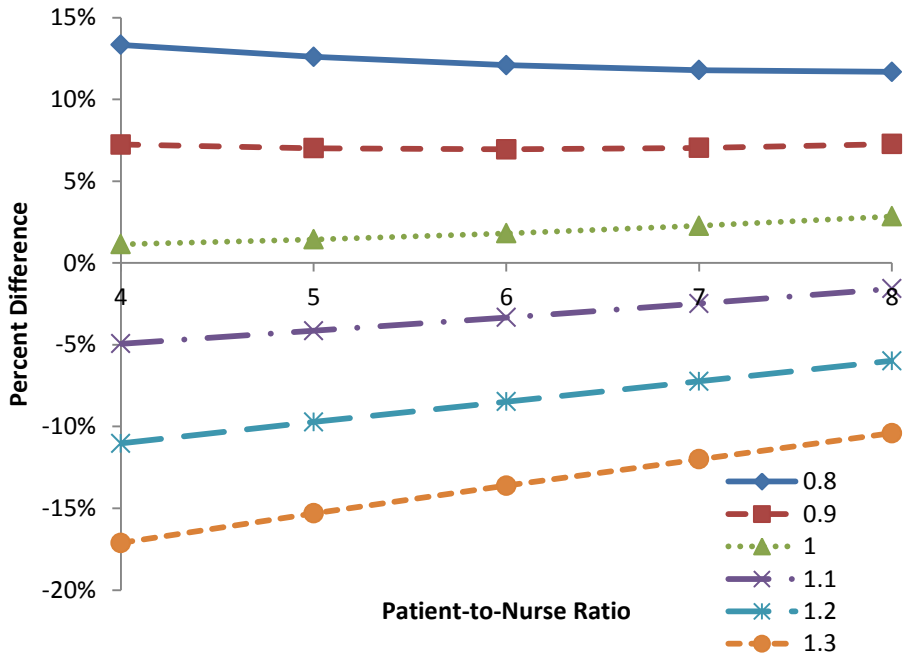


Figure 2.3. Comparison of Problem (P2) Relative to Problem (P4) for Various δ Values

2.5. Sensitivity Analysis

In this section, we perform sensitivity analysis to gain better insights into some of the parameters whose values are more difficult to estimate. We focus on Problem (P4) given its relative simplicity and previously reported accuracy. One key parameter that may vary is the cost per nurse turnover. We initially use $30,000 per turnover, which is a conservative figure according to various studies. Jones (2005) estimates turnover cost per registered nurse to be $62,100 - $67,100, whereas Kosel and Olivo (2002) state that it costs, on average, $46,000 to replace a medical/surgical nurse and about $64,000 to replace a critical care nurse. The Maryland Hospital Association (2000) estimates that it costs between $30,000 and $50,000 per registered nurse, and other estimates include the nurses' annual salary which was $64,690 in 2010 as reported by the Bureau of Labor Statistics, U.S. Department of Labor (2012). Thus, we report results for turnover cost per

nurse ranging between $20,000 and $60,000. Figure 2.4 shows that the results from solving Problem (P4) are robust with respect to the turnover cost per nurse.

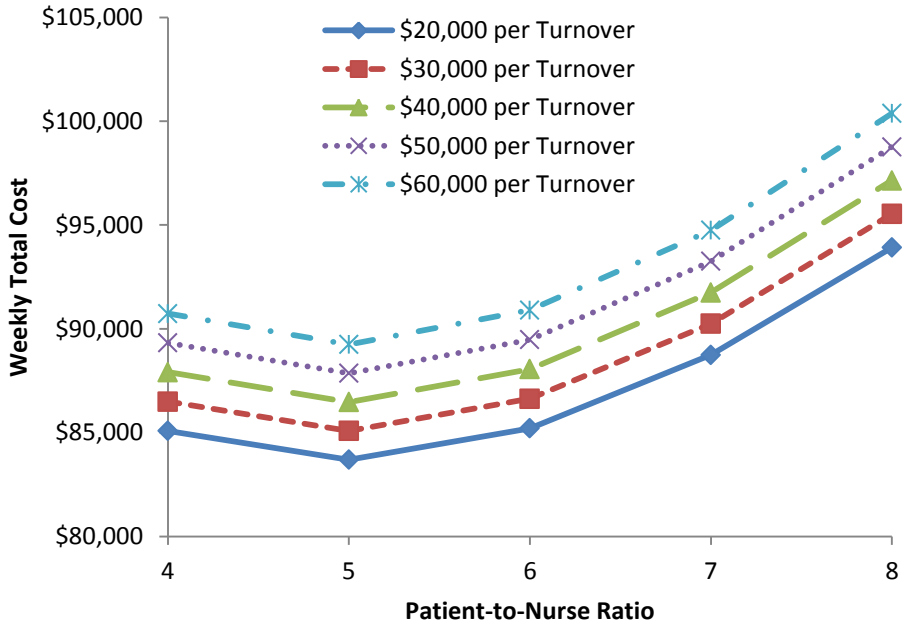


Figure 2.4. Total Costs for Different Nurse Turnover Cost Values

Next we consider the sensitivity of the results to the value of the odds ratio for patient length of stay. Figure 2.5 presents the impact of various patient length of stay odds ratio values (β_{LOS}) on total costs. Although the cost savings from a decrease in patient length of stay resulting from lowering the patient-to-nurse ratio is almost always significant at higher ranges of ratios (i.e. 7:1 or 8:1), it does not become high enough to offset the increase in staffing costs for the lower ranges of ratios (i.e. 4:1 or 5:1) until we use the odds ratio value of 1.13, which is the value we calculated from Kane et al. (2007b). However, any value greater than 1.13 would intensify the cost savings, thereby making the lower ratios cost-effective.

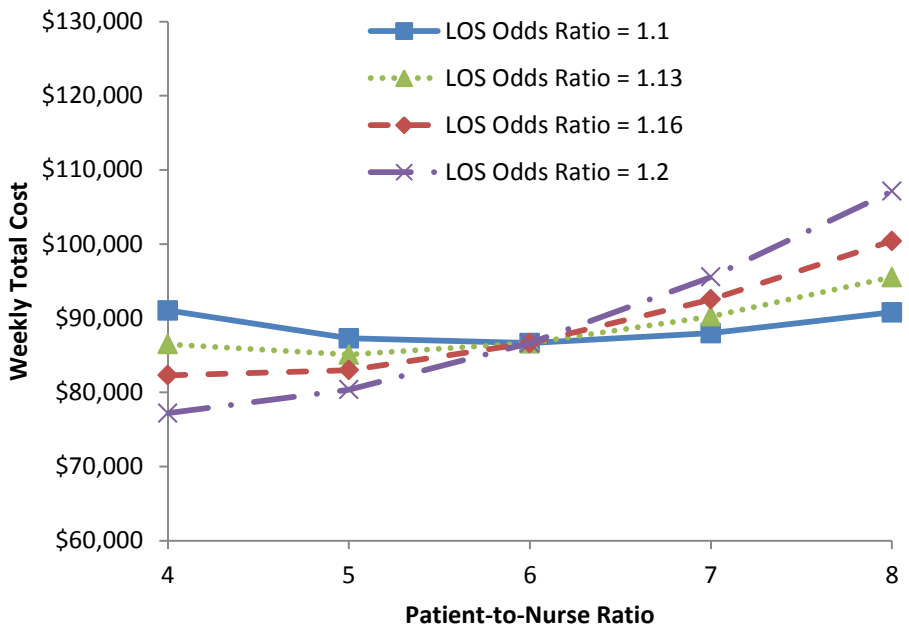


Figure 2.5. Total Costs for Various Length of Stay Odds Ratios (β_{LOS})

We also perform sensitivity analysis on the nurse burnout odds ratio, which is directly related to nurse turnover, and the nurse burnout to turnover conversion factor. It has been reported that an increase of one patient per nurse is associated with a 23% increase in nurse burnout ($\beta_{TO} = 1.23$) (Aiken et al. 2002), and this change in nurse burnout needs to be converted into change in nurse turnover. A conversion factor of 1 means a 10 percent increase in nurse burnout would lead to 10 percent increase in nurse turnover. We test the nurse burnout odds ratios from 1.1 to 1.4 and burnout to turnover (abbreviated as TO in Figure 2.7) conversion factor range from 0.6 to 1.2 to examine the robustness of the results regarding patient-to-nurse ratio and total costs. Figure 2.6 and Figure 2.7 show that the results presented in Section 2.4.3 are quite robust with regards to the various burnout odds ratios and conversion factor values.

We have used a conservative estimate of $100 per patient day for non-wage length of stay cost. Given the complications of identifying actual hospital costs versus what a hospital charges, and hospital variable costs versus fixed costs with respect to a change in patient length of stay, this is a difficult cost to estimate. For example, Rothberg et al. (2005) mention a patient length of stay cost of $1,000 per day which includes nurse wages, but it is difficult to know what exactly is included in this value and how the components are determined. Figure 2.8 shows the results for the non-wage length of stay cost varying between $100 and $500 per day.

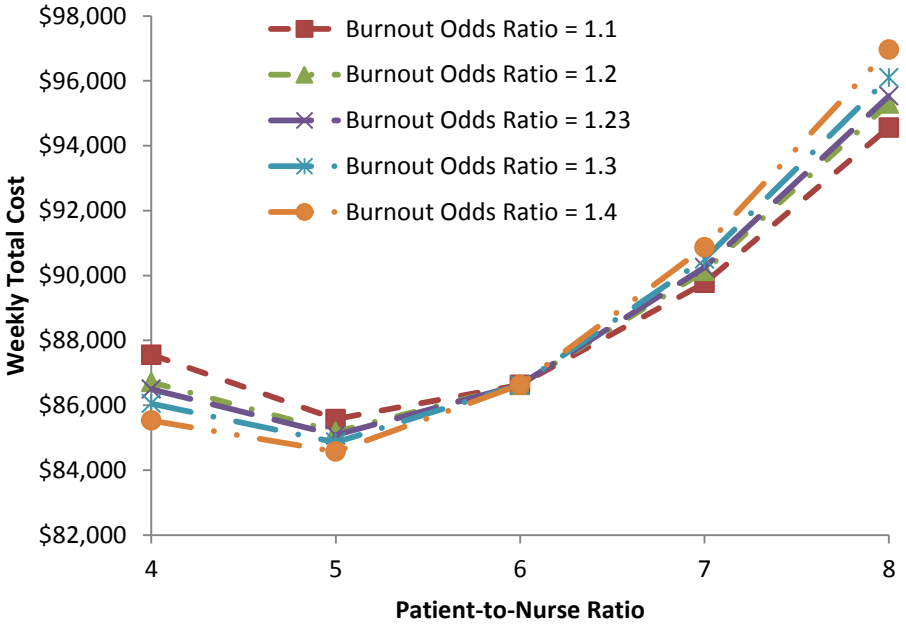


Figure 2.6. Total Costs for Various Burnout Odds Ratios

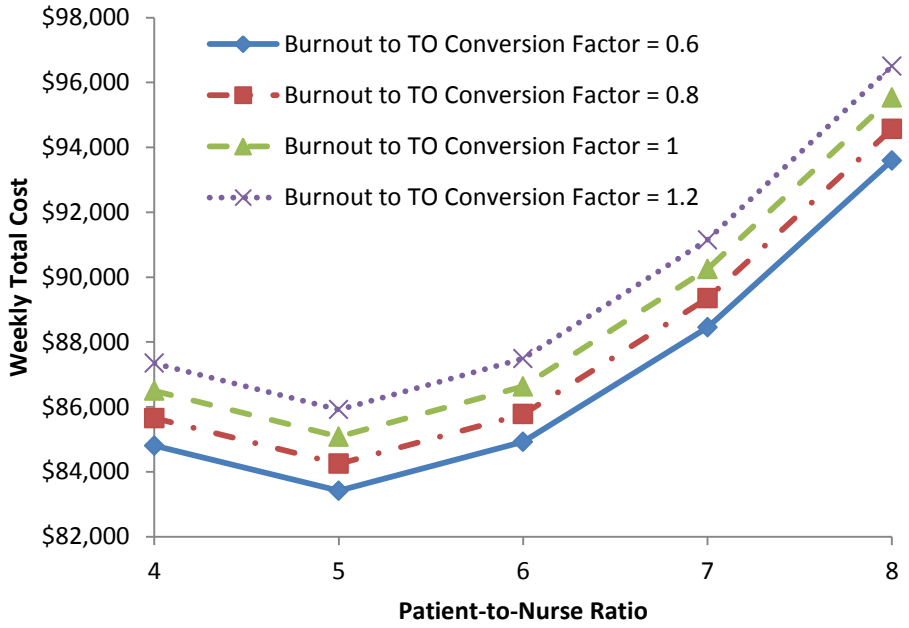


Figure 2.7. Total Costs for Various Nurse Burnout to Turnover (TO) Conversion Factors

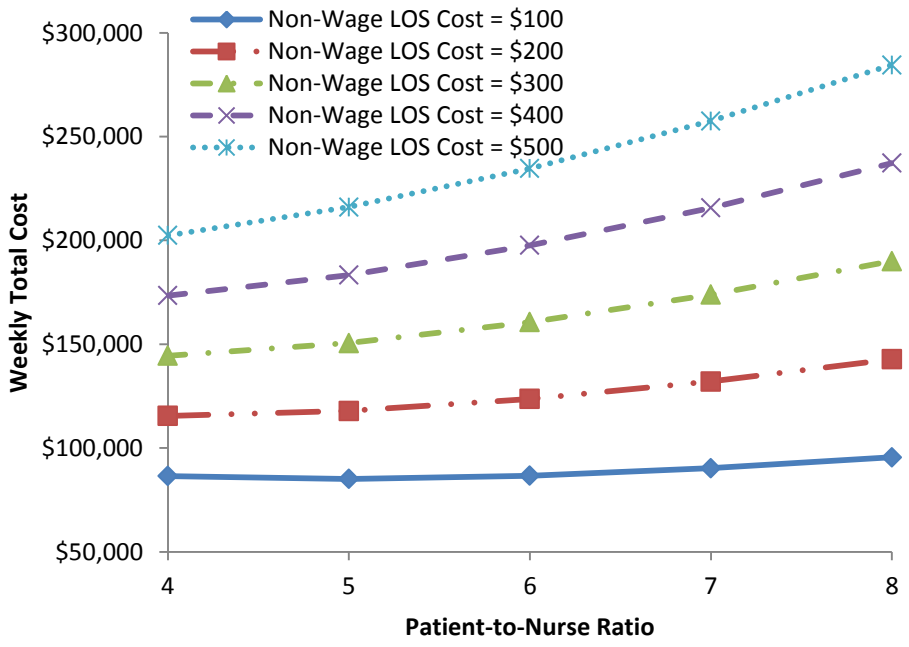


Figure 2.8. Total Costs for Various Non-Wage Occupied Bed Cost per Patient Day

2.6. Conclusions and Future Research

Hospital capacity planning is one of the difficult challenges facing the healthcare industry. In particular, the patient-to-nurse ratio decision has been extensively debated and studied by policy makers, hospital managers, and healthcare professionals. It has been widely assumed that lower patient-to-nurse ratios would increase the costs incurred by hospitals while improving the quality of patient care. Thus, the debate has been to find the optimal ratio that allows hospitals to provide adequate quality of care without incurring excessive costs. In addition, patient and nurse outcomes have been mostly ignored in traditional operations management approaches to hospital capacity planning. Consequently, the relevance of healthcare operations management decision models may be limited and potential cost savings from higher quality of care may have been neglected.

To address these issues, we develop nurse staffing models that incorporate adverse patient outcomes and nurse satisfaction, and show that lower nurse staffing ratios do not necessarily lead to higher costs. Our models show that lower staffing ratios can be more cost-effective even for conservative estimates of costs for patient length of stay and nurse turnover. Of course, the results will depend on the particular hospital setting under consideration. Our goal is to present a methodology for identifying optimal nurse staffing ratios for a given setting. Our numerical results suggest the need to incorporate medical and nurse outcomes into operations management models to correctly portray the tradeoffs faced by healthcare management.

While we focus on a few adverse outcomes such as patient length of stay and nurse turnover, there are many more outcomes that need to be considered. For example, it

would be interesting to investigate how including the effects of agency nurse usage, patient mix, and hospital volume on quality of care costs and nurse turnover would further change the optimal patient-to-nurse ratio. System shocks such as flu season and reduced capacity of new nurses from nurse turnover may also have impacts on the staffing decision. Moreover, this study can be extended by distinguishing among different hospital units according to their level of care. In addition, approaches other than higher nurse staffing levels could be studied for improving patient care, such as nurse-driven process improvement (Sims 2003). Such extensions would provide valuable insights to the healthcare industry to further improve its efficiency, financial health, societal impacts, and the ability to provide high quality patient care.

CHAPTER 3

BEHAVIOR-AWARE WORKFORCE STAFFING

Abstract

Empirical studies of service systems have demonstrated evidence of speedup and slowdown, which are defined as an increase and decrease, respectively, of service rate as a result of changes in workload. We model speedup and slowdown in a very general way to represent many possible joint effects of these behavioral phenomena. We use this model to study the impact of speedup and slowdown on a multi-period workforce staffing problem with recourse. We identify conditions under which the optimal workload, defined to be the ratio of requests to workers, is independent of the number of customer requests in the system, but we show that in general, a dynamic recourse policy is optimal. When a dynamic recourse policy is optimal, we show that the slowdown effect is strong enough to cause the firm to aggressively utilize expensive on-call workers to avoid future system congestion. This effect exists even in the presence of discounted future costs. Using a numerical study, we demonstrate that a one-step look-ahead policy performs very well and is a viable alternative when the optimal policy is not practical to compute.

3.1. INTRODUCTION

In many service systems, such as hospitals, call centers, and restaurants, service is provided by staff whose speed of work is affected by the amount of work assigned to each worker. Recent empirical work has shown that such servers frequently employ a changing service rate (either deliberately or as an unintended consequence of system congestion) depending on the workload at any given time (KC and Terwiesch 2012, Tan

and Netessine 2014). Specifically, two countervailing effects have been observed: *speedup*, where servers increase the service rate as the amount of work increases, and *slowdown*, where servers decrease the service rate as the amount of work increases. There are many possible reasons why these phenomena may occur. For example, speedup may occur when workers feel more motivated by a high demand for their service, while slowdown may occur when congestion in the service system due to high demand interferes with workers' ability to get their job done in a timely manner.

Regardless of the causal mechanisms behind speedup and slowdown, their effects on the performance of a service system should not be ignored. Speedup of servers such as waiters in a restaurant or nurses in a hospital can decrease the number of servers needed to provide a target level of service, but slowdown can increase the number of servers needed. Therefore, management decisions about staffing and scheduling should reflect an understanding of how workload and service rate are related. Traditionally, staffing decisions are made in advance and in the presence of *demand uncertainty*. Because the number of requests from customers (demand) is uncertain, staffing decisions must balance the possibility of understaffing (i.e., having too few workers to handle the demand, which leads to poor service quality or long service times) against the possibility of overstaffing (i.e., having too many workers, which leads to idle time or wasted staff resources). A schedule lead time is necessary because workers generally expect to be given their work schedules in advance: while workers prefer to be notified of their schedules as early as possible, a longer schedule lead time means more uncertainty in demand for the established schedule.

To mitigate the effects of demand uncertainty, many firms employ recourse actions to counter overstaffing and understaffing. Specifically, firms may depart from the established schedule and either obtain extra workers when understaffed or send workers home when overstaffed. Extra workers may be obtained through external employment agencies, cross-trained workers, or on-call workers. For simplicity, we will simply refer to these additional workers as "on-call workers". On-call workers often incur a wage premium, paid either to the worker to mitigate the inconvenience of being on call or to the agency in exchange for the staff procurement service. When overstaffed, the firm may be able to reduce the number of workers and recoup at least some proportion of their wages. For example, it may be possible to find some volunteers who are willing to leave before the shift ends, without being paid for the rest of their shift. Workers who are sent home involuntarily may receive pay for part of their shift. Employers in some states, such as California, are required to pay their employees for certain unworked but regularly scheduled time (California Department of Industrial Relations 2001). These recourse actions are illustrative and similar recourse actions are available in various forms in many different industries.

Staffing decisions, along with realized demand (i.e., the number of requests from customers), determine the *workload*, which we define to be the ratio of requests to workers. The impact of workload on the workers' service rate should be considered when making both initial scheduling and recourse decisions. At the beginning of each shift, as realized demand is compared to the established schedule, the incumbent workload may be characterized as "low", "moderate", or "high" when the staffing is "overstaffed", "adequately staffed" or "understaffed", respectively. In Table 3.1, we summarize the

recourse actions with respect to the workload level and the potential behavioral impacts that should be considered in the staffing decisions. Because the workload resulting from staffing decisions has an impact on workers' tendency to speedup or slowdown, considering these behaviors could lead to different staffing and recourse decisions compared to the case where these behaviors are ignored. By incorporating the workers' behavior into the overall staffing decision making process, firms can improve their workers' productivity and overall performance.

Workload	Staffing Status	Staffing Adjustments	Behavioral Issues to Consider
Low	Overstaffed	Send some workers home	• Speedup may increase the number of workers to send home • Slowdown may reduce the number of workers to send home
Moderate	Adequately Staffed	No adjustments	
High	Understaffed	Obtain additional workers	• Speedup may reduce the additional staffing needs • Slowdown may increase the additional staffing needs

Table 3.1. Recourse Staffing Decisions and Behavioral Issues to Consider

3.1.1. Incorporation of Speedup and Slowdown

Figure 3.1(a) and Figure 3.1(b) show examples of monotonic relationships between the workload and the workers' service rate. These S-shaped relationships represent the cases where only speedup is present (Figure 3.1a) and where only slowdown is present (Figure 3.1b). Figure 3.2 shows examples of non-monotonic relationships between the workload and the workers' service rate that are inverse U-shaped (if speedup occurs before

slowdown) and U-shaped (if slowdown occurs before speedup as workload increases), which can represent the joint effect of speedup and slowdown. Both U-shaped and inverse U-shaped relationships have been demonstrated in empirical studies (Batt and Terwiesch 2012, Jaeker and Tucker 2013, KC 2013, Tan and Netessine 2014). Intuitively, the nature of the relationship between workload and service rate should affect the magnitude of recourse actions taken and/or the quantitative interpretation of "low", "moderate", and "high" number of requests (in Table 3.1). We investigate the thresholds that separate these regions and their behavior with respect to changes in system parameters. We also study the staffing and recourse decisions when overstaffed and understaffed under the joint effects of speedup and slowdown. We characterize how the optimal workload changes with respect to the total amount of work that needs to be processed (number of requests in the system). In the "moderate" range where the system is adequately staffed, the workload is not constant in the number of requests (by definition since the number of workers is constant while the number of requests increases). We find that the workers' tendency to slowdown causes the optimal workload when overstaffed or understaffed to vary in number of requests. Lastly, we examine when a fixed request-to-server ratio policy, which is commonly employed in practice, should be used by firms and the potential benefits in switching to the dynamic policy.

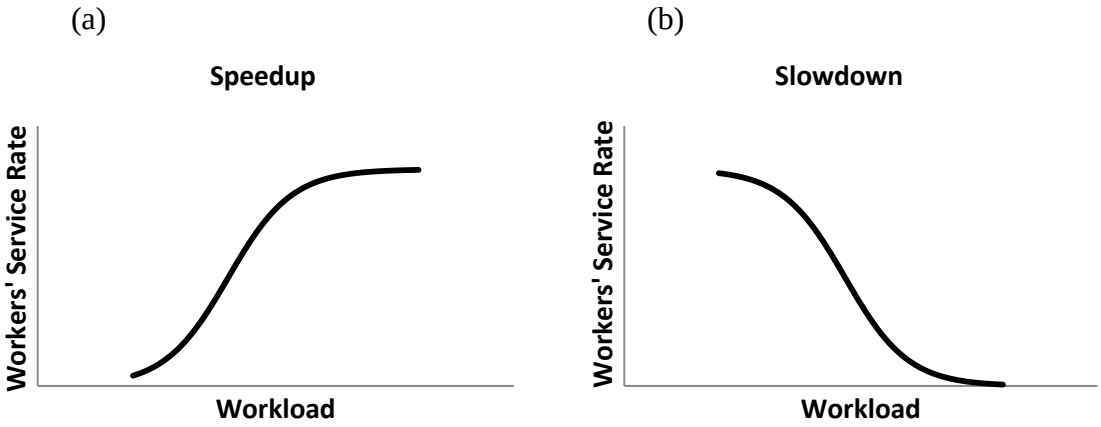


Figure 3.1. Functional Forms of Service Rate under (a) Speedup or (b) Slowdown Alone

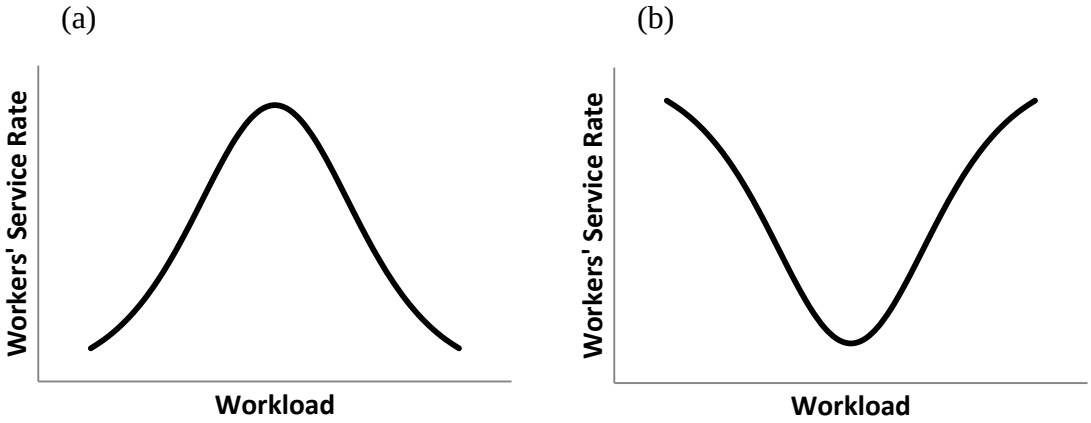


Figure 3.2. Functional Forms of Service Rate under Both Speedup and Slowdown. (a) Inverse U-Shape Due to Speedup Occurring First; (b) U-Shape Due to Slowdown Occurring First

3.1.2. Example: Hospital Nurse Staffing

While behavior-aware service staffing is applicable in many industries and settings, we will illustrate the phenomenon by focusing on the health care industry and we use hospital nurse staffing to demonstrate more clearly the staffing procedures and issues introduced in the previous section. The presence of workload measures such as patient-to-nurse (PTN) ratios and well-known recourse options such as "float" and agency nurses

make the healthcare industry a prime candidate for the consideration of behavioral issues in the staffing decision process. We also note that most nurses are scheduled in advance and on-call, agency, and cross-trained nurses require a wage premium.

At one of our partner hospitals in Indiana, a nurse manager creates a monthly advance schedule that dictates how many and which nurses are scheduled for each shift in an inpatient unit. Although the manager has some idea of the patient demand distribution, there are uncertainties involved in new patient arrivals and patient discharges for each shift. The hospital targets a PTN ratio of 5 patients per nurse. While Indiana does not have any legislation on laws requiring a specific PTN ratio, it is noteworthy that California state law prevents hospitals from exceeding the PTN ratio of 5:1 in medical/surgical units, and other states are considering similar legislation as well.

As each shift begins, the nurse manager on duty gains a much better understanding of the patient demand in the ward. The majority of the demand consists of patients who have already been admitted to the unit but have not yet been discharged. In addition, the manager anticipates new patient arrivals based on elective schedules and input from the emergency department and upstream units in the hospital (e.g., the Intensive Care Unit). Each day, there is a meeting between the nurse managers from each unit and the nurse manager of the float pool. They discuss the needs of each unit and the availability and skills of the float nurses, who represent the available pool of on-call workers for the nurse managers. During this meeting, the available float nurses are assigned to the units based on need.

One factor that the managers do not explicitly consider when making initial nurse schedule and recourse staffing decisions is the behavioral impact of the workload. Examples of speedup include anticipating test results and providing patient care proactively. Examples of slowdown include nurses taking longer to pick up test results because they are busy handling other duties, or nurses being interrupted more frequently by requests from doctors leading to more setups or other inefficiencies. If the effects of speedup and slowdown are taken into account, mangers may make different scheduling and recourse decisions than would otherwise be chosen. Note that nurse workload can affect patient outcomes, which impacts patient length of stay (i.e., service rate). Therefore, our modeling of service rate as a function of workload can also capture quality of care effects. For these reasons, the behavioral phenomena of speedup and slowdown need to be considered to make sure that optimal staffing decisions are made by the manager. In this essay, we do not attempt to alter the slowdown or speedup dynamic—rather, we note that its presence has been widely observed, and we show how to account for its reality in the staffing process.

The remainder of the chapter is organized as follows. Section 3.2 discusses literature relevant to this study. Section 3.3 presents the speedup/slowdown models and Markov decision process model. We begin by formulating a general problem for determining the optimal number of workers needed for each period and later consider special cases. Section 3.4 presents analytical results and Section 3.5 discusses numerical experiments and results. Section 3.6 provides some concluding remarks and opportunities for future research.

3.2. LITERATURE REVIEW

Before presenting our model to integrate speedup and slowdown effects into a staffing model with recourse, we review relevant literature in three streams: speedup and slowdown effects on service rate, staffing with recourse actions, and patient-to-nurse (PTN) ratios. The third stream of research is highly relevant because recent studies on PTN ratios combined with controversial legislation mandating maximum PTN ratios in some states provide a strong motivation for studying workload in the healthcare setting.

3.2.1. Speedup and Slowdown

Empirical studies suggest that in many operational settings, the service rate provided by workers depends on the workload. In Table 3.2, we summarize recent articles demonstrating speedup and/or slowdown in service and manufacturing settings. For each article, we list the type of operational setting, the independent variable representing the workload, and the dependent variable representing the effect of speedup and/or slowdown.

Author	Setting	Independent Variables	Dependent Variables	Results
Schultz et al. (1998)	Production lines (laboratory experiment)	Inventory level	Processing speed	Workers speed up when they are the cause of idle time on the line
Schultz et al. (1999)	Laboratory experiment	Inventory level	Processing speed	Slowest worker works faster under low-inventory than high-inventory systems
KC and Terwiesch (2009)	Hospital transportation; cardiac Intensive Care Unit (ICU) care	Workload	Length of stay, transport time	Increase in load reduces length of stay/transport time, but long periods of increase load decreases the service rate
Armony et al. (2015)	Emergency Department (ED)	Occupancy level	Service rate	Inverse U-shaped effect
Batt and Terwiesch (2012)	ED	Waiting room census	Service time (from when a patient is placed in a treatment bed to when treatment in the ED is complete	Inverse U-shaped effect
KC and Terwiesch (2012)	Intensive Care Unit	Occupancy level	Patient length of stay	16% shorter length of stay for a patient discharged from a busy ICU than that for a comparable patient discharged from a low-occupancy ICU
Jaeker and Tucker (2013)	Acute care hospitals	Occupancy level	Length of stay	As the hospital occupancy level of patients of the same type increases, LOS increases
			Probability of discharge	As same type inpatient workload increases towards around 85% occupancy, the probability of discharge increases by about 7%. Once workload exceeds approximately 85% occupancy, the probability of discharge decreases and LOS increases. As incoming patient load of the same type increases, the probability of discharge on the day before expected discharge also increases.

KC (2013)	ED	Level of physician multitasking	Total time taken to discharge a given number of patients	U-shaped response: initially reduces the time taken, but only up to a certain threshold level, after which it increases
			Throughput rate	Inverse U-shaped effect on throughput rate (the number of patients discharged per hour)
			Number of patient diagnoses	Inverse U-shaped effect
			Number of patient revisits	U-shaped effect
Tan and Netessine (2014)	Restaurant	Workload: number of tables or diners that a server simultaneously handles	Servers' performance: sales and meal duration	Workload has inverse-U-shaped relationship with meal duration
				Workload has inverse-U-shaped relationship with sales
Jaeker and Tucker (2015)	Acute care hospitals	Occupancy level	Length of stay	Patient length of stay increases as occupancy increases until a tipping point, when patients are discharged early to alleviate congestion. There is a second tipping point beyond which additional occupancy leads to a longer length of stay

Table 3.2. Empirical Studies Demonstrating Speedup and/or Slowdown in Service and Manufacturing Systems

Speedup has been shown to exist in many operational settings. In a manufacturing setting, Schultz et al. (1998) use a laboratory experiment to show that workers speed up when they are the cause of idle time on a production line. In the healthcare industry, KC and Terwiesch (2009) show that a 10% increase in system load reduces patient length of stay by two days for cardiothoracic surgery patients, and a 20% increase in the load for patient transporters reduces the transport time by 30 seconds. In each of these cases, workers speed up when the workload is high. KC and Terwiesch (2012) also find that the length of stay for patients in the ICU is influenced by the occupancy level of the unit. Their analysis shows that a patient is likely to be discharged early when the occupancy is high.

On the other hand, workers tend to slow down if the system is too congested, especially for systems with shared resources. Schultz et al. (1999) find from their experiments that the slowest workers in a production line work faster in low-inventory situations than in high-inventory situations. The joint effects of speedup and slowdown have been shown by a number of empirical studies as well. Armony et al. (2015) study detailed patient flow data from a large Israeli hospital and find that the service rate first increases and then decreases as the emergency department (ED) occupancy increases. Batt and Terwiesch (2012) conduct a detailed econometric analysis of an ED at a major U.S. hospital and identify and test for mechanisms that generate speedup and slowdown. They find that service time in a hospital ED first increases then decreases with workload. Jaeker and Tucker (2013) analyze inpatient data from California acute care hospitals and find that the probability of patient discharge increases until the occupancy level increases to approximately 85% and decreases with workload exceeding 85% occupancy. Tan and

Netessine (2014) analyze a data set from a restaurant chain and show that service speed first decreases with the increase in workload, but above a certain workload threshold, service speed increases with the further rise in workload. Jaeker and Tucker (2015) show from inpatient data of 203 California acute care hospitals that patient length of stay initially increases as occupancy increases until a tipping point, when patients are discharged early to alleviate congestion. They also find a second tipping point beyond which additional occupancy leads to a longer length of stay. These increases and decreases in service rate can also be explained using the context of multitasking. KC (2013) shows that the total time taken to discharge a given number of patients has a U-shaped response to the level of physician multitasking. Multitasking initially helps in improving productivity but it eventually overwhelms the worker once the level of multitasking passes a certain threshold.

3.2.2. Staffing with Recourse

Service enterprises in general and nursing in particular have been widely examined by researchers studying the problem of staffing with recourse. Hur et al. (2004) formulate a mathematical model and heuristics for the real-time schedule adjustment decisions for service manager, and they test the proposed procedures in the quick service restaurant industry. Mehrotra et al. (2010) develop a methodology to make real-time schedule adjustments in call center operations, and they test the model using data from an actual call center. The results show that the adjustments provide significant value when the call center is understaffed.

Nurse staffing with recourse has also been studied. Moz and Pato (2003, 2004, 2007) and Clark and Walker (2011) study the rerostering problem in the context of a public hospital, where modifications to the original schedule are necessary to replace nurses unable to work shifts that were previously assigned to them. Bard and Purnomo (2005) formulate an integer programming model for reactively scheduling nurses with daily adjustments and individual preference consideration. Punnakitikashem et al. (2008) develop a two-stage stochastic integer programming model with recourse for nurse assignment, in which each patient is assigned to a nurse at the beginning of a shift, with the goal of minimizing excess workload and developing balanced workloads for nurses while taking new admissions into account. In contrast, our work studies initial staffing and real-time adjustments in number of staff. Wright and Bretthauer (2010) study both internal recourse approaches (using overtime, float pool nurses, etc.) and external recourse approaches (using agency nurses) in combating nurse shortages while controlling the number of undesirable shifts. However, our essay incorporates the impact of the staffing decisions on the service rate, which is not considered in any of the papers listed above.

3.2.3. Patient-to-Nurse Ratios

Since the 1999 passage of California Assembly Bill 394 (AB394), which established minimum staffing levels for registered nurses in California hospitals, the patient-to-nurse (PTN) ratio has received significant attention in the literature. See Spetz (2004) for an examination of the early implementation of fixed staffing ratios in acute-care hospitals in California and an overview of the debate on the issue. Aiken et al. (2002) analyze the association between the PTN ratio and patient mortality, failure-to-rescue among surgical

patients, and factors related to nurse retention. Their analysis shows that each additional patient per nurse is associated with a 7% increase in patient mortality and 23% increase in the odds of nurse burnout. Needleman et al. (2006) study the benefit of increasing hospital nurse staffing and find that increasing total nursing hours leads to a decrease in patient length-of-stay, adverse outcomes, and patient deaths. Kane et al. (2007b) examine the association between nurse staffing and patient outcomes in acute care hospitals and find increased staffing to be associated with lower mortality, lower length of stay, and decreased odds of adverse medical outcomes. Aiken et al. (2011) show that the effect of decreasing workloads by one patient per nurse has virtually no impact on deaths and failure-to-rescue in hospitals with poor work environments, but decreases the odds on both deaths and failures in hospitals with average and good environments. Cook et al. (2012) evaluate the impact of AB394 and find evidence that AB394 had the intended effect of decreasing PTN ratio in hospitals, but they do not find the improved PTN ratio to be associated with improvements in measured patient safety. Conversely, Cimiotti et al. (2012) find significant association between PTN ratio and urinary tract infection and surgical site infection. McHugh et al. (2013) examine the relationship between nurse staffing and readmission penalties and find that hospitals with higher nurse staffing had 25 percent lower odds of being penalized compared to otherwise similar hospitals with lower staffing. We contribute to this stream of literature by providing valuable insights on the optimal workload and staffing levels in service industries and study how the decisions change when behavioral phenomena of speedup and slowdown are incorporated.

3.3. MODEL

We incorporate the impact of speedup and slowdown on the multi-period workforce staffing problem with recourse. As we noted in Section 3.2, empirical studies have demonstrated evidence of speedup and slowdown, but to the best of our knowledge there is no existing literature on how to model the effects of speedup and slowdown on service rate and incorporate these effects into a staffing decision. In Section 3.3.1, we describe functional forms to model speedup and slowdown separately, then combine the two functions in a very general way to represent many possible joint effects of speedup and slowdown. In Section 3.3.2, we formulate a stylized staffing model, which is an infinite-horizon stochastic dynamic program for multi-period staffing with recourse. In our analysis in Section 3.4 and subsequent numerical studies in Section 3.5, we will use these two models together to obtain results and insights on the multi-period staffing problem. A summary of notation is given in Appendix B.1 and proofs are presented in Appendix B.3.

3.3.1. Speedup and Slowdown

Depending on the operational setting, either speedup or slowdown may be present in isolation or they may have joint effects on the service rate. Even when only one effect is *explicitly* present, both effects may actually be present, with one dominating the other. For example, speedup due to high workload may unintentionally lead to reduced quality requiring correction or rework and causing slowdown, but not at a sufficient level to offset the speedup. Of course, if this effect increases sufficiently, the time needed for correction or rework may result in slowdown in overall service rate of the system. In order to model these joint effects in a very general way, we first model speedup and

slowdown separately and then use a convex combination of the two functions to represent the joint effects. Denote the workload, or *ratio* of requests to servers, as $r \in (0, \bar{r}]$, where $\bar{r}$ is the maximum allowable ratio. Then, define

$$\text{(Speedup)} \quad \gamma(r) = \frac{\bar{\mu}_{sp}}{1 + e^{-\theta_{sp}(r - r_{sp})}}, \text{ and} \tag{3.1}$$

$$\text{(Slowdown)} \quad \tau(r) = \frac{\bar{\mu}_{sl}}{1 + e^{\theta_{sl}(r - r_{sl})}}, \tag{3.2}$$

where $\bar{\mu}_{sp}$ and $\bar{\mu}_{sl}$ are scale parameters representing the upper bounds of the service rate that a worker can achieve, θ_{sp} and θ_{sl} are shape parameters representing the degree of speedup and slowdown effects, and r_{sp} and r_{sl} are location parameters representing the positions of the speedup and slowdown curves with respect to the workload (they are the levels of workload where a service rate of $\bar{\mu}_{sp}/2$ and $\bar{\mu}_{sl}/2$ are achieved). The joint effect of speedup and slowdown is modeled as a convex combination of $\gamma(r)$ and $\tau(r)$, which is written as

$$\mu(r) = \beta \cdot \gamma(r) + (1 - \beta) \cdot \tau(r) \tag{3.3}$$

where $\beta \in [0,1]$. When $\beta = 0$, only slowdown is present, and when $\beta = 1$, only speedup is present. As workload increases, speedup occurs before slowdown if $r_{sp} < r_{sl}$ and slowdown occurs before speedup if $r_{sp} > r_{sl}$.

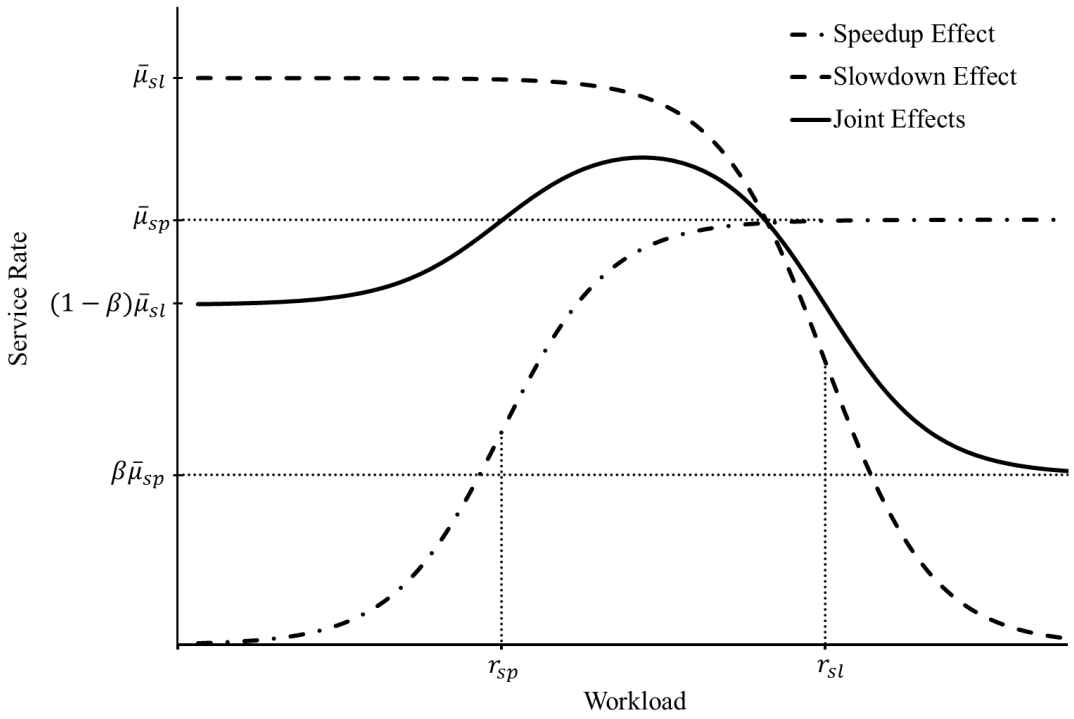


Figure 3.3. Modeling the Joint Effects of Speedup and Slowdown Using Logistic Functions

Using (3.3), we can model different relationships between the workload and the service rate that have been observed in various empirical studies. Figure 3.3 shows an example of $\gamma(r)$, $\tau(r)$, and $\mu(r)$ representing an inverse U-shaped relationship between the workload and the service rate. For analytical tractability, we assume that $\theta_{sp} = \theta_{sl} = \theta$ and consider the resulting four cases from the $\mu(r)$ function: monotonically increasing, monotonically decreasing, inverse U-shaped, and U-shaped. Specifically, when the impact of one behavioral phenomenon is much greater than that of the other and the joint effect mainly reflects the dominant phenomenon, the function may be either monotonically increasing or decreasing, even when β is neither zero nor one. Figure 3.1(a) and Figure 3.1(b) show examples of cases in which speedup and slowdown

dominate the other phenomenon, respectively. We formally define these two cases, and Lemmas 3.1 and 3.2 characterize sufficient conditions for the two cases.

DEFINITION 3.1. *Speedup "dominates" slowdown when* $\mu(r)$ *is nondecreasing for* $r \in (0, \bar{r}]$.

DEFINITION 3.2. *Slowdown "dominates" speedup when* $\mu(r)$ *is nonincreasing for* $r \in (0, \bar{r}]$.

LEMMA 3.1. *Suppose* $\theta_{sp} = \theta_{sl} = \theta$ *and* $\frac{(1-\beta)}{\beta} < \frac{\left(e^{\theta r_{sl}} + e^{\theta r}\right)^2 \bar{\mu}_{sp}}{\left(e^{\theta r_{sp}} + e^{\theta r}\right)^2 e^{\theta(r_{sl} - r_{sp})} \bar{\mu}_{sl}}$ *for* $0 < r < \bar{r}$.

Then $\mu(r)$ *is monotonically increasing and the impact of speedup dominates that of slowdown.*

LEMMA 3.2. *Suppose* $\theta_{sp} = \theta_{sl} = \theta$ *and* $\frac{(1-\beta)}{\beta} > \frac{\left(e^{\theta r_{sl}} + e^{\theta r}\right)^2 \bar{\mu}_{sp}}{\left(e^{\theta r_{sp}} + e^{\theta r}\right)^2 e^{\theta(r_{sl} - r_{sp})} \bar{\mu}_{sl}}$ *for* $0 < r < \bar{r}$.

Then $\mu(r)$ *is monotonically decreasing and the impact of slowdown dominates that of speedup.*

The other two cases arise when neither behavioral phenomenon dominates the other. In particular, when $\theta_{sp} = \theta_{sl} = \theta$, we have either inverse U-shaped or U-shaped relationships between the workload and the service rate depending on which phenomenon occurs first, as shown in Figure 3.2(a) and Figure 3.2(b). We define and present sufficient conditions for these two cases.

DEFINITION 3.3. $\mu(r)$ *is "inverse U-shaped" if* $\mu(r)$ *is not monotonic and there exists* r_0 *such that* $\mu(r)$ *is nondecreasing for* $0 < r \leq r_0$ *and nonincreasing for* $r_0 \leq r \leq \bar{r}$.

DEFINITION 3.4. $\mu(r)$ *is "U-shaped" if* $\mu(r)$ *is not monotonic and there exists* r_0 *such that* $\mu(r)$ *is nonincreasing for* $0 < r \leq r_0$ *and nondecreasing for* $r_0 \leq r \leq \bar{r}$.

LEMMA 3.3. *If* $\theta_{sp} = \theta_{sl} = \theta$ *and* $\mu(r)$ *is not monotonic,* $\mu(r)$ *is inverse U-shaped when* $r_{sp} < r_{sl}$ *and U-shaped when* $r_{sp} > r_{sl}$.

Lemma 3.3 shows that there are four cases to be considered; when $r_{sp} = r_{sl}$ the service rate function is quasilinear, in which case one behavioral effect dominates the other (see Definitions 3.1 and 3.2).

3.3.2. Markov Decision Process Model

We model the workforce staffing recourse decision as an infinite-horizon discounted Markov decision process (MDP) where $\alpha \in (0,1)$ is the one-shift discount rate. The firm makes an initial decision, which is an *a priori* staffing schedule for each shift based on forecasted demand. At each decision epoch (i.e., the beginning of a shift), the *state* is the number of requests in the system. The state space is the set of non-negative integers. The firm's *action* is a target workload (i.e., the request-to-worker ratio). Thus, the action space is the set of positive reals. Although in practice the number of workers must also be an integer, we relax this requirement in the stylized model and are interested in understanding the firm's "ideal" workload. That is, an initial number of workers is assigned at the beginning of the planning horizon to each shift and subsequently the manager makes real-time schedule adjustments to achieve the ideal workload (ratio) based on the number of requests at the beginning of each shift. In particular, if at the beginning of a shift the workload is higher than the optimal level, the manager schedules additional workers at a premium, for example by calling in on-call workers. Although the

wages for these on-call workers are typically higher than those for the regular workers, in this case the reduction of future variable costs and backorder costs (associated with the number of service requests in the system, and elaborated on below) may outweigh the extra immediate costs incurred for these workers. On the other hand, if the unit has too many workers scheduled in a given shift, the manager sends some workers home. In this case, the firm may still have to pay some proportion of the wages for the workers sent home.

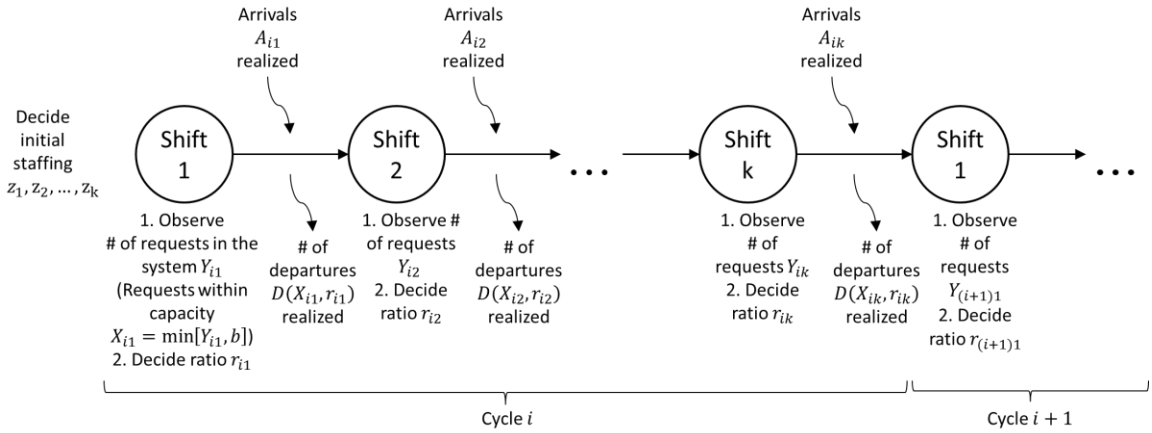


Figure 3.4. Timeline of Decisions and Events

We consider an infinite time horizon with a cyclic arrival pattern consisting of k shifts in each cycle (a cycle might represent a week, for example). Figure 3.4 presents the timeline of decisions and events in a cycle. *A priori*, the firm assigns z_s workers to work in shift s for $s = 1, \ldots, k$ in each cycle. While the realized number of arrivals and departures in a given shift can be different from those in a corresponding shift of a different cycle, the distributions of arrivals and departures are independent of the cycle number. Consequently, z_s is independent of the cycle number as well. Each worker receives a wage of c_w per shift, and the system incurs a variable cost c_v per shift for each

request in service. If the number of requests exceeds the service capacity of the system b, the excess requests are backlogged and the system incurs a holding cost c_h per backlogged unit per shift.

At the beginning of shift s of cycle i, the manager observes the number of service requests Y_{is} and the number of requests that are placed in service (i.e., not backlogged) is X_{is}, which is determined as $\min\{Y_{is}, b\}$. The manager decides on the workload, which is the ratio of requests in service to workers, denoted $r_{is} \in (0, \bar{r}]$ by adjusting the original workforce z_s as needed. She has an ability to obtain extra workers at the per-worker wage of $(1 + \psi)c_w$, where $\psi \geq 0$ is the on-call premium relative to the regular worker wage. She also can send some workers home and recoup a proportion of their wages, denoted $\phi \in [0,1]$. If $\phi = 0$, the firm still has to pay the full wage for the workers sent home. If $\phi = 1$, the firm is able to recoup the entire wage for the workers sent home. After the adjustments are made, the workload for the shift is set at r_{is}. The service rate (which is a function of r_{is}) determines how many requests will depart the system by the end of the shift and depends on the presence of speedup and/or slowdown. In order to satisfy the Markovian requirement of MDP, we assume an independent and identically distributed departure probability $\mu(r_{is})$ for each request in service. Thus, the number of requests that depart the system at the end of a given shift, denoted as $D(X_{is}, r_{is})$, follows the binomial distribution with X_{is} trials and success probability $\mu(r_{is})$. This formulation of $\mu(r_{is})$ as a probability requires that the average length of stay to be at least one shift (for industries where the average length of stay is shorter than a shift, we can further subdivide the shift into intervals of time small enough so that the requests would stay in the system for at least one time period on average). Finally, there are new arrivals of requests A_{is}, which

follows a general distribution with $E[A_{is}] = \lambda_s$. The number of requests for the subsequent shift is the convolution of a new arrival of requests in the shift A_{is} and requests still remaining in the system after the previous shift $Y_{is} - D(X_{is}, r_{is})$.

Let $V_s(y)$ denote the minimum expected total discounted cost at the beginning of shift s if y requests are in the system. We formulate the following infinite-horizon stochastic program, whose optimal objective value will be $V_s(y)$.

$$V_s(y) = \min_{\mathrm{r_{ij}}} E\left[\sum_{j=s}^{k} \alpha^{j-s}\left(c_w z_j + c_v X_{1j} + c_h\left(Y_{1j} - X_{1j}\right) + (1+\psi)c_w\left(\frac{X_{1j}}{r_{1j}} - z_j\right)^+\right.\right.$$

$$\left. - \phi c_w\left(z_j - \frac{X_{1j}}{r_{1j}}\right)^+\right) + \sum_{i=2}^{\infty}\sum_{j=1}^{k} \alpha^{(i-1)k+j-s}\left(c_w z_j + c_v X_{ij} + c_h\left(Y_{ij} - X_{ij}\right)\right.$$

$$\left.\left. + (1+\psi)c_w\left(\frac{X_{ij}}{r_{ij}} - z_j\right)^+ - \phi c_w\left(z_j - \frac{X_{ij}}{r_{ij}}\right)^+\right)\right] \tag{3.4}$$

subject to:

$$Y_{1s} = y \tag{3.5}$$

$$Y_{i1} = Y_{(i-1)k} + A_{(i-1)k} - D\left(X_{(i-1)k}, r_{(i-1)k}\right) \text{ for } i = 2, 3, \ldots \tag{3.6}$$

$$Y_{ij} = Y_{i(j-1)} + A_{i(j-1)} - D\left(X_{i(j-1)}, r_{i(j-1)}\right) \text{ for } i = 1, 2, \ldots; j = 2, 3, \ldots, k \tag{3.7}$$

$$X_{ij} = \min\{Y_{ij}, b\} \text{ for } i = 1, 2, \ldots; j = 1, 2, \ldots, k \tag{3.8}$$

$$r_{ij} \in (0, \bar{r}] \text{ for } i = 1, 2, \ldots; j = 1, 2, \ldots, k \tag{3.9}$$

In the above formulation, the decision variables are r_{ij}, the ratio of requests to servers in shift j of the ith cycle. Equivalently, we can write $V_s(y)$ recursively as the solution to a set of k Bellman's equations as follows. Note that the i subscript is eliminated because

decisions are independent of past cycles and only dependent on the current number of requests Y (i.e., the model has the Markov property).

$$V_s(y) = \min_{r\in(0,\bar{r}]}\{v_s(y,r)\}, \tag{3.10}$$

where

$$v_s(y,r) = c_w z_s + c_v \min\{y,b\} + c_h(y-b)^+ + (1+\psi)c_w\left(\frac{\min\{y,b\}}{r} - z_s\right)^+ - \phi c_w\left(z_s - \frac{\min\{y,b\}}{r}\right)^+ + \alpha E\left[V_{s^{++}}\left(y + A_s - D(\min\{y,b\},r)\right)\right] \tag{3.11}$$

$$s^{++} = \begin{cases} s+1 & \text{if } s < k \\ 1 & \text{if } s = k \end{cases} \tag{3.12}$$

3.4. ANALYSIS

In this section, we develop analytical insights into optimal staffing policies for firms when speedup and slowdown are taken into account. Specifically, Section 3.4.1 presents closed form solutions for the case when speedup dominates slowdown or the case of a U-shaped curve when $\mu(\bar{r}) \geq \mu(r)$ for all $r \in (0,\bar{r}]$. Section 3.4.2 presents results for the case where slowdown dominates speedup or the case of an inverse U-shape curve. While closed form solutions are not possible for the cases analyzed in Section 3.4.2, we are able to show that the optimal workload will always be in the slowdown region. Section 3.4.3 allows any of the four cases (speedup dominates, slowdown dominates, U-shape, and inverse U-shape) while making additional assumptions on the problem structure. In this case, we are able to show that the optimal workload is independent of customer census. In Section 3.5, we present numerical experiments to further analyze optimal staffing decisions for the cases where there is not a closed form solution.

3.4.1. "Speedup Dominates" and "U-Shape" Cases

We first consider cases where the speedup effect dominates the slowdown effect (Figure 3.1a) or slowdown occurs before speedup (U-Shape, Figure 3.2b) as defined in Section 3.3.1. We present the optimal staffing policy for both cases.

PROPOSITION 3.1. *When speedup dominates slowdown, or when $\mu(r)$ is U-shaped and $\mu(\bar{r}) \geq \mu(r)$ for all $r \in (0, \bar{r}]$, the optimal decision is to always keep the workload at $\bar{r}$ and staff the minimum number of workers allowed. The firm's decision is independent of the number of requests in the system y.*

In both cases, the firm would maximize the speedup effect and workers' productivity by increasing the workload as much as possible, resulting in minimum staffing costs. For industries in which there are limits on how much workload can be assigned to workers (e.g., for safety reasons), Proposition 3.1 suggests that a fixed workload policy (at $\bar{r}$) is optimal for the speedup-dominates case, independent of the amount of work present in the system. For the U-shaped case, a fixed workload policy is optimal if the right tail of the U-shaped case is sufficiently high such that $\mu(\bar{r}) \geq \mu(r)$ for all $r \in (0, \bar{r}]$. When $\mu(\bar{r}) < \mu(r)$, the optimal workload is either $\bar{r}$ or the optimal solution for the region of the U-shape curve in which slowdown dominates speedup, which we discuss further in the next subsection.

3.4.2. "Slowdown Dominates" and "Inverse U-Shape" Cases

In this subsection, we are interested in cases where the impact of slowdown dominates that of speedup (Figure 3.1b) or speedup occurs before slowdown (Inverse U-Shape, Figure 3.2a). Note that in both cases there is a level of workload $r_0 < \bar{r}$ that maximizes

the service rate. For workloads higher than r_0, the service rate decreases with the workload. Furthermore, for the "Inverse U-shape" case, the service rate increases with the workload for workload lower than r_0. We observe that for both cases, the optimal workload would always be at least the workload level that maximizes the service rate. The following proposition formalizes the above discussion.

PROPOSITION 3.2. *Let $r_0 \leq \bar{r}$ be a global maximizer for $\mu(r_s)$. Then when slowdown dominates speedup or $\mu(r)$ is inverse U-shape, the optimal workload $r_s^*(y) \geq r_0$.*

Later in the chapter, we show examples both analytically and numerically where $r_s^* > r_0$ and thus the optimal workload does not maximize the departure rate. It follows from Proposition 3.2 that the optimal workload would always be in a *slowdown region*, where the service rate decreases with the workload. For every level of workload in a *speedup region*, where the service rate increases with the workload, the firm can find a workload level in a slowdown region that achieves the same or better service rate with fewer workers.

3.4.3. Shift-by-Shift Staffing with No Capacity Constraint

While we know from Proposition 3.2 that the optimal workload will always be in a slowdown region, fully characterizing the optimal solution beyond Proposition 3.2 is not possible, and in Section 3.5 we provide a numerical analysis. In this subsection we gain more insights on the optimal staffing decision by analyzing a special case in which extra on-call workers do not require any premium ($\psi = 0$) and the firm can recoup the entire wage for workers sent home ($\phi = 1$). We call this case "shift-by-shift staffing" since this is equivalent to a situation in which the manager has the ability to assign workers at the

beginning of every shift without any penalty and thus the predetermined number of workers is irrelevant. We also assume that there is no capacity constraint.

Under these simplifying assumptions, the optimality equation is given by

$$V_s(y) = \min_{r\in(0,\bar{r}]}\{v_s(y,r)\}, \tag{3.13}$$

where

$$v_s(y,r) = c_w\frac{y}{r} + c_v y + \alpha E\left[V_{s^{++}}\left(y + A_s - D(y,r)\right)\right] \tag{3.14}$$

Denote by $r_s{}^*(y)$ the optimal workload to assign to each worker in shift s when there are y requests in the system. In other words, $r_s{}^*(y) = \arg\min_{r\in(0,\bar{r}]}\{v_s(y,r)\}$. In the following, we show that in this special case, the optimal workload does not change with y. That is, there is a ratio (r_s^*) that is optimal independent of the number of customer requests and at the start of each shift, staffing is set to achieve this ratio.

PROPOSITION 3.3. *Suppose that* $\psi = 0$, $\phi = 1$, *and* $b = \infty$. *Then for each* $s \in \{1, 2, \dots, k\}$, *there exists ratio* r_s^* *such that* $r_s{}^*(y) = r_s^*\ \forall y$.

One consequence of Proposition 3.3 is that $V_s(y)$ is a linear function. Knowing this, we can solve for an explicit expression for $V_s(y)$. To do so, note that there must exist reals ξ_s, η_s such that $V_s(y) = \xi_s y + \eta_s$. Using the definition of $V_s(y)$, we obtain

$$\xi_s y + \eta_s = \left(\frac{c_w}{r_s^*} + c_v\right)y + \alpha\left(\xi_{s^{++}}\left(y\left(1-\mu(r_s^*)\right)\right) + \xi_{s^{++}}\lambda_s + \eta_s\right), \tag{3.15}$$

from which we can obtain a system of $2k$ linear equations that can be solved for $\xi_s, \eta_s, s \in \{1, 2, \dots, k\}$. For example, when $k = 1$, we can solve to obtain $\xi_1 = \frac{\frac{c_w}{r_1^*} + c_v}{1-\alpha\left(1-\mu(r_1^*)\right)}$ and $\eta_1 = \frac{\alpha}{1-\alpha}\lambda_1\xi_1$. Therefore, we conclude that

$$V_1(y) = \frac{\frac{c_w}{r_1^*} + c_v}{1-\alpha\left(1-\mu(r_1^*)\right)}\left(y + \frac{\alpha}{1-\alpha}\lambda_1\right). \tag{3.16}$$

Because $V_1(y)$ is the *minimum* expected total discounted cost, r_1^* must minimize the expression $\frac{\frac{c_w}{r_1^*} + c_v}{1-\alpha\left(1-\mu(r_1^*)\right)}$. When slowdown dominates speedup or $\mu(r)$ is inverse U-shape, we can see that this is a case in which r_1^* is strictly greater than r_0 that maximizes $\mu(r)$ as mentioned in Section 3.4.2. We also note that when speedup dominates slowdown or $\mu(r)$ is U-shape and $\mu(\bar{r}) \geq \mu(r)$ for $r > 0$, Proposition 3.3 is consistent with Proposition 3.1 and the firm would always staff the minimum number of workers allowed. For $k > 1$, it is more complicated to solve for the optimal workload and thus we defer to the numerical studies.

The optimal staffing decision depends on the properties of $\mu(r)$. We now show that under a reasonable assumption on the shape of the function $\mu(r)$, the model has a global minimum in the domain under consideration, which allows us to use the first-order condition to solve directly for the optimal workload. We note that the slowdown and inverse U-shaped cases of our model meet this more general criteria.

PROPOSITION 3.4. *If $\mu(r)$ is quasiconcave with global maxmimum r_0 and there exists r_1 for which $\mu(r)$ is concave for $r \leq r_1$ and convex for $r \geq r_1$, there is a global minimum $r_s^* \in (0, \bar{r}]$ for $v_s(y, r)$.*

For cases not covered by Proposition 3.1 or shift-by-shift staffing, the optimal staffing decision in general is dependent on the number of requests in the system and difficult to characterize analytically. In the next section, we perform numerical experiments to gain more insights on optimal staffing policies for those cases.

3.5. NUMERICAL EXPERIMENTS

In this section, we numerically analyze the model presented in Section 3.3 for cases where a closed form solution is not available. Specifically, we numerically study the optimal staffing decision beyond Proposition 3.2 for cases when slowdown dominates or $\mu(r)$ is inverse U-shaped. We also compare the performance of several heuristics that are much easier to implement to that of the optimal policy. Finally, we investigate the importance of having the optimal advance schedule when employing different staffing policies. For sensitivity analysis, we test a wide range of parameter values to examine the robustness of the optimal policy. In particular, we set worker costs per shift to \$400 and consider variable costs of service ranging from \$100 to \$700 per shift, backorder costs ranging from 1/20 to 2 times variable cost, wage premiums of 25 to 100 percent, recaptured wages ranging from 0 to 80 percent, and a discount rate of 0.99. We consider a cycle size of 1 or 2 with arrivals per shift λ=15 or $\{\lambda_1, \lambda_2\}$={18,12} or {25,5}. In each of the arrival scenarios, the average number of arrivals per shift is 15 service requests. The theoretically achievable service rate (in the absence of speedup or slowdown effects) is 1/6, meaning that when speedup and slowdown are incorporated, the average service time is at least 6 shifts, resulting in average occupancy of at least 90, compared to capacities of 120, 130, or 140 (although since service rates change from shift to shift

according to the workload, the actual average occupancy may be substantially above 90). A complete list of values is shown in Table 3.3.

Parameter	Values
c_w	400
c_v	100, 300, 500, 700
c_h	$0.05c_v$, $0.2c_v$, c_v, $2c_v$
ψ	0.25, 0.5, 1
ϕ	0, 0.4, 0.8
b	120, 130, 140
α	0.99
$\bar{\mu}_{sp}$, $\bar{\mu}_{sl}$	1/6
r_{sp}	3
r_{sl}	7
$\bar{r}$	8
λ (for $k = 1$)	15
$\{\lambda_1, \lambda_2\}$	{18,12}, {25,5}

Table 3.3. Experimental Design

3.5.1. Optimal Recourse Actions

We first study the structure of the optimal recourse action. Throughout this section, we assume the firm uses the optimal number of staff in the advance schedule with respect to recourse actions. (We study the effect of the advance scheduling decision in Section 3.5.3). When slowdown dominates speedup or $\mu(r)$ is inverse U-shape, Proposition 3.3 tells us that the optimal workload is independent of y if $\psi = 0$, $\phi = 1$, and $b = \infty$. When these conditions are not met, the optimal staffing policy is dynamic and depends on the number of requests in the system. Figure 3.5 illustrates the optimal recourse policy for a sample set of parameters which is generally representative of our experiments. At the beginning of a shift, there is an initial staffing level z and a realized number of requests y. The line in region 2 represents the resulting workload if no adjustments are made to the

workforce for the current shift. Note that if no recourse is taken anywhere, the resulting graph (which would no longer be optimal) would be an extension of the line in region 2 to regions 1 and 3. However, this would result in a very low (suboptimal) workload in region 1 or a very high (suboptimal) workload in region 3. Consistent with Table 3.1, we thus divide the optimal decision into three different regions based on the number of requests relative to the established schedule: overstaffed (region 1), adequately staffed (region 2), and understaffed (region 3). We call the point separating regions 1 and 2 the "overstaffing threshold" and the point separating regions 2 and 3 "understaffing threshold". In region 1, some workers are sent home to maintain the optimal workload. In region 2, no recourse action is taken—staffing is held constant; the benefit of maximizing the worker productivity is outweighed by the cost of wage changing capacity (in the form of wage premiums or unrecovered wages) and the workload grows linearly in the number of requests. Finally in region 3, additional workers are obtained to keep the workload down to the optimal level. Thus, to achieve the optimal workload in region 1, workers must be sent home, while to achieve the optimal workload in region 3, workers must be added.

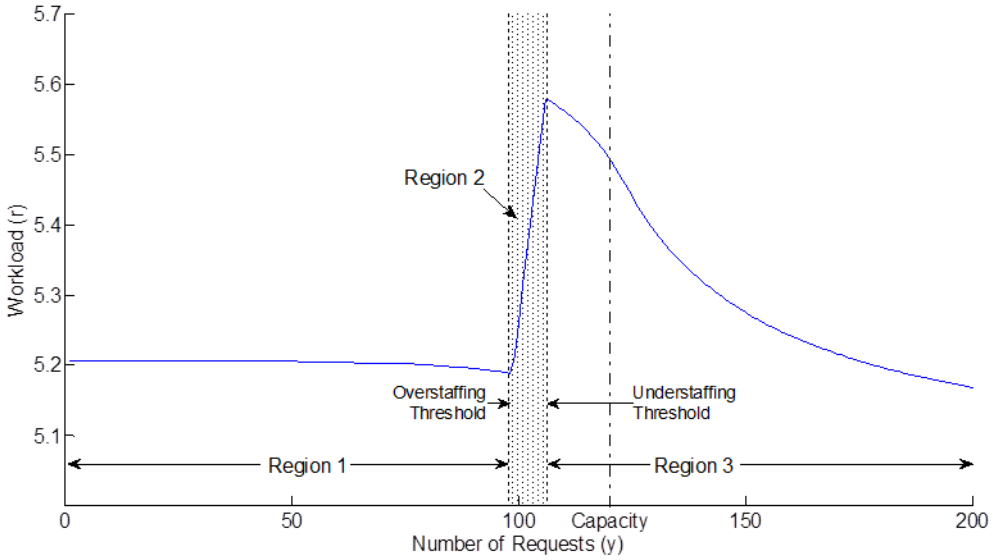


Figure 3.5. Optimal Staffing Policy ($c_v = 300, \phi = 0.4, \psi = 0.5, c_h = c_v, b = 120$)

We observe that if the number of requests at the beginning of the period is in region 1, the optimal workload is nearly constant (although slightly non-increasing) in the number of requests. Thus, workers are sent home to maintain a nearly constant ratio of requests per worker. If the number of requests is in region 2, no workers are added or subtracted to the existing schedule and so the workload grows linearly in the number of requests. Interestingly, in region 3 the optimal workload (requests per worker) is not approximately constant (as in region 1), but decreases as the number of requests in the system increases, even though staffing to a lower workload requires high utilization of expensive on-call workers. This is because the benefit provided by additional workers obtained when the number of requests is very high is twofold, and this benefit is high enough to incentivize the firm to increase staffing in the current shift instead of waiting to add staff in future shifts. First, as intuition suggests, the additional staff increase the throughput of the system. In addition, and perhaps more subtly, their presence results in lower workload assigned to each worker, thereby reducing the degree of slowdown

experienced by each worker. Recall that Proposition 3.2 tells us that for the cases of slowdown dominates or inverse U-shape, the optimal workload will always be in a slowdown region, where the service rate decreases with the workload. As the number of requests increases, slowdown becomes ever more costly because with slowdown, fewer requests will depart, resulting in continued congestion in future time periods. To avoid this propagation of congestion, the firm should employ even more workers in the current shift, which explains the decreasing optimal workload in region 3. The decrease in optimal workload becomes even steeper once the number of requests is greater than the capacity of the system (shown as a dotted line in Figure 3.5). Because backlogged requests incur costs but cannot be processed until admitted into service, the firm benefits from aggressively staffing to reduce the number of requests in the system below capacity and preventing propagation of backlogs into future time periods.

3.5.1.1. Sensitivity Analysis

We perform sensitivity analysis to examine the behavior of the optimal policy in response to changes in the system parameters. Figure 3.6 compares the optimal recourse policies when $\phi = 0$, 0.4, and 0.8. We observe that even when $\phi = 0$ (meaning the firm does not recoup any wage for the workers sent home), it is still optimal for the firm to send workers home to keep the workload high enough to achieve peak productivity from workers on duty. Region 1 becomes wider and the optimal workload in region 1 increases as ϕ increases. Because the firm recoups a larger portion of the wages for the workers sent home for higher values of ϕ, the firm is more willing to send workers home for a given number of requests in the system. Similarly, Figure 3.7 shows that as ψ increases, the optimal workload in region 3 increases and the understaffing threshold increases to

larger y value. From Proposition 3.3 we know that for $\phi = 1, \psi = 0$, and no capacity constraint, the optimal workload is constant. In Figure 3.6 and Figure 3.7, we can see that the workload indeed converges to a horizontal line as we approach those values.

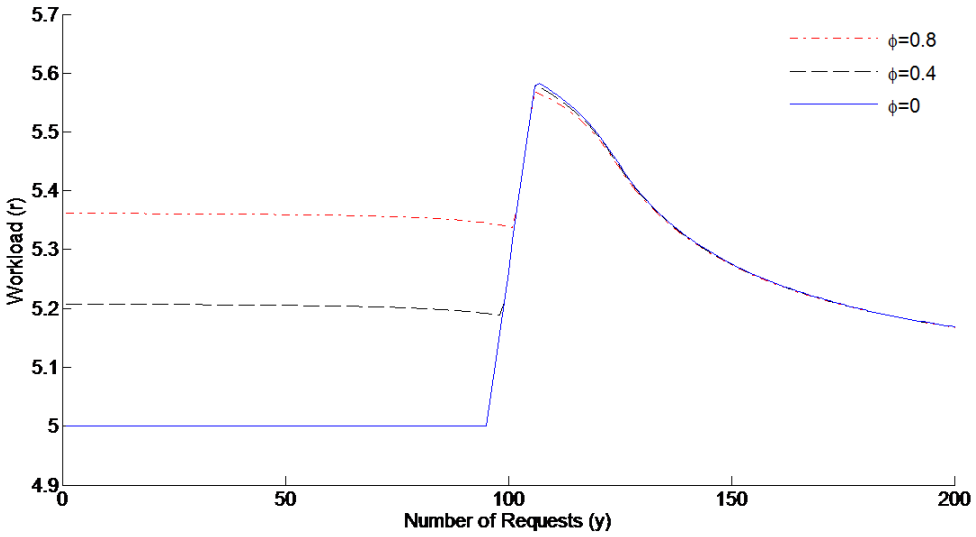


Figure 3.6. Optimal Staffing Policy for Various ϕ Values ($c_v = 300, \psi = 0.5, c_h = c_v, b = 120$)

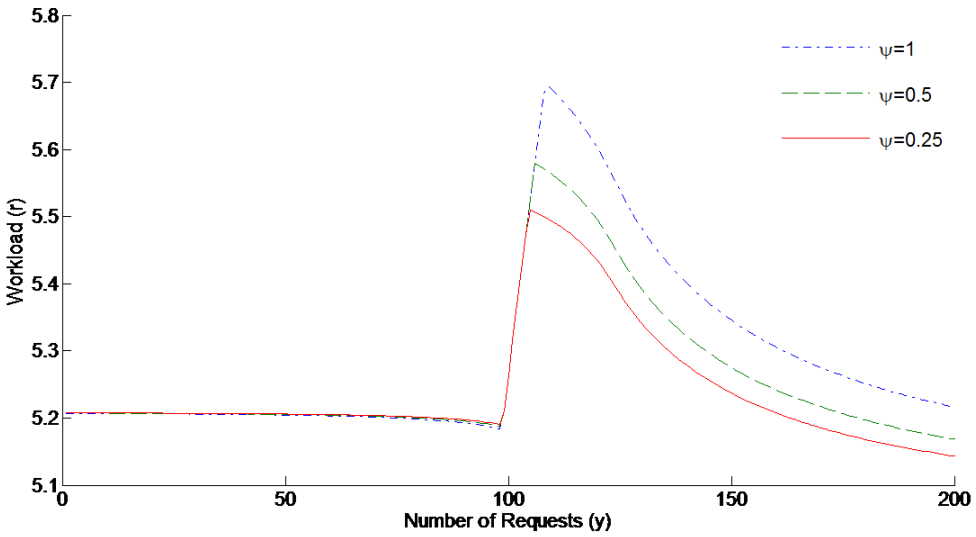


Figure 3.7. Optimal Staffing Policy for Various ψ Values ($c_v = 300, \phi = 0.4, c_h = c_v, b = 120$)

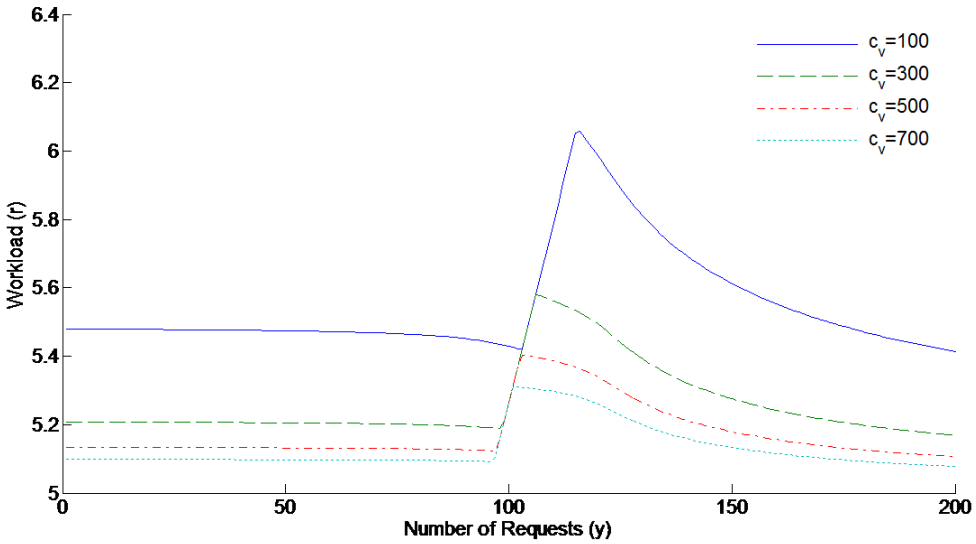


Figure 3.8. Optimal Staffing Policy for Various c_v Values ($\phi = 0.4$, $\psi = 0.5$, $c_h = c_v$, $b = 120$)

Figure 3.8 shows that under an optimal policy the firm staffs more aggressively and the workload decreases in the variable cost per request c_v. The number of workers sent home is greatest for low values of c_v while the number of on-call workers utilized is greatest for large values of c_v. This is because the cost of staffing is more easily outweighed by the future expected variable costs when c_v is larger. Thus when c_v is larger, the firm would staff more workers to process more requests in the current period, either by sending fewer workers home (in region 1) or bring in more on-call workers (in region 3).

Figure 3.9 shows that the firm's policy on on-call workers changes drastically with changes in holding cost per backlogged request. As the holding cost per request becomes more expensive, the firm becomes more aggressive in hiring on-call workers to reduce the number of requests below capacity and avoid backlogs in future time periods (see the steep decrease in region 3 when $c_h = c_v$ or $c_h = 2c_v$). However, when holding

cost is extremely low compared to the variable cost, the optimal workload starts to increase when the system is over capacity as the firm slows its hiring of on-call workers once the system is full. In part, this phenomenon is due to the fact that while hiring extra workers mitigates slowdown, backlogged work cannot be placed into service, no matter how many additional workers are obtained, and so the benefit of extra on-call workers accrues only to future periods (by increasing the departure of existing requests). Once the system is full, it is more cost-effective for the firm to reduce its current-period cost by bringing on fewer on-call workers, which increases the workload (note that workload is the number of requests *in service* per worker; the backlogged requests are not counted). Because there is a reasonably high probability that the system would stay full for the foreseeable future, it is optimal for the firm to enjoy savings from a decrease in the number of on-call workers while incurring very low holding costs rather than trying to reduce the number of requests as quickly as possible.

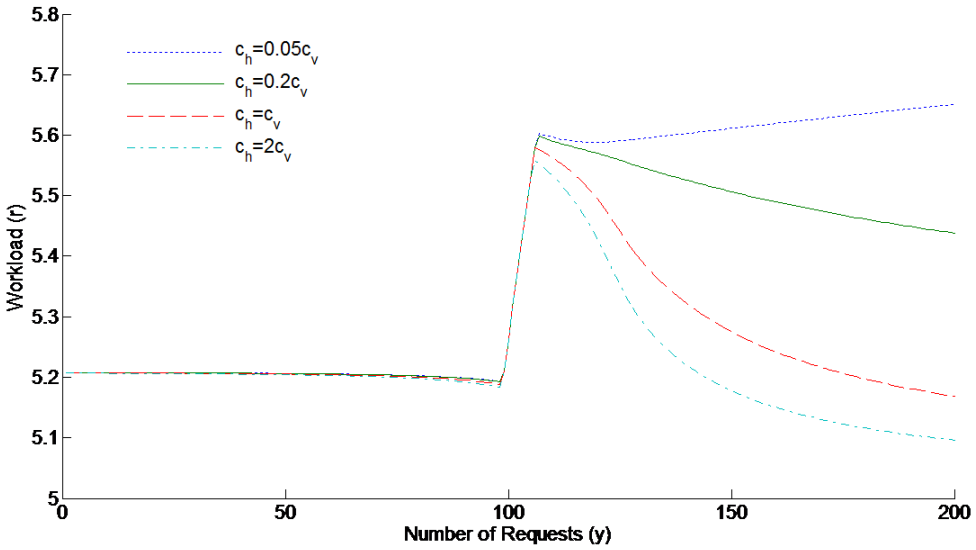


Figure 3.9. Optimal Staffing Policy for Various c_h Values ($c_v = 300$, $\phi = 0.4$, $\psi = 0.5$, $b = 120$)

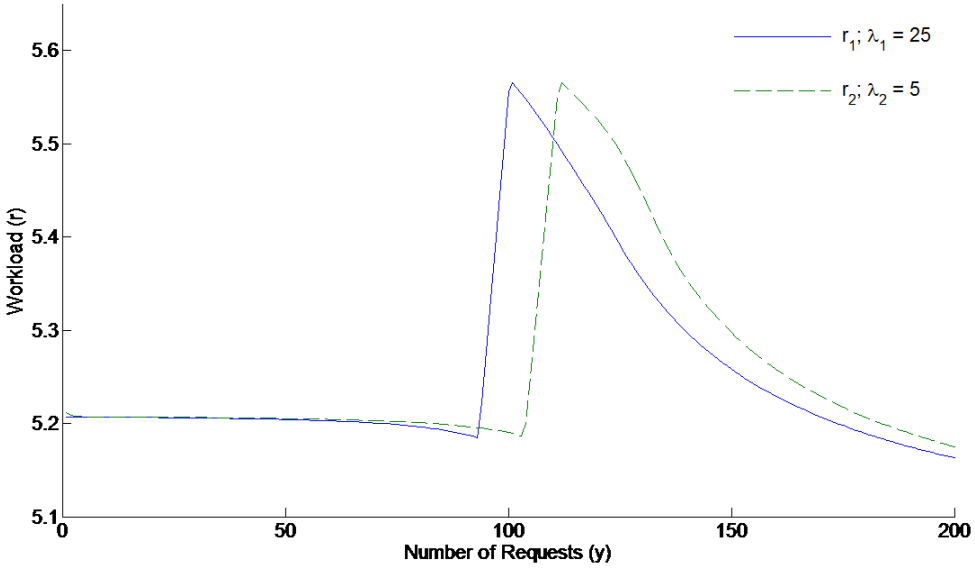


Figure 3.10. Optimal Staffing Policy when Arrivals are Cyclic
($c_v = 300, \phi = 0.4, \psi = 0.5, c_h = c_v, b = 120$)

3.5.1.2. Cyclic Arrivals ($k > 1$)

The characteristics of the optimal policy remain similar to the non-cyclic case when $k > 1$ and arrivals are cyclic. Figure 3.10 presents the optimal policies for two shifts in a cycle ($k = 2$) and $\{\lambda_1, \lambda_2\} = \{25,5\}$. The shape of the optimal policy for each shift is almost identical to that under single arrival pattern with corresponding λ=15. We remind the reader that in this section we assumed the firm uses an optimal advance schedule: in our experiments with cyclic arrivals, we find that the main difference in staffing for the two shifts come from the advance schedule instead of the recourse actions, which also explains why r_1 and r_2 have essentially identical shapes but with r_2 shifted to the right to account for the different advance schedule. The primary difference is that the $\lambda_1 = 25$ shift anticipates the current period's high demand compared to the capacity while $\lambda_2 = 5$ shift anticipates lower demand and responds accordingly. Consequently, the behavior of the optimal policy with respect to changes in system parameters discussed in the previous

subsection still apply when arrivals are cyclic. In Section 3.5.2, we exploit this similarity with the non-cyclic case to develop a "one-step look-ahead" heuristic.

3.5.2. Comparison with Different Heuristics

While the optimal recourse policy guarantees the best performance, it may not be practical to implement because of the difficulty in obtaining the exact solution. We now compare the performance of various heuristics that are easier to implement than the optimal policy. We assume the firm uses the optimal advance schedule, which may be different for each policy (we study the importance of an optimal advance schedule in Section 3.5.3), and only compare the performances of different recourse actions. First, we test the performance of the single-ratio policy. For this policy, the chosen workload does not depend on the number of requests in the system. When arrivals are cyclic, we assume that the firm employs the same ratio throughout the cycle to keep the workload at a constant level. Second, we test the performance of three myopic policies: single-shift with zero terminal cost, single-shift with non-wage terminal cost, and two-shift with zero terminal cost (after second shift). Single-shift myopic policy with zero terminal cost looks only at the cost of the current shift and thus always fix the workload at $\bar{r}$. Single-shift myopic policy with non-wage terminal cost treats every shift as if it is last shift of the finite horizon and the firm incurs terminal cost for each request remaining at the end. Two-shift myopic policy with zero terminal cost treats every shift as if it is second-to-last shift of the finite horizon with zero terminal cost after the last shift. We present the results for two-shift myopic policy with zero terminal cost in the figures because it is the best-performing myopic policy. Third, we test the "one-step look-ahead" policy implemented on a rolling horizon, for which the firm minimizes the expected total

discounted cost assuming it will revert back to the single-ratio policy starting next shift. The one-step look-ahead policy is equivalent to applying one step of the policy improvement algorithm starting from the single-ratio policy.

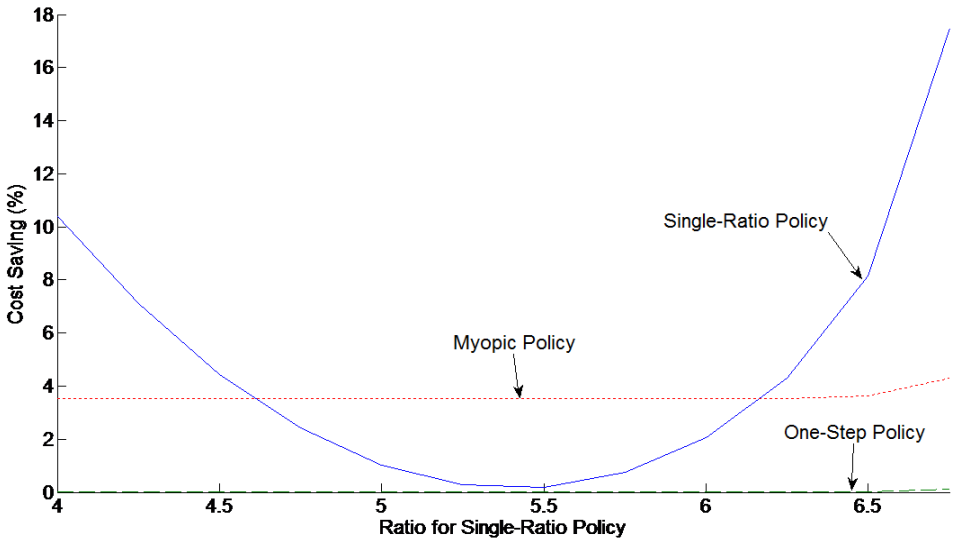


Figure 3.11. Possible Cost Savings by Using the Optimal Policy over Other Heuristics for the Non-Cyclic Case
($c_v = 300, \phi = 0.4, \psi = 0.5, c_h = c_v, b = 120$)

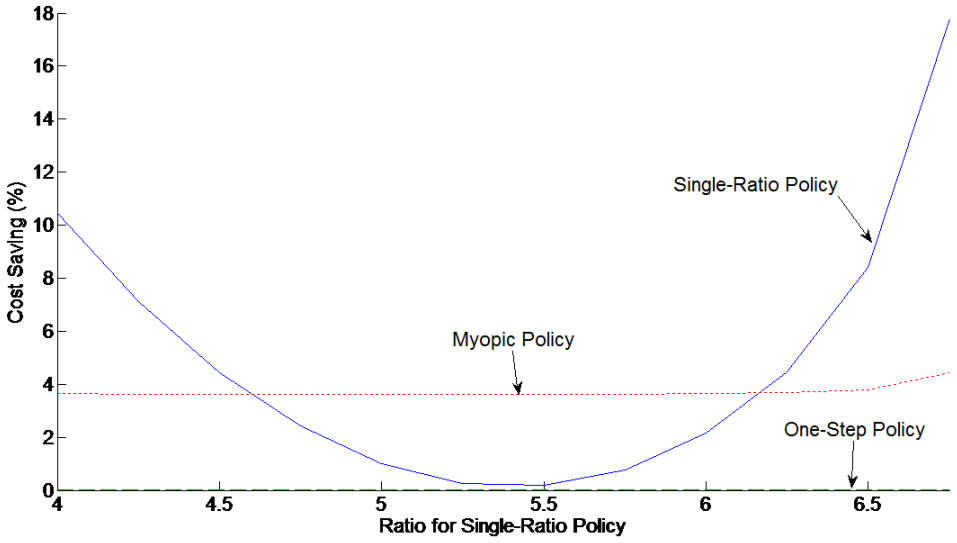


Figure 3.12. Possible Cost Savings by Using the Optimal Policy over Other Heuristics for the Cyclic Case
($c_v = 300, \phi = 0.4, \psi = 0.5, c_h = c_v, b = 120, \lambda_1 = 25, \lambda_2 = 5$)

Figure 3.11 and Figure 3.12 illustrate the percent improvement the firm can experience by employing the optimal policy over the three heuristics listed above when arrivals are non-cyclic ($k = 1$) and cyclic ($k = 2$), respectively. Note that the one-step policy curve is at a cost savings very close to zero and therefore barely visible in Figure 3.11 and Figure 3.12. To compute the expected cost, we needed an initial census distribution, which we computed assuming the firm currently uses a single-ratio policy. The horizontal axis represents the various fixed workloads that the firm can use for its current single-ratio policy while the vertical axis represents the percent improvement the firm experiences in costs. Because we assume the single-ratio policy to be the current policy, our comparison provides a built-in advantage to the single-ratio recourse policy. In spite of this, the firm can still improve its costs substantially by switching to the optimal recourse policy unless its current single-ratio recourse policy uses a ratio very close to 5.5. We note that even determining the best single-ratio policy requires accounting for speedup or slowdown. Since we already know that $r_s^*(y) \geq r_0$ from Proposition 3.2 for the cases of slowdown or inverse U-shape, there is a chance that the firm's conjectured ratio for the single-ratio policy would be close to optimal if the firm has sufficient knowledge on the effects of speedup and slowdown on $\mu(r)$ of its employees. On the other hand, there are substantial negative cost consequences if a firm selects an inappropriate ratio for the single-ratio policy—for example, the optimal policy has as much as 12% lower cost than a single ratio policy of 4.0.

The myopic policy performs reasonably well when ϕ is small or ψ is large, but the performance deteriorates as ϕ increases or ψ decreases. This observation can be explained by the fact that when ϕ is small or ψ is large, it is much less likely for the firm

to take any recourse actions for both the myopic and the optimal policies. Thus, the myopic policy resembles the optimal policy much more than it would otherwise. As ϕ increases or ψ decreases, the firm now has extra incentives to take recourse actions in addition to improvement in worker productivity, and consequently the performance of the myopic policy suffers.

Figure 3.11 and Figure 3.12 also clearly demonstrate that the one-step look-ahead recourse policy (barely visible in the graphs with values very close to zero) is a potential substitute for the optimal policy. In our numerical experiments, the difference between the performances of the optimal and the one-step policies was less than 1% for all system parameter values. Since the one-step policy greatly reduces the solution complexity of the MDP model, it is an attractive candidate as a recourse policy for firms to enjoy "near-optimal" performance without extensive computational effort. The exceptional performance of the one-step policy shows that consideration of just a few shifts into the future can be acceptable in the presence of discounting (recall that we use $\alpha = 0.99$). This observation coincides with the fact that the optimal policy under cyclic arrivals is very similar to a combination of k different optimal policies with varying λ values under single arrival pattern. Considering the infinite-horizon problem with k shifts in each cycle as multiple problems with k different λ values can still lead to adequate performance. On the other hand, the difference in performance between the myopic and the one-step policies emphasizes the importance for firms to consider future implications, even for a few shifts, of their actions taken in the current period.

3.5.3. Sensitivity to a Suboptimal Advance Schedule

The results of the numerical experiments from Sections 3.5.1 and 3.5.2 are based on an optimal advance schedule, which is of course the best case in practice. Therefore, we examine the consequences of suboptimal advance scheduling when various recourse actions are available to compensate. Figure 3.13 shows the percent increase over the optimal cost for a range of advance schedule decisions when *optimal recourse* actions are taken. We see that as long as the advance schedule is relatively close to the optimal schedule, the subsequent optimal recourse actions can compensate for most of the cost incurred by suboptimal advance scheduling. As expected (but not shown on the figure), higher values of ϕ increase the ability of recourse actions to offset a suboptimal advance schedule. Similarly, as ψ decreases the firm gains more ability to correct its mistake in the advance schedule due to lower cost of utilizing on-call workers. Thus, optimal advance scheduling is less important as ϕ increases or ψ decreases (i.e., as recourse becomes less expensive). Furthermore, the impact of a suboptimal advance schedule decreases as c_v increases because wage costs represent relatively smaller portion of the total costs when c_v is high, and hence recourse actions appear less expensive.

Figure 3.14 compares the single-ratio policy and the one-step policy to the optimal policy for ranges of initially scheduled staff z. For the most part, the performance of the single-ratio policy does not depend on the initial staffing level z. Higher z values perform somewhat better for lower ratios because they reduce the need for on-call workers to meet the requirements of a low workload per worker. On the other hand, a firm with a high single-ratio policy prefers lower z values to minimize the number of

workers to send home. Figure 3.14 shows that the one-step policy performs very well even when the advance schedule is not optimal.

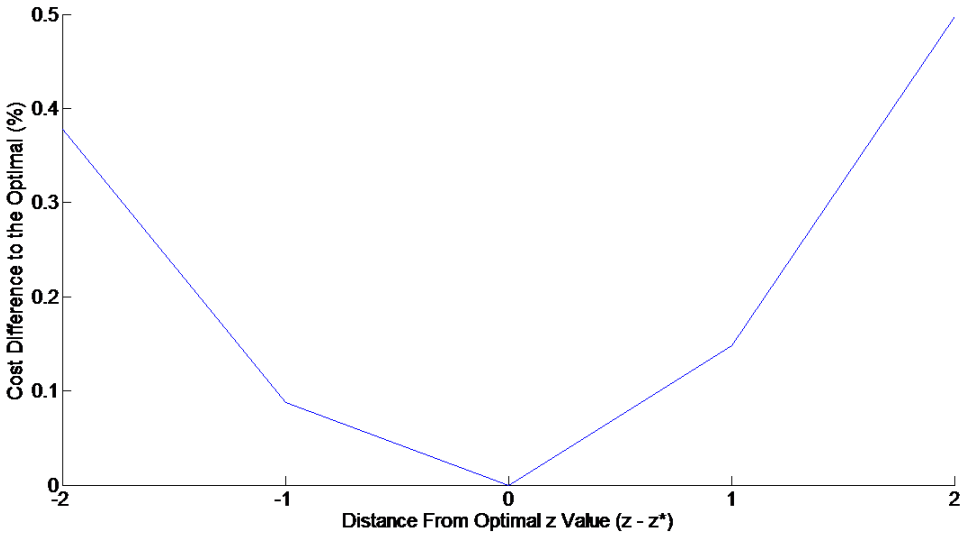


Figure 3.13. Cost Savings Achieved by Using the Optimal z Over Other Values ($c_v = 300, \phi = 0.4, \psi = 0.5, c_h = c_v, b = 120$)

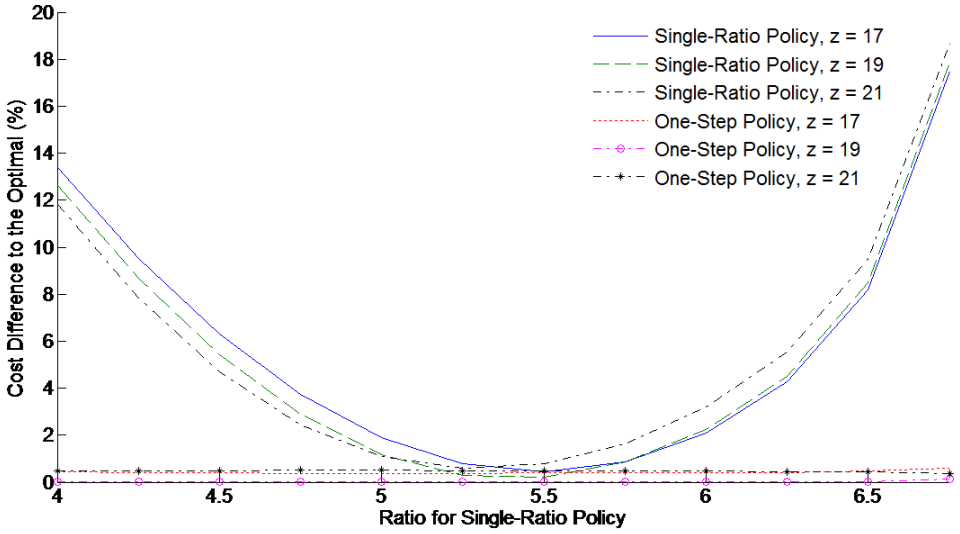


Figure 3.14. Performance of the Single-Ratio and the One-Step Policies for Different z Values ($c_v = 300, \phi = 0.4, \psi = 0.5, c_h = c_v, b = 120$)

It is worth noting that the impact of the firm's decision from the advance schedule is very much related to the effectiveness of single-ratio policy. If the firm has a good understanding of the productivity of its employees with respect to workload and therefore is able to set a reasonable ratio for its single-ratio policy, it can use the same knowledge to determine the range of z whose performances are close to that of the optimal advance schedule. This observation again underscores the importance of incorporating the workers' tendency to speedup or slowdown in staffing. There is an opportunity for firms to understand the implications of their staffing decisions on the productivity of their employees in order to avoid an inferior advance schedule that cannot be rectified even with optimal recourse actions.

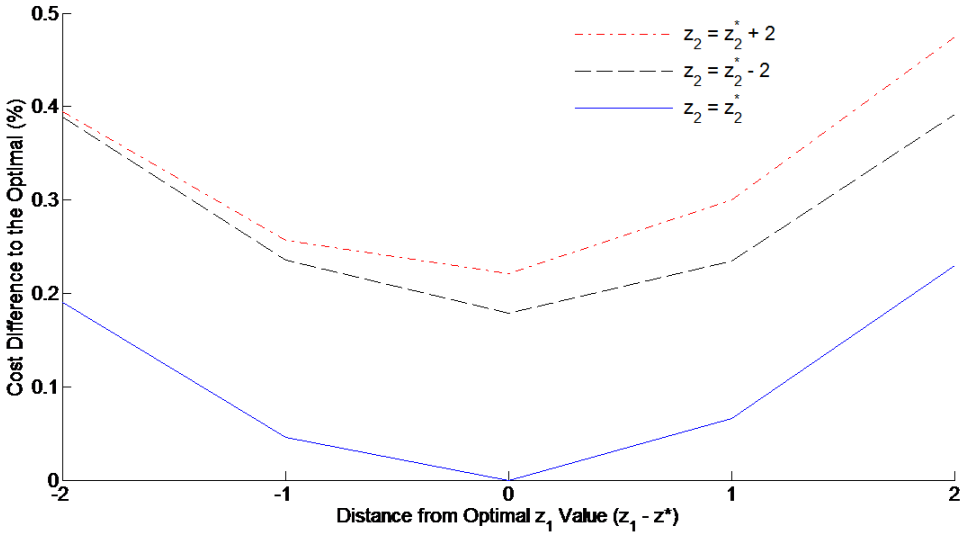


Figure 3.15. Cost Savings Achieved by Using Optimal z_1 over Other Values for Different z_2 Values in the Cyclic Case
($c_v = 300, \phi = 0.4, \psi = 0.5, c_h = c_v, b = 120, \lambda_1 = 25, \lambda_2 = 5$)

We close this section by observing that when arrivals are cyclic, the results are very similar to those of the non-cyclic case. Figure 3.15 shows the percent increase over the optimal cost for a range of z_1 values when optimal recourse actions are taken in three

different scenarios of z_2 ($z_2 = z_2^*$, $z_2 = z_2^* - 2$, and $z_2 = z_2^* + 2$). We observe that the firm can perform very well through the optimal recourse actions even when z_1 and z_2 deviate from the optimal values.

3.6. CONCLUSION AND FUTURE RESEARCH

Recent empirical studies of service systems have shown that servers employ a changing service rate depending on the workload at any given time. In this chapter, we modeled speedup and slowdown separately and used a convex combination of the two functions to represent many possible joint effects of these behavioral phenomena. We then incorporated the impact of speedup and slowdown into workforce staffing problem with recourse through the model. We showed that the optimal workload is independent of the number of customer requests under certain conditions and characterized the optimal dynamic recourse policy when these conditions are not met. When the optimal policy is dynamic, we showed that the workers' tendency to slow down causes the firm to aggressively use on-call workers even at a high premium. Similarly, the workers' tendency to speed up can incentivize the firm to send workers home even when it does not recoup any wages for unworked time. When the arrivals for customer requests are cyclic, the optimal policy for each shift has essentially identical shapes but with a different advance schedule. Exploiting this result, we developed a one-step look-ahead heuristic that greatly reduces the solution complexity of the MDP model and showed that it performs almost as well as the optimal recourse policy.

While our model for speedup and slowdown is able to represent many possible joint effects, we acknowledge that speedup may not be sustainable over extended periods

of time and slowdown may exacerbate over long periods of time. Consequently, a firm may experience changes in the degree of the joint effects over time even when the workload is maintained at the constant level. The consideration of this additional time dimension for behavioral phenomena could be an attractive direction for future research.

CHAPTER 4

THE PATIENT PATIENT: THE PERFORMANCE OF TRADITIONAL VERSUS OPEN-ACCESS SCHEDULING POLICIES

Abstract

We compare traditional and open-access scheduling policies for outpatient medical practices in terms of the number of patients served and financial performance. Under a traditional scheduling policy, a patient schedules an appointment in advance and there is a significant possibility of patient no-shows. In response, doctors overbook patients to reduce idle time created by no-shows. Under an open-access scheduling policy, all appointments are scheduled the day of the appointment, thereby eliminating patient no-shows but creating more randomness in the daily number of appointments. In contrast to earlier works, we consider the optimal average number of patients served and find that while the traditional policy may be more profitable by providing doctors more control over their schedule and ability to limit overtime, the open-access policy may lead doctors to serve a greater number of patients.

4.1. Introduction

The U.S. health care industry is facing various challenges with rising costs and limited capacity. Access to health care is least secure among affluent countries (Gulliford and Morgan 2003), and there are widespread shortages of nurses and primary-care doctors. According to the Association of American Medical Colleges (2010), the aging population of both doctors and patients as well as the Affordable Care Act will lead to a shortage of 45,000 primary care physicians and 46,000 surgeons and medical specialists in the United

States by 2020. These trends suggest that hospitals and outpatient medical practices need to be very efficient with limited resources in providing services to patients to meet the increasing demand and to maximize the number of patients seen each day. It also behooves policy makers to provide appropriate incentives to entice hospitals and outpatient medical practices to serve more patients.

Patient no-shows have been widely recognized as a major factor that interferes with the industry's effort in becoming more efficient. Traditionally, patients schedule their appointments weeks or even months in advance to enable doctors to have full control of their workday schedules. When a patient makes an appointment well in advance under a traditional scheduling policy, there is a significant possibility that he or she might not make it to the scheduled appointment. These patient no-shows can result in doctor idle time, which is ever more undesirable in the context of the aforementioned shortages. Various approaches have been attempted to reduce the number of patient no-shows: calling patients on the waiting list, providing reminder postcards, letters, or phone calls, and imposing financial penalties for no-shows. In response to potential patient no-shows, many doctors who employ traditional policies practice overbooking to varying degrees. The extent to which a doctor tries to avoid idle time and cover expected patient no-shows through overbooking depends on how much she values patient goodwill, since overbooking increases the probability that the patient will have to wait after arriving at the office.

Recently, the open-access scheduling policy has been proposed as a remedy for patient no-shows because it minimizes the appointment delay, or time until an appointment. Under an open-access policy, patients call in on the morning of their

preferred day and schedule an appointment for the same day. By minimizing the appointment lead time, an open-access policy nearly eliminates patient no-shows because appointment delay has a significant impact on rate of no-shows (Gallucci et al. 2005).

We compare the performance of traditional and open-access policies in order to provide insights as to their effects on doctors and patients. While there are other advantages and disadvantages related to implementing both appointment policies, we focus on the financial implications for outpatient medical practices to examine the doctors' incentives in the policy decision. At the same time, we compare the number of patients served under the two policies to investigate which policy is preferable for policy makers and the general public.

The performance of the different appointment scheduling policies depends on the patient demand an outpatient medical practice receives. As noted earlier, hospitals and outpatient medical practices often face the favorable setting where demand exceeds supply, allowing doctors to fully book any preferred schedule. Traditional scheduling policies allow the doctors to experience a lesser degree of randomness in overall daily patient demand while open-access policies reduce the risk of idle time for doctors by eliminating (or at least greatly reducing) patient no-shows.

In contrast to recent articles (Murray and Tantau 2000, Robinson and Chen 2010) that concluded an open-access policy is preferable because it eliminates patient no-shows (which result in doctors' idle time), our model finds that the comparison is more nuanced. We observe that the randomness in daily patient demand plays a role in determining which policy outperforms the other. In particular, when an open-access policy results in

relatively high randomness in patient demand, the traditional policy can be more effective in utilizing the doctor's time. On the other hand, it is likely that more patients are served under the open-access policy. Our results provide a possible explanation to the current situation, in which most of the doctor's offices employ a traditional policy while the patients prefer the open-access policy. We also provide insights that can help policy makers to better incentivize the doctors to implement the socially-optimal policy.

The remainder of the chapter is organized as follows. First, we discuss the relevant literature. Then we present profit-maximization models for both traditional and open-access scheduling policies. We perform an analytical comparison of special cases of the two policies. Next, we compare the two policies using numerical experiments that elucidate tradeoffs and we report the results. Finally, we provide concluding remarks and future research opportunities. All proofs are provided in the Appendix C.

4.2. Literature Review

Appointment scheduling in the health care industry has been studied extensively by the operations research literature. Bailey (1952) and Lindley (1952) are among the pioneers in research on outpatient appointment scheduling. Cayirli and Veral (2003) and Gupta and Denton (2008) provide comprehensive surveys of research on appointment scheduling.

Overbooking has received much attention from researchers in revenue management, specifically in airline and other transportation industries (Hillier and Lieberman 2001, Barnhart et al. 2003, Van Ryzin and Talluri 2003). In the health care industry, overbooking has been viewed both as a source of patient dissatisfaction and an

approach to compensate for patient no-shows. Lau and Lau (2000), Denton and Gupta (2003), and Robinson and Chen (2003) focus on minimizing patient waiting time, physician idle time, and staff overtime. More recently, LaGanga and Lawrence (2007) examine the use of appointment overbooking to improve overall clinic performance while Chakraborty et al. (2010) show that a scheduling policy using overbooking provides an optimal stopping rule that determines how many patients are scheduled in a given day. Tsai and Teng (2014) develop a stochastic overbooking model that considers patients' call-in sequence for outpatient clinics with multiple resources that outperforms traditional appointment policy.

Lately, an open-access scheduling policy has been gaining popularity both in application and research. Murray and Tantau (1999) are credited as the first to present what is now known as advanced or open-access system. Kopach et al. (2007) develop a simulation model to study the effects of clinic parameters on implementation of the open-access policy, and Green and Savin (2008) present a single-server queueing model to identify maximum patient panel sizes for medical practices using open-access policy. Liu et al. (2010) propose heuristic dynamic policies for scheduling patient appointments and find that the open-access policy can be a reasonable choice when the patient load is relatively low. On the other hand, Patrick (2012) develops a Markov decision process model that shows that a short booking window performs better than an open-access policy. Samorani and LaGanga (2015) study the problem of optimally overbooking appointments given no-show predictions that depend on the individual characteristics and on the appointment day, and suggest a heuristic that should be preferred to open-access under most parameter configurations.

The work most relevant to this essay is Robinson and Chen (2010) (denoted RC), who compare the performance of traditional and open-access policies. Employing a cost minimization model, RC conclude that the performance of an open-access policy dominates that of a traditional policy. Importantly, the average number of patients served is exogenous when comparing the two methods in RC. As in RC, we compare two methods of appointment scheduling: a traditional scheduling policy under which patients schedule well in advance, and an open-access policy under which patients schedule a same-day appointment at the beginning of the day. While our model and research focus is similar to RC, we examine the more general problem by expanding the comparison and increasing the number of decision variables. RC implicitly assume that it is preferable and beneficial for doctors to leave before the end of the day. Our formulation allows us to generalize implicit assumptions of RC, and capture the possible impact of having different average number of patients served on the overall performance of the scheduling policies. Having a higher average number of patients served allows more revenues and possibly more profits but may require additional overtime. In contrast to RC, we find that the relative financial performance of a traditional policy (versus that of an open-access policy) does not always suffer from the behavior of patient no-shows. Rather, our model finds that the traditional policy outperforms the open-access policy in terms of the doctor's profits in many scenarios, but the open-access policy encourages the doctors to serve more patients on average.

4.3. Model

In this section, we develop stylized models for two contrasting scheduling policies and discuss basic assumptions. Our models are direct extensions of the models developed by

RC, and retain many of their assumptions. However, our approach fundamentally differs in that the objective of our model is to maximize profits instead of minimizing costs (LaGanga and Lawrence 2007, Muthuraman and Lawley 2008). Following the nomenclature used by RC, we refer to the two policies as "traditional" and "open-access."

We make the following assumptions. Each patient pays a price of r for service that costs c for the doctor. Without loss of generality, we normalize c to zero to simplify the analysis. As a result, r can be thought of as profit margin. For the remainder of this chapter, we use the subscript "T" and "OA" to represent the traditional and open-access policy, respectively. The doctor sees D_T patients in a day under the traditional policy, and D_{OA} patients in a day under the open-access policy. A further characterization of D_T and D_{OA} is detailed in the next two subsections. Each day consists of n regular (i.e., not overtime) appointment slots. Scheduling appointments beyond n requires overbooking and/or overtime. Define θ_i as the number of slots beyond n that the doctor stays to complete all service under policy i. We assume overtime costs $c_{ot}(\theta_i)$ to be convex in θ_i to capture the increasing marginal cost of overtime (LaGanga and Lawrence 2012, Truong 2015). Consistent with approaches that have been taken by many other researchers (e.g., Ho and Lau 1992, Cayirli et al. 2006, Robinson and Chen 2010), we assume that service times are constant, patients arrive on-time for their appointments, and there are no emergency patients that the doctor needs to accommodate. Consequently, each appointment slot is equal in length, which is equal to the service time. These assumptions allow us to focus and concentrate on the effect of the no-show rate on the average number of patients served and profit.

4.3.1. Traditional Scheduling Policy

Under the traditional policy, patients make an appointment well in advance of their preferred time. Given that their appointment is well in advance, there is a strong possibility that he or she might not make it to the scheduled appointment. We assume an exogenous probability $p \in [0,1]$ that each patient will not show up. The number of patients the doctor actually sees in a day D_T is $Q - X$, where Q is (the decision for) the number of appointments she schedules in a day and X is the (random) number of patient no-shows. There are n available regular-time appointment slots and the doctor can schedule more than n appointments using overbooking and/or overtime. We assume that there is sufficient demand such that all Q appointments can be booked. The number of patient no-shows X is binomially distributed with population of Q and probability p. Therefore, the expected realized daily demand $E[Q - X] = \mu_T$ is $Q(1 - p)$.

If a patient fails to show up, the doctor—if she does not overbook–is idle and incurs an opportunity cost of lost revenue. Hence, the doctor has an incentive to overbook in order to reduce idle time. The doctor overbooks to cover a proportion $\gamma \in [0,1]$ of the expected no-shows Qp. The amount of overbooking is determined by how much she values her patients' time relative to her idle time. If she values her patients' time significantly more than her own (i.e., she is very patient-friendly), she would never overbook any patients. A doctor who chooses to overbook no one is characterized by $\gamma = 0$, although this patient-friendly doctor could still choose to schedule $Q > n$ through overtime. Alternatively, if she wants to reduce her expected idle time (i.e., she is less patient-friendly), she can overbook to cover some or all of her expected no-shows. Thus, γ represents the "patient friendliness" of the doctor with zero being the most patient-

friendly (with no overbooking) and one being the least patient-friendly (with overbooking that covers every no-show). Under this definition, the number of overbooked patients is $m = Qp\gamma$ and the number of appointment slots the doctor needs to stay to complete service is $\theta_T = [Q - n - \min\{X, m\}]^+$. Note that the overbooking process is imperfect—while the doctor knows the expected number of patient no-shows, she cannot know which patients will not show. As a result, the doctor may still experience idle time despite overbooking if a patient scheduled in an earlier slot does not show up. We refer to this phenomenon as "idleness mismatch". Our equation for θ_T assumes that there is no idleness mismatch—that is, the doctor's idle time will not occur so long as overbookings exist. For example, if three slots are overbooked, idleness will not occur until there are four no-shows. For low ranges of γ, this is a reasonable assumption because in that case idleness mismatches occur with low probability—the number of overbooked patients m is small and it is not very difficult for the doctor to ensure the availability of overbooked patients at the precise time of the patient no-shows by scheduling the (few) overbooked appointments towards the beginning of the day. For higher values of γ, our model can overstate profits as it fails to account for idleness that might arise in an early period before an overbooked patient arrives in a later period. Given this limitation of our model (and because we believe that in general doctors are not callously unfriendly), we only consider low values of $\gamma \leq 0.3$ in our analysis and experiments. The doctor incurs a cost of lost goodwill $c_{gw}(g)$ for patients who are overbooked or booked in the overtime slots, $g = Q - n$, and we assume a convex cost function to reflect an increasing marginal cost. An increasing marginal cost reflects a dynamic where a small number of overbooked patients are easily absorbed by the idle slots resulting from no-shows, while large

overbookings can result in significant queueing effects as appointments stack up. We treat γ as an exogenous descriptor of the doctor but in a later section explore the sensitivity of our model's results to various values of γ.

Before overbooking, we assume that a doctor single-books the first n patients to the n appointment slots. As shown in RC, the optimal traditional policy contains no "holes" in the schedule; i.e., if it is optimal to schedule a patient for a given time slot, then it will be optimal to schedule patients for every earlier time slot. Our objective is to maximize the expected profit for doctors. The expected profit for a doctor under the traditional policy is

$$\Pi_T(Q) = E\left[rD_T(Q) - c_{ot}(\theta_T) - c_{gw}(g)\right]. \tag{4.1}$$

In order to facilitate comparisons between traditional and open-access policies, for the remainder of this chapter, we assume $c_{ot}(\theta_i)$ and $c_{gw}(g)$ to be quadratic functions such that $c_{ot}(\theta_i) = k_{ot}\theta_i^2$ and $c_{gw}(g) = k_{gw}g^2$, where k_{ot} and k_{gw} are an overtime cost parameter and a cost parameter for loss of goodwill, respectively. For a patient-friendly doctor with $\gamma = 0$, who does not overbook but still may choose to work overtime, Proposition 4.1 presents the optimal number of appointments she schedules under a traditional scheduling policy.

PROPOSITION 4.1. *Under a traditional scheduling policy and when* $\gamma = 0$*, the optimal number of appointments the doctor schedules in a day is* $n + \frac{r(1-p)}{2(k_{ot}+k_{gw})}$.

As expected, the optimal number of appointments increases with revenue per patient and number of regular-time appointment slots, while it decreases with the cost of

overtime and loss of goodwill. The second term $\frac{r(1-p)}{2(k_{ot}+k_{gw})}$ is the amount of overtime the doctor works to maximize her profit. The increasing marginal costs of overtime and loss of goodwill make any additional overtime too costly for the doctor. For $\gamma > 0$, this problem has no closed-form solutions for optimality conditions, but can be studied numerically, and we do so in Section 4.5 where we compare the performance of traditional and open-access policies.

4.3.2. Open-Access Scheduling Policy

Under the open-access policy, patients call in on the morning of their preferred day and schedule an appointment for the same day. Therefore, the doctor does not know the exact number of appointments she will have on a given day. We model this demand as D_{OA}, a random variable with distribution $\phi(\cdot)$. On the other hand, the doctor still asserts influence on patient demand by deciding the panel size, which is a number of unique patients for which a doctor is responsible, and the panel size essentially determines the average daily demand μ_{OA}. That is, D_{OA} is a function of μ_{OA}, which is a decision variable for the doctor. As in RC, because patients only make an appointment for the same day, we assume the absence of "no-shows" under open-access policy. This assumption highlights an additional benefit of the open-access policy: it is possible for the doctor to gain additional goodwill from her patients because the open-access policy minimizes the appointment delay. Our model can easily account for this extra benefit by using $r' \geq r$ for the margin per patient under the open-access policy.

In contrast to RC, we implicitly assume a negative impact from idleness in the open-access scheduling policy. In particular, RC minimize cost and assume that doctors

go home early if demand is low (with no cost penalty). In contrast, in our profit maximization model the doctor incurs an opportunity cost of lost revenue per appointment slot if the demand on a given day does not reach n, and incurs an overtime cost which is a quadratic function if the demand exceeds n. Since the daily patient demand is random, the optimal average daily demand maximizes the doctor's expected profit:

$$\max_{\mu_{OA}} \Pi_{OA} = E[rD_{OA}(\mu_{OA}) - c_{ot}(\theta_{OA})] = r\mu_{OA} - k_{ot}\int_{n}^{\infty}(y-n)^2\phi(y)dy \qquad (4.2)$$

As in RC, we assume that the number of patients D_{OA} requesting appointments under an open-access policy follows the binomial distribution, which converges to a Poisson distribution for arrivals. Consequently, the doctor's expected overtime cost is expressed as

$$\begin{aligned} E[k_{ot}[(D_{OA}-n)^+]^2] &= k_{ot}\sum_{s=n+1}^{\infty}(s-n)^2 p(s|\mu_{OA}) \\ &= k_{ot}\big[\mu_{OA}^2[1-P(n-2|\mu_{OA})] + \mu_{OA}[1-P(n-1|\mu_{OA})] \\ &\quad - 2n\mu_{OA}[1-P(n-1|\mu_{OA})] + n^2[1-P(n|\mu_{OA})]\big], \end{aligned} \qquad (4.3)$$

where $p(s|\mu_{OA})$ and $P(s|\mu_{OA})$ are the probability mass and cumulative distribution functions of the Poisson distribution with mean μ_{OA}, respectively.

The expected profit function for the open-access policy with Poisson Arrivals is

$$E[\Pi_{OA}] = r\mu_{OA} - E[k_{ot}[(D_{OA}-n)^+]^2]. \qquad (4.4)$$

4.3.3. Comparison of the Two Policies: Profitability and Number of Patients Served

When comparing the two policies, it is important to consider the optimal average number of patients served under each appointment policy. Figure 4.1 illustrates the hazard of comparing the two policies without considering how many patients were served. In this example, the optimal number of patients served is greater under the open-access policy than the traditional policy. When comparing the maximum profits of the two policies that occur at their respective optimal number of patients served, in this example we can see that the traditional policy outperforms the open-access policy. However, we see that for any patient volume greater than μ_{EQ}, the open-access policy results in higher profit than the traditional policy. We must compare the two policies at their optimal patient volumes to avoid a faulty comparison.

The optimal decisions of the doctor for both traditional and open-access policies are difficult to characterize analytically. In the next section, we gain insights into the doctor's decision by analyzing and comparing the two policies for a special case of a very friendly doctor (with $\gamma = 0$) and uniform (rather than Poisson) patient demand for the open-access policy. Although the uniform distribution is not an accurate representation of patient demand, it allows us to learn structurally about how the demand variability affects the desirability of the two policies both in the number of patients served and profit. In Section 4.5 we perform numerical experiments for the Poisson-demand case and find in these experiments that the structural insights from the uniform case are consistent with what we observe in the Poisson case.

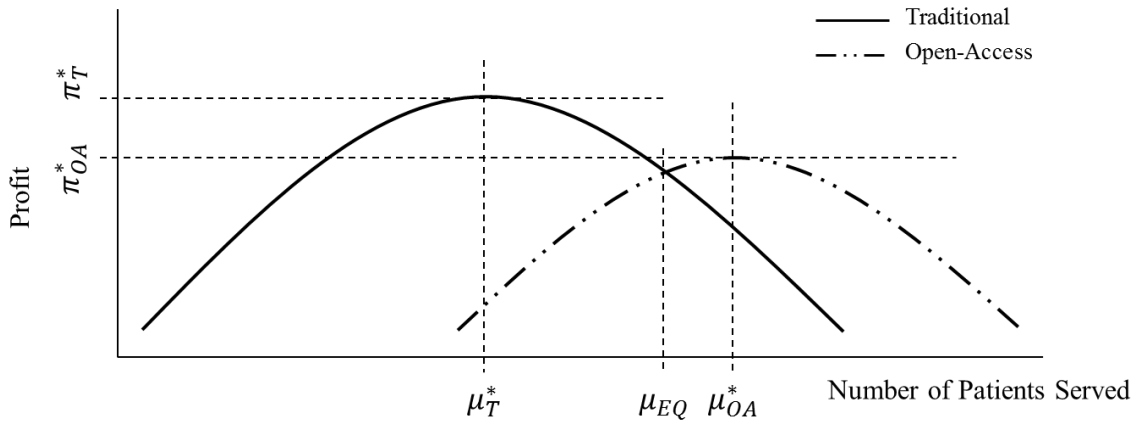


Figure 4.1. Example of when the Number of Patients Served is Greater under a Traditional Policy

4.4. A Patient-Friendly Doctor ($\gamma = 0$) and Uniform Demand under the Open-Access Policy

Assume that $\gamma = 0$ and that the daily patient demand under open-access scheduling policy is uniformly distributed between $\mu_{OA} - l$ and $\mu_{OA} + l$. The parameter l acts as a proxy for variability—the larger l, the more variation.

PROPOSITION 4.2. *Under uniform demand, the optimal average daily demand (which is directly related to the panel size) under open-access scheduling policy is:*

$$n + \frac{r}{2k_{ot}} \qquad \text{for } n \leq \mu_{OA} - l, \text{ and}$$

$$n - l + \sqrt{\frac{2lr}{k_{ot}}} \quad \text{for } \mu_{OA} - l \leq n \leq \mu_{OA} + l.$$

Proposition 4.2 shows that the optimal average number of patients the doctor sees in a day under open access decreases with variability of the demand and the overtime cost parameter. Note that Proposition 4.2 does not include the case where $\mu_{OA} + l < n$ because it is never optimal for μ_{OA} to be less than $n - l$.

We now compare the two scheduling policies on their performances at their optimal patient volumes when the daily patient demand under open-access scheduling policy is uniformly distributed. For accurate comparison, we compare them at their optimal number of patients served (for maximum profits). For simplicity of exposition, define:

$$\xi \equiv np + \frac{r}{k_{ot}} - \frac{r(1-p)^2}{2(k_{ot}+k_{gw})} + \sqrt{\frac{r[2k_{ot}(k_{ot}+k_{gw})np + r(k_{gw}+2k_{ot}p-k_{ot}p^2)]}{k_{ot}^2(k_{ot}+k_{gw})}} \quad (4.5)$$

PROPOSITION 4.3. *When the daily patient demand under an open-access scheduling policy is uniformly distributed and* $\gamma = 0$, *the optimal number of patients served is greater under an open-access policy than under a traditional policy when* l *is less than* ξ .

When demand variability is large under an open-access policy, the doctor will be more conservative and thus prefer lower average daily demand because of the increasing marginal cost of overtime. Consequently, the optimal number of patients served is greater under a traditional policy if the demand variability is relatively large. On the other hand, when demand variability is small, under an open-access policy the doctor would serve more patients on average. In this case it is much more likely for the realized daily demand to be equal or very close to the optimal average number of patients served that the doctor desires under open access, whereas a traditional policy continues to force her to be more conservative because of the uncertainty in patient no-shows. Figure 4.2 shows an example of the value of threshold ξ for a case with $l = 6$. It can be seen that the threshold ξ is greater than l for the majority of the values of r and k_{ot} shown in the figure. Note that $l = 6$ is comparable in variability to Poisson demand with $\mu_{OA}^* = 12$ (equal to the number of slots n). Thus, for this moderate level of demand variability, more

patients are served on average under the open-access policy compared to the traditional policy. Figure 4.2 also shows that the optimal number of patients served under the open-access policy is likely to be higher than that of the traditional policy when revenue per patient is high relative to the overtime cost parameter. This is consistent with our results under Poisson demand, shown in Section 4.5.

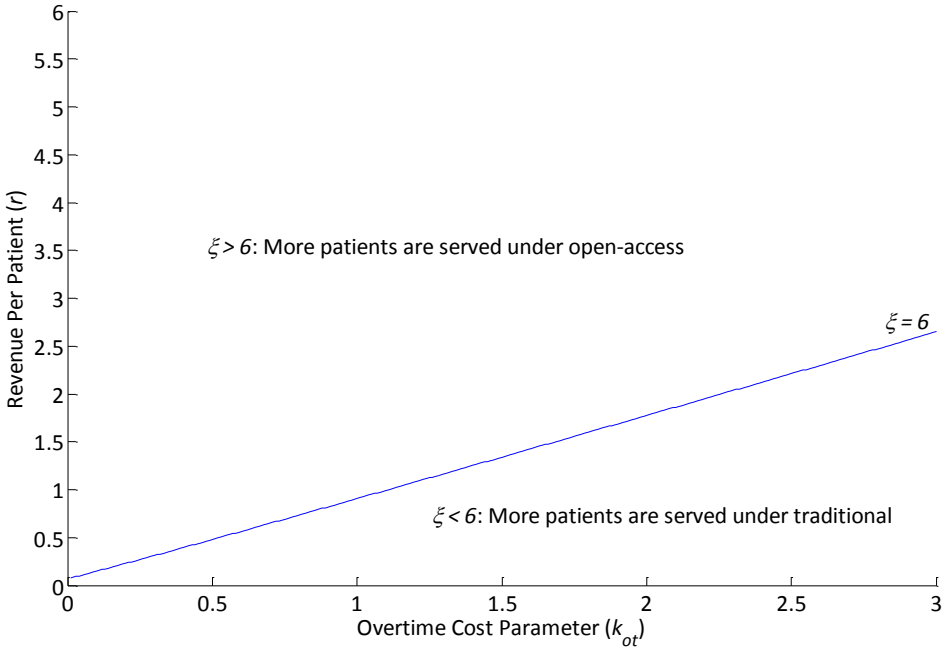


Figure 4.2. Number of Patients Served: Threshold ξ ($p = 0.25$, $n = 12$, $k_{gw} = 0.5$, $l = 6$)

We next compare the profits under both policies to examine which policy is more preferable for doctors. Proposition 4.4 shows there exists a threshold degree of demand variability that determines which policy is more profitable. Smaller values of l again lead to a preference for open-access but there remain many regions where a traditional policy performs better than an open-access policy. This result is in contrast to the conclusion of RC (that open-access dominates) and arises because we compare the two policies at their

respective optimal number of patients served instead of requiring the average number of patients served to be equal.

Define:

$$\alpha \equiv np + \frac{4r}{9k_{ot}} - \frac{r(1-p)^2}{4(k_{ot}+k_{gw})} + \frac{1}{9k_{ot}}\sqrt{\frac{72k_{ot}rnp(k_{ot}+k_{gw})+16r^2k_{gw}-2r^2k_{ot}[1-9p(2-p)]}{k_{ot}+k_{gw}}} \quad (4.6)$$

$$\beta \equiv \frac{1}{2k_{ot}}\sqrt{\frac{[12k_{ot}rnp(k_{ot}+k_{gw})+3r^2[k_{gw}+k_{ot}p(2-p)]}{k_{ot}+k_{gw}}}, \text{ and} \quad (4.7)$$

$$\tau \equiv -12k_{ot}np(k_{ot}+k_{gw}) - 2k_{gw}r + k_{ot}r - 3k_{ot}rp(2-p). \quad (4.8)$$

PROPOSITION 4.4. *When the daily patient demand under open-access scheduling policy is uniformly distributed with $l > 0$ and $\gamma = 0$, at the respective optimal patient volumes, the open-access scheduling policy is more profitable than the traditional scheduling policy if:*

(i) $l < \alpha$ *for* $\tau \leq 0$, *and*

(ii) $l < \beta$ *for* $\tau > 0$.

Proposition 4.4 shows that the doctors prefer the traditional policy when there is moderate variability in patient demand under the open-access policy. Figure 4.3 shows an example of the relative profitability of open access or traditional based on whether the thresholds α and β are greater than or less than 6 when $l = 6$. (α is used when $\tau \leq 0$ and β is used when $\tau > 0$.) The graph of the profit threshold of the two policies has similarities to that of patient volume that we saw in Figure 4.2. In particular, Figure 4.3 shows that the open-access policy is more profitable when revenue per patient is relatively high and overtime cost is relatively low. However, Figure 4.2 and Figure 4.3 show that the threshold for number of patients served is generally higher than the profit

threshold for the same parameter values of r and k_{ot}. Thus, it is possible that, for a given scenario, more patients are served on average under the open-access policy while the traditional policy is more profitable for the doctor.

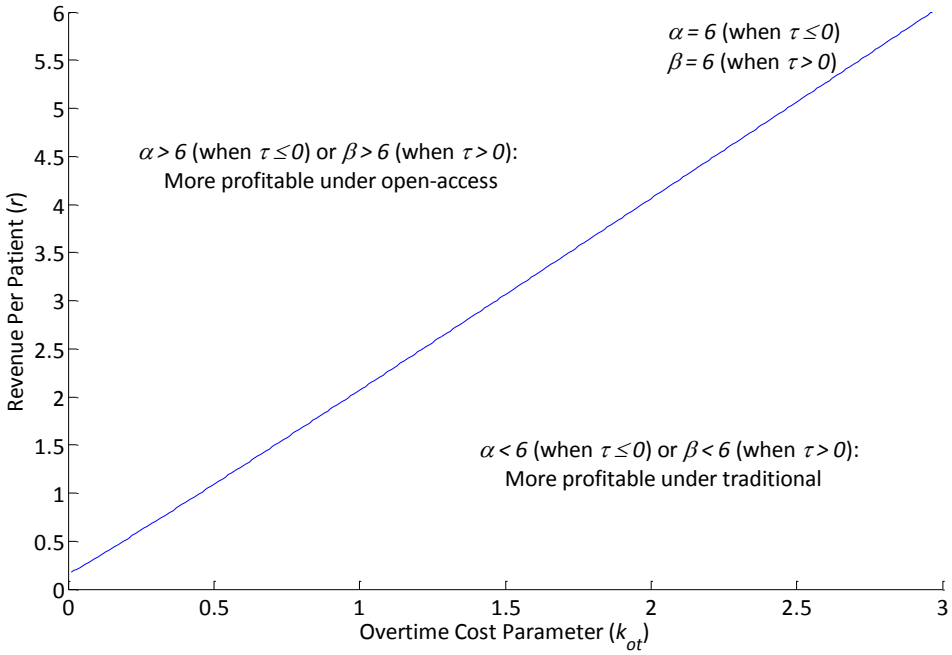


Figure 4.3. Profits: Threshold α (when $\tau \leq 0$) or β (when $\tau > 0$) ($p = 0.25$, $n = 12$, $k_{gw} = 0.5$, $l = 6$)

Propositions 4.3 and 4.4 assume a uniform distribution for patient demand under an open-access policy. We observe from numerical results in the next section that the insights from the propositions continue to hold when we use Poisson arrivals for the demand under an open-access policy.

4.5. Numerical Comparison of Traditional and Open-Access Policies with Poisson Arrivals

In this section, we use numerical experiments to compare the two appointment scheduling policies under Poisson demand for open access. We use a wide range of

parameter values for revenue and costs to account for a broad range of circumstances for doctors. Revenue per patient is considered relative to the cost of overtime and loss of goodwill. We consider values for the revenue per patient r from 0.1 to 6 and values for the overtime cost parameter k_{ot} from 0.1 to 3 and we fix the loss of goodwill cost parameter k_{gw} at 0.5. For all experiments, we assume the patient no-show rate is 25%. A regular working day for doctors consists of 12 equal-length time slots and thus any patients seen by the doctor after slot 12 is considered as overtime. Table 4.1 presents the experimental design for various parameters.

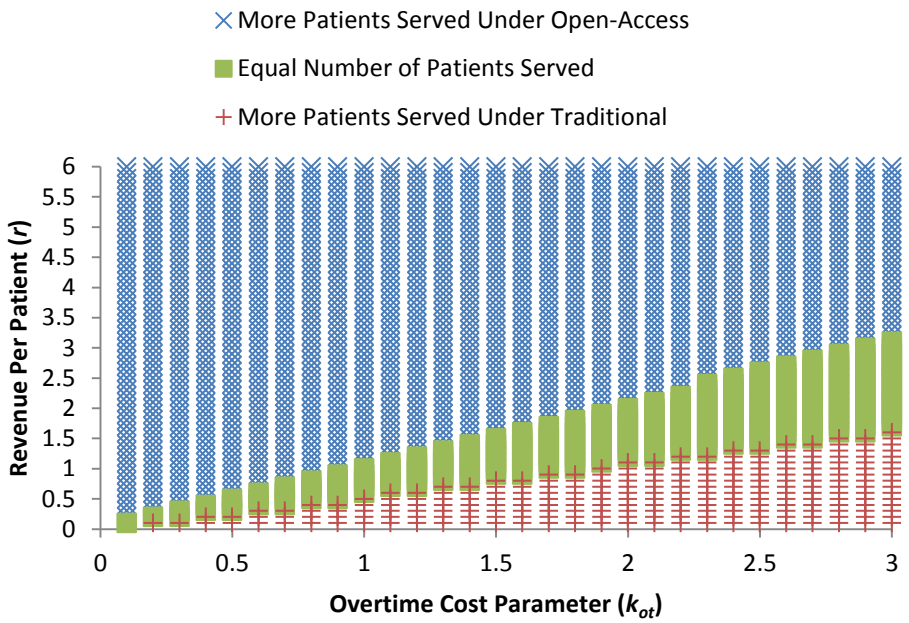
Parameter	**Symbol**	**Values**
Patient no-show rate	p	25%
Number of regular slots	n	12
Revenue per patient	r	0.1 – 6
Overtime cost parameter	k_{ot}	0.1 – 3
Loss of goodwill cost parameter	k_{gw}	0.5
Patient friendliness	γ	0 – 0.3

Table 4.1. Experimental Design

We first compare the number of patients served at optimality for traditional and open-access policies for various parameter values for overtime cost and revenue per patient. We present results for $\gamma = 0$ and 0.3 to compare doctors with different levels of patient friendliness. Figure 4.4 confirms that the optimal numbers of patients served under the two policies are generally different and underscores the importance of comparing the policies at their optimal patient volumes. Also, Figure 4.4 shows that more patients are generally served under an open-access policy when revenue per patient is high relative to the cost of overtime, while more are served under a traditional policy when overtime is very costly. A traditional policy allows the doctor to set an upper limit

on the number of patients she sees on a given day (and thus caps overtime) while an open-access policy does not provide the doctor any control over her daily demand once she decides on her panel size. This control provided by the traditional policy is beneficial when overtime is very expensive but can limit profitability when each patient served represents high profit margin (revenue). Consequently, as overtime becomes more expensive, a traditional policy allows the doctor to serve more patients because under an open-access policy, the doctor needs to be more cautious in setting her average daily demand to prevent potential overtime. On the other hand, an unexpectedly high number of patients is more desirable when revenue per patient is high and the overtime cost parameter is low, and thus the open-access policy incentivizes the doctor to serve more patients. As γ increases, the risk of overtime decreases due to the doctor's utilization of overbooking, which allows the doctor to serve more patients on average in same number of appointment slots by reducing the idle time. Higher values of γ offset (to a certain extent) the main benefit provided by the open-access policy—the elimination of patient no-shows and resulting doctor's idle time—while enabling the doctor to still enjoy the ability of the traditional policy to set an upper limit on the number of patients served. While an open-access policy provides the doctor less control over her daily demand, more patients are served on average under the open-access policy than the traditional policy even when $\gamma = 0.3$, as shown in Figure 4.4. The greater number of patients served and the reduced patient wait time make the open-access policy attractive to policy makers and patients.

(a) $\gamma = 0$



(b) $\gamma = 0.3$

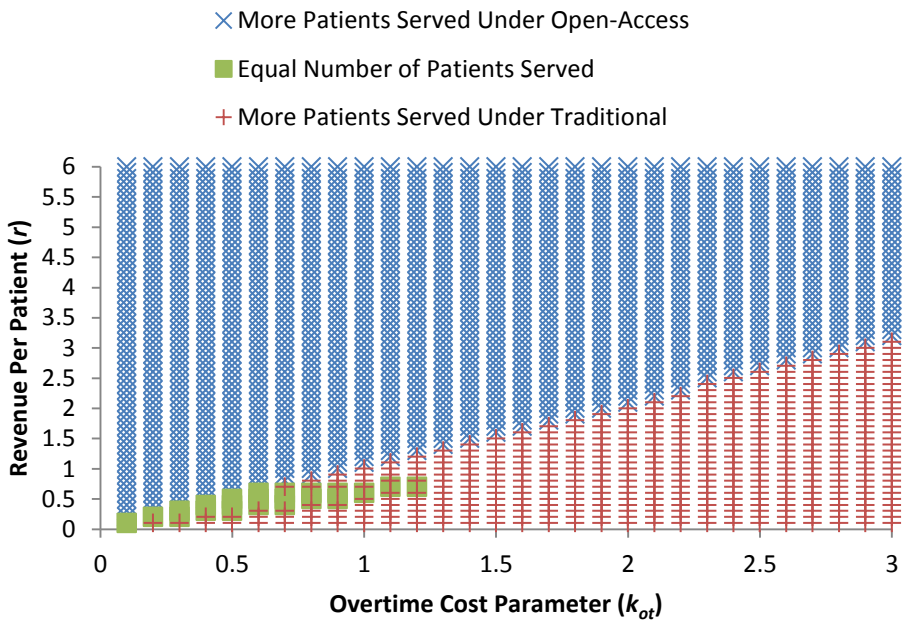


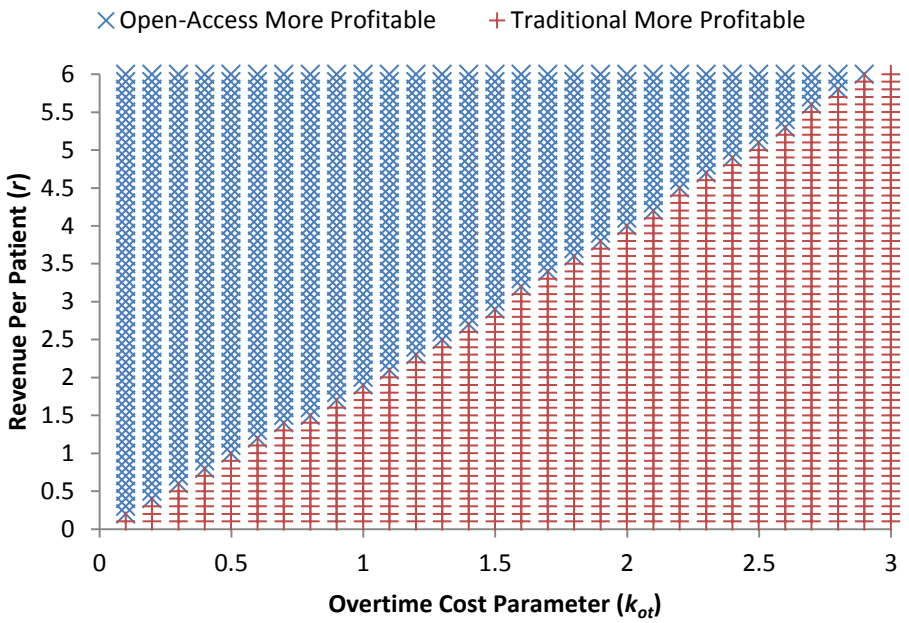
Figure 4.4. Comparison of Patients Served

Given the optimal number of patients served for both scheduling policies, we now compare the profits of the two scheduling policies at their optimal patient volumes.

Figure 4.5 shows that the traditional policy is more profitable when revenue per patient is relatively low and overtime cost is relatively high, similar to the pattern we saw in Figure 4.4 for the number of patients served. When revenue per patient is high and overtime is not expensive, the advantage an open-access policy provides in eliminating the doctor's idle time caused by patient no-shows and allowing the doctor to see more patients during the normal working hours becomes magnified. At the same time, its weakness in controlling potential overtime is not as important as it is when overtime is more expensive.

Figure 4.5 illustrates a consideration for policy makers as they strive to encourage doctors to employ an open-access policy and serve more patients: because the inability of the open-access policy to control overtime is its major drawback, the policy maker might consider various initiatives that lower the cost of overtime for doctors. Figure 4.6 further emphasizes this opportunity by highlighting the difference between Figure 4.4 and Figure 4.5, which is observed in the boundaries—for parameters in the center of the graphs there are many scenarios where more patients are served under open access but traditional is more profitable. Figure 4.6 shows that as long as cost of overtime is moderate relative to revenue per patient, doctors and patients would have different preferences for appointment scheduling policy. As seen in panel (b), the number of scenarios with different preferences for doctors and patients is even greater with less patient-friendly doctors.

(a) $\gamma = 0$



(b) $\gamma = 0.3$

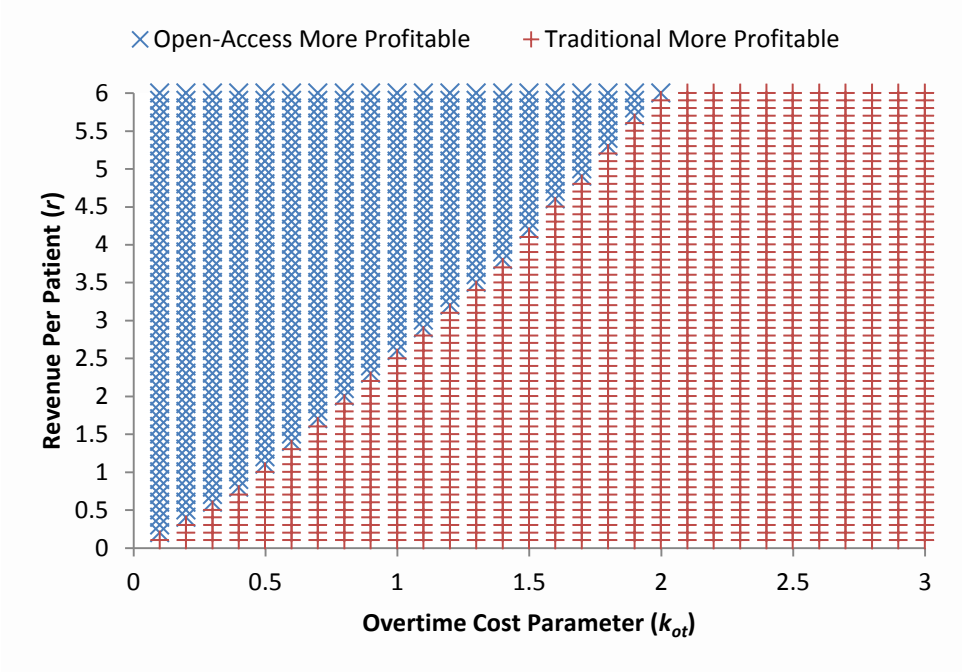
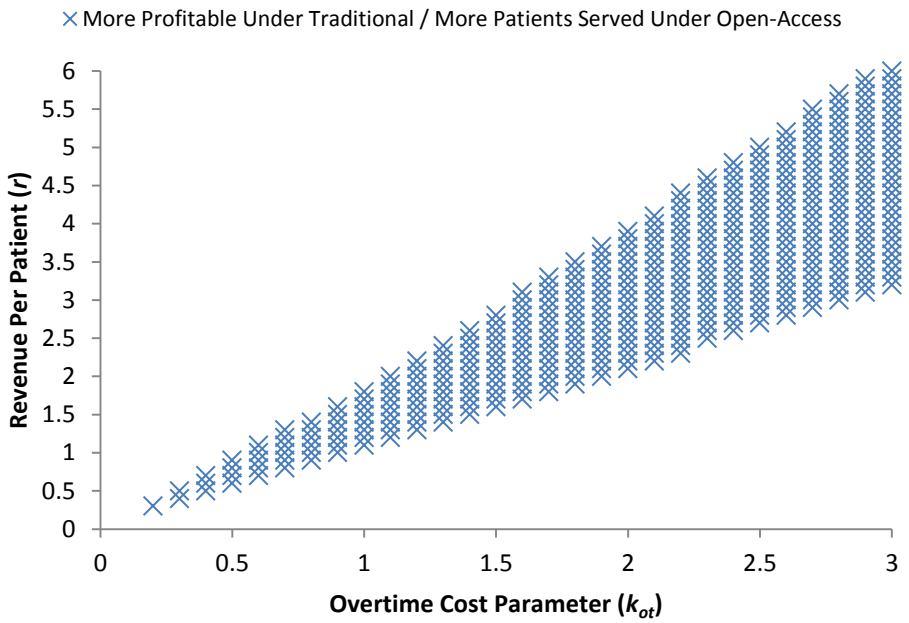


Figure 4.5. Comparison of Profits

(a) $\gamma = 0$



(b) $\gamma = 0.3$

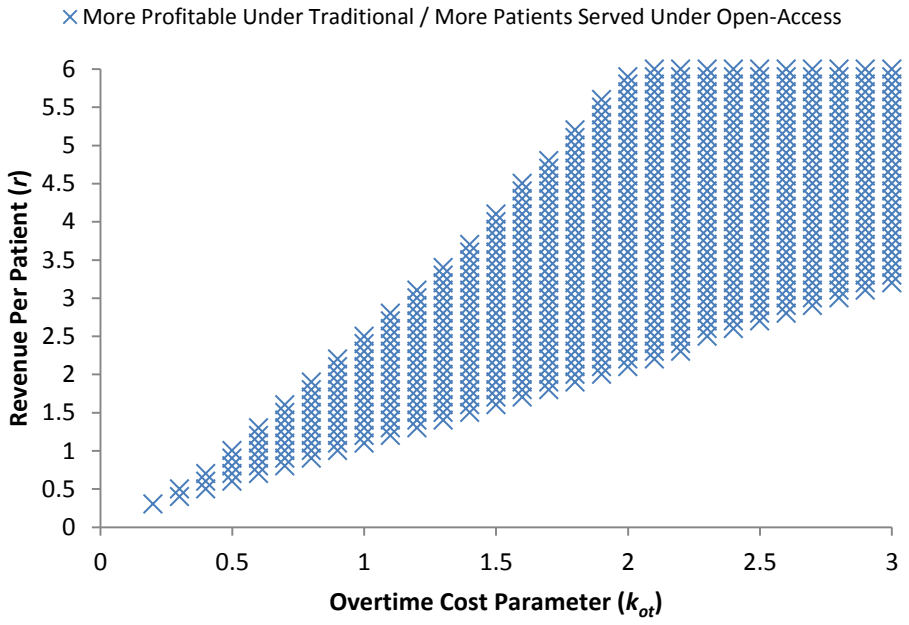
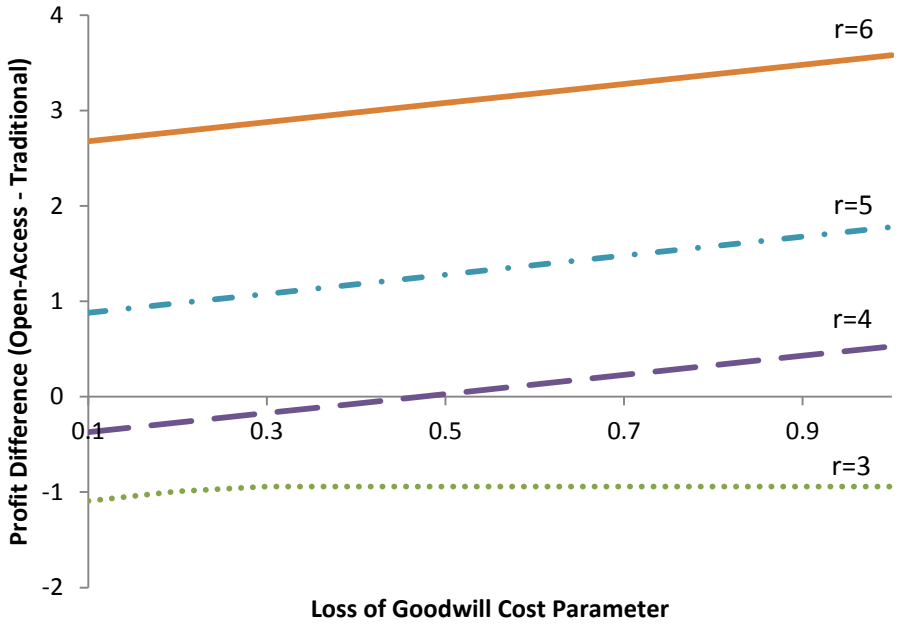


Figure 4.6. Traditional Policy More Profitable while Open-Access Serves More Patients

(a) $\gamma = 0$



(b) $\gamma = 0.3$

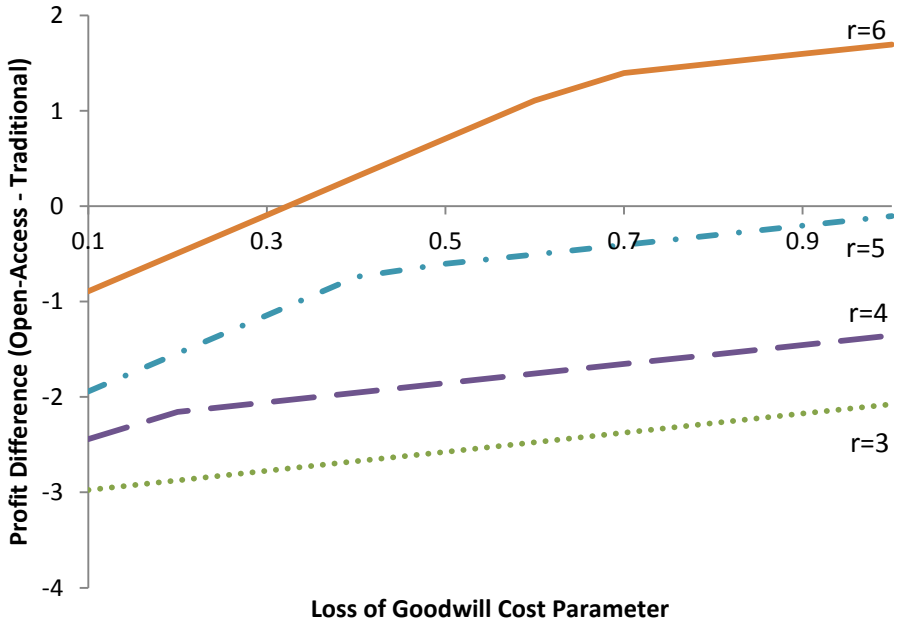


Figure 4.7. Comparison of Profits ($k_{ot} = 2$)

Figure 4.7 suggests another potential approach for policy makers. Doctors would start to favor the open-access policy over the traditional policy as the cost of goodwill

becomes more expensive. This suggests the importance of ensuring that doctors understand and account for (or even explicitly charged for) patient goodwill costs. Also, Figure 4.7 shows that it becomes much more difficult to incentivize doctors who are less patient-friendly to employ an open-access policy. Even for a relatively low value of $\gamma = 0.3$, the traditional policy outperforms the open-access policy for most of the cases shown in the figure. This result signifies the challenges for the policy makers and potentially explains why most doctors still employ the traditional policy while the open-access becomes more popular among the general public. In order to increase the incentive for doctors to choose open access, the doctors must incur higher costs from patient waiting—policy makers might want to consider ways to make these costs more visible or explicit.

4.6. Conclusion

We compare two appointment scheduling policies in the health care industry to gain insights on their overall performance and impact on the number of patients served. Under a traditional scheduling policy, a patient makes an appointment well in advance of his or her preferred time, but there is a possibility that the patient will not show up. If a patient does not arrive for his or her appointment, there is a possibility that the doctor remains idle and incurs an opportunity cost and idle cost. Alternatively, the doctor has an option of overbooking which, although not perfect, can reduce the adverse impact of patient no-shows. The doctor decides how many patients to schedule in a given day. Under an open-access policy, a patient makes an appointment in the morning of his or her preferred day, thereby eliminating the possibility of no-shows. On the other hand, the doctor loses some control over the patient demand because the policy requires the doctor to see every

patient on the day he or she calls in. Although the doctor is not able to decide the exact demand of each day, she still asserts influence on patient demand by deciding her panel size, which essentially determines the average daily demand.

By analyzing the optimal average number of patients served under both scheduling policies, we show that the traditional policy can allow doctors to maintain a higher profit for more settings while the open-access policy incentivizes doctors to serve more patients. Our results provide insights into a current challenge of the U.S. health care industry: society and policy makers prefer an open-access policy, but doctors do not necessarily have incentive to switch from a traditional policy. The traditional policy is more efficient in controlling a doctor's overtime (although it suffers from patient no-shows) because it allows the doctor more control over daily patient demand. The open-access policy is subject to more variability in daily patient arrivals because it guarantees same-day service for all the patients who call in the morning. The traditional policy requires patient patients who face a larger lead time to see the doctor; the open-access policy requires patient doctors who have less control over schedule variability. Many doctors view their own time as more valuable than patients' time, especially in the context of a shortage of doctors, and it is also likely that the marginal cost of overtime increases at a fairly high rate. Because of the ability to better control the overtime and the value of the doctors' time, many doctors may find the traditional policy more profitable than the open-access policy, capitalizing on "patient patients". Consequently, for our society to achieve its optimum, policy makers should strive to provide appropriate incentives and inducements to doctors so that they would be willing to serve more patients through the open-access policy.

Possible extensions to our current work might explore hybrid scheduling policies for doctors, in which she would set aside certain appointment slots for advance booking while reserving the rest for same-day appointments or urgent walk-ins (Dobson et al. 2011). It would be interesting to investigate the optimal strategies for building hybrid policies, and answer questions such as how many and which slots in a day to set aside for a particular policy. It is probable that many doctors in reality are already using some form of hybrid policy, and this extension would provide valuable insights in achieving both goals of higher profit and improved service.

CHAPTER 5

CONCLUSION

This dissertation showed that it is important for firms to take behavior, quality, and flexibility into consideration when making capacity planning decisions for their service systems. Capacity planning decisions and the firm's utilization of flexible capacity present in the system have a significant impact on the quality of the work and may also change the service rate of the workers through speedup and slowdown. In Chapter 2, we demonstrated that it is possible for hospitals to decrease total costs by increasing their staffing because lower workload for nurses leads to better quality of care and decreases the costs incurred by adverse patient and nurse outcomes. In Chapter 3, we showed that the optimal workload generally varies in the number of customer requests in the system when the behavioral impacts of the capacity planning decision are incorporated. The workers' tendency to slow down when understaffed causes the firm to aggressively utilize expensive on-call workers, and their tendency to speed up incentivizes the firm to send some workers home even when it does not recoup any wages for unworked time. In Chapter 4, we compared two appointment scheduling policies with different levels of flexibility and provided a possible explanation to the current situation, in which most of the doctor's offices employ a traditional policy, which generally is more profitable for the doctor, while the patients and society prefer the open-access policy, which is likely to serve more patients.

While we studied the impacts of capacity planning decisions on workers' behavior and quality of outcome separately, considering these important attributes concurrently and studying the correlation between the behavior, flexibility, and quality in

service capacity planning decisions are natural and promising directions for future research.

Appendix A Proofs for Chapter 2

Proof of Proposition 2.1.

$$F_1 = c_s n + \sum_{i=1}^{m} (\beta_i^{r-r_b}) \gamma_i \lambda c_i + c_A \int_{nr}^{\infty} \left(\frac{t}{r} - n\right) \phi(t) dt$$

$$\frac{\partial F_1}{\partial n} = c_s - c_A \left[\int_{nr}^{\infty} \phi(t) dt\right] = c_s - c_A[1 - \Phi(\mathrm{nr})] = \mathrm{c_s} - c_A \overline{\Phi}(nr)$$

$$\frac{\partial^2 F_1}{\partial n^2} = c_A r \phi(nr) \geq 0$$

$$\frac{\partial^2 F_1}{\partial n \partial r} = c_A n \phi(nr)$$

$$\frac{\partial F_1}{\partial r} = \sum_{i=1}^{m} (\beta_i^{r-r_b}) \gamma_i \lambda c_i \ln \beta_i - c_A \left[\int_{nr}^{\infty} \frac{t}{r^2} \phi(t) dt\right]$$

$$\frac{\partial^2 F_1}{\partial r^2} = \sum_{i=1}^{m} (\beta_i^{r-r_b}) \gamma_i \lambda c_i (\ln \beta_i)^2 + c_A \left[\frac{n^2}{r} \phi(nr) + \int_{nr}^{\infty} \frac{2t}{r^3} \phi(t) dt\right] \geq 0$$

$$H = \begin{bmatrix} c_A r \phi(nr) & c_A n \phi(nr) \\ c_A n \phi(nr) & \sum_{i=1}^{m} (\beta_i^{r-r_b}) \gamma_i \lambda c_i (\ln \beta_i)^2 + c_A \left[\frac{n^2}{r} \phi(nr) + \int_{nr}^{\infty} \frac{2t}{r^3} \phi(t) dt\right] \end{bmatrix}$$

Since

$$c_A r \phi(nr) \times \left\{ \sum_{i=1}^{m} (\beta_i^{r-r_b}) \gamma_i \lambda c_i (\ln \beta_i)^2 + c_A \left[\frac{n^2}{r} \phi(nr) + \int_{nr}^{\infty} \frac{2t}{r^3} \phi(t) dt\right] \right\}$$

$$\geq c_A r \phi(nr) \times c_A \frac{n^2}{r} \phi(nr) = [c_A n \phi(nr)]^2$$

$$|H| \geq 0$$

Therefore, function F_1 is jointly convex in n and r. ■

Proof of Proposition 2.2.

$$F_1 = \frac{2c_s c_A \bar{g} - c_s^2(\bar{g} - \underline{g})}{2c_A r} + \sum_{i=1}^{m} (\beta_i^{r-r_b}) \gamma_i \lambda c_i$$

$$\frac{\partial F_1}{\partial r} = \frac{-2c_s c_A \bar{g} + c_s^2(\bar{g} - \underline{g})}{2c_A r^2} + \sum_{i=1}^{m} (\beta_i^{r-r_b}) \gamma_i \lambda c_i \ln \beta_i = 0$$

$$r^2 = \frac{2c_s c_A \bar{g} - c_s^2(\bar{g} - \underline{g})}{2c_A \sum_{i=1}^{m} (\beta_i^{r-r_b}) \gamma_i \lambda c_i \ln \beta_i}$$

Since $\beta_i \geq 1$ for all i, $\beta_i^{r-r_b}$ would always be greater than or equal to 1 for $r^* \geq r_b$. Therefore,

$$r^{*2} \leq \frac{2c_s c_A \bar{g} - c_s^2(\bar{g} - \underline{g})}{2c_A \sum_{i=1}^{m} \gamma_i \lambda c_i \ln \beta_i}$$

$$r^* \leq \sqrt{\frac{2c_s c_A \bar{g} - c_s^2(\bar{g} - \underline{g})}{2c_A \sum_{i=1}^{m} \gamma_i \lambda c_i \ln \beta_i}}$$

Also note that because agency nurses typically earn at least as much as unit nurses (i.e., $c_A \geq c_s$), we know that $2c_s c_A \bar{g} - c_s^2(\bar{g} - \underline{g}) \geq 0$. Thus, Proposition 2.2 follows. ■

Proof of Proposition 2.3. Proposition 2.3 is easily verified through the first and second derivatives of (2.17) with respect to r. ■

Proof of Proposition 2.4. Proposition 2.4 is easily verified through the first and second derivatives of (2.19) with respect to r. ■

Appendix B More Details and Proofs for Chapter 3

Appendix B.1. Notations

r	Workload (ratio of requests to servers)
$\bar{r}$	Maximum allowable workload
$\gamma(r)$	Speedup effect as a function of workload
$\tau(r)$	Slowdown effect as a function of workload
$\mu(r)$	Joint effects of speedup and slowdown as a function of workload
$\bar{\mu}_{sp}$	Upper bound of the service rate for speedup effect
$\bar{\mu}_{sl}$	Upper bound of the service rate for slowdown effect
θ_{sp}	Degree of speedup effect
θ_{sl}	Degree of slowdown effect
r_{sp}	Level of workload where a service rate of $\bar{\mu}_{sp}/2$ is achieved
r_{sl}	Level of workload where a service rate of $\bar{\mu}_{sl}/2$ is achieved
β	Weight for speedup effect
$1-\beta$	Weight for slowdown effect
α	One-shift discount rate
k	Number of shifts in a cycle
z_s	Initial number of workers assigned to work in shift s
c_w	Wage per worker per shift
c_v	Variable cost per shift for each request in service
c_h	Holding cost per backlogged unit per shift
b	Service capacity of the system
Y	Number of service requests in the system
X	Number of requests that are placed in service (i.e., not backlogged)
ψ	On-call premium relative to the regular worker wage
ϕ	Proportion of wage the firm can recoup for workers sent home
D	Number of requests that depart the system at the end of a given shift
A	Number of new arrivals of requests in a given shift
λ	Expectation of A

Appendix B.2. Finite-Horizon Model

Throughout Chapter 3, we consider the finite-horizon version of the model presented in Section 3.3 to prove various results. Thus, we formally define the finite-horizon problem with t cycles remaining. (We denote the first cycle as cycle t and last cycle as cycle 0.)

Denote by $V_{st}(y)$ the minimum expected total discounted cost when the system starts with a y requests in shift s of cycle t. Each cycle represents a cycle with k shifts. $V_{st}(y)$ must satisfy the following optimality equation:

$$V_{st}(y) = \min_{r\geq 0}\Bigg\{c_w z_s + c_v \min\{y,b\} + c_h(y-b)^+ + (1+\psi)c_w\left(\frac{\min\{y,b\}}{r} - z_s\right)^+ - \phi c_w\left(z_s - \frac{\min\{y,b\}}{r}\right)^+ + \alpha E\left[V_{\oplus(st)}\left(y + A_s - D(y,r)\right)\right]\Bigg\},$$

$$\oplus\,(st) = \begin{cases} s+1, t & \text{if } s < k \\ 1, t-1 & \text{if } s = k \end{cases}$$

where $D(y,r)$~Binomial$\left(y\,,\mu(r)\right)$, and with the convention that $V_{s(-1)}(y) = 0$ for all y.

We also define the finite horizon version of the problem with t cycles remaining for shift-by-shift staffing with no capacity constraint case presented in Subsection 3.4.3. Denote by $V_{st}(y)$ the minimum expected total discounted cost when the system starts with a y requests in shift s of cycle t. Each cycle has k shifts. $V_{st}(y)$ must satisfy the following optimality equation:

$$V_{st}(y) = \min_{r\geq 0}\left\{c_w\frac{y}{r} + c_v y + \alpha E\left[V_{\oplus(st)}\left(y + A_s - D(y,r)\right)\right]\right\},$$

$$\oplus\,(st) = \begin{cases} s+1, t & \text{if } s < k \\ 1, t-1 & \text{if } s = k \end{cases}$$

where $D(y,r)$~Binomial$\left(y\,,\mu(r)\right)$, and with the convention that $V_{s(-1)}(y)=0$ for all y.

Appendix B.3. Proofs

Proof of Lemma 3.1. The first derivative of $\mu(r)$ is

$$\mu'(r) = \theta\left[\frac{\bar{\mu}_{sp}e^{\theta(r_{sp}+r)}\beta}{\left(e^{\theta r_{sp}}+e^{\theta r}\right)^2} - \frac{\bar{\mu}_{sl}e^{\theta(r_{sl}+r)}(1-\beta)}{(e^{\theta r_{sl}}+e^{\theta r})^2}\right],$$

which is positive when $\frac{(1-\beta)}{\beta} < \frac{\left(e^{\theta r_{sl}}+e^{\theta r}\right)^2\bar{\mu}_{sp}}{\left(e^{\theta r_{sp}}+e^{\theta r}\right)^2 e^{\theta(r_{sl}-r_{sp})}\bar{\mu}_{sl}}$. ■

Proof of Lemma 3.2. The first derivative of $\mu(r)$ is negative when $\frac{(1-\beta)}{\beta} > \frac{\left(e^{\theta r_{sl}}+e^{\theta r}\right)^2\bar{\mu}_{sp}}{\left(e^{\theta r_{sp}}+e^{\theta r}\right)^2 e^{\theta(r_{sl}-r_{sp})}\bar{\mu}_{sl}}$. ■

Proof of Lemma 3.3. We need to show that when $\mu(r)$ is not monotonic, the function is unimodal with a single local maximum when $r_{sp} < r_{sl}$ and unimodal with a single local minimum when $r_{sp} > r_{sl}$. To show unimodality, we need to show that for some workload r_0, $\mu(r)$ is monotonically increasing (decreasing) for $r \le r_0$ and monotonically decreasing (increasing) for $r \ge r_0$ when $r_{sp} < r_{sl}$ ($r_{sp} > r_{sl}$). The first derivative of $\mu(r)$ is

$$\mu'(r) = \theta\left[\frac{\bar{\mu}_{sp}e^{\theta(r_{sp}+r)}\beta}{\left(e^{\theta r_{sp}}+e^{\theta r}\right)^2} - \frac{\bar{\mu}_{sl}e^{\theta(r_{sl}+r)}(1-\beta)}{(e^{\theta r_{sl}}+e^{\theta r})^2}\right],$$

which is positive when $e^{\theta(r_{sl}-r_{sp})}\frac{\bar{\mu}_{sl}(1-\beta)}{\bar{\mu}_{sp}\beta} < \frac{\left(e^{\theta r_{sl}}+e^{\theta r}\right)^2}{\left(e^{\theta r_{sp}}+e^{\theta r}\right)^2}$ and negative when $e^{\theta(r_{sl}-r_{sp})}\frac{\bar{\mu}_{sl}(1-\beta)}{\bar{\mu}_{sp}\beta} > \frac{\left(e^{\theta r_{sl}}+e^{\theta r}\right)^2}{\left(e^{\theta r_{sp}}+e^{\theta r}\right)^2}$. The left side of the inequality is a constant while the right-side decreases with r when $r_{sp} < r_{sl}$ and increases with r when $r_{sp} > r_{sl}$. Therefore,

the function will always be decreasing or unimodal when $r_{sp} < r_{sl}$ and will always be increasing or unimodal when $r_{sp} > r_{sl}$, and the result follows. ■

Lemma B.3.1, proved in Ross (1983), enables us to conclude that the results for finite-horizon problem apply to the infinite-horizon version of the model as well. We use Lemma B.3.2 to prove Propositions 3.1 and 3.2.

LEMMA B.3.1. (Ross 1983; Chapter 2.3, Proposition 3.1). $V_{st}(y) \to V_s(y)$ uniformly as $t \to \infty$.

Proof of Lemma B.3.1. For a proof, see Ross (1983). ■

LEMMA B.3.2. $V_s(\cdot)$ is an increasing function.

Proof of Lemma B.3.2. We prove by induction over $t \geq 0$ on the finite horizon version of the problem presented in Appendix B.2. We first note that the base case is increasing in the number of requests since $V_{k0}(y) = \min_{r\in(0,\bar{r}]}\left\{c_w z_s + c_v \min\{y,b\} + c_h(y-b)^+ + (1+\psi)c_w\left(\frac{\min\{y,b\}}{r} - z_s\right)^+ - \phi c_w\left(z_s - \frac{\min\{y,b\}}{r}\right)^+\right\}$ and thus $r^* = \bar{r}$. Because $x = \min\{y,b\}$ is nondecreasing in y, $V_{k0}(\cdot)$ is an increasing function since it is a sum of expressions that are either nondecreasing or increasing. For the induction, assume $V_{\oplus(st)}(\cdot)$ is an increasing function. Then $E\left[V_{\oplus(st)}(\cdot)\right]$ is an increasing function, and $V_{st}(\cdot)$ is a sum of expressions that are either nondecreasing or increasing, making it an increasing function as well. By using Lemma B.3.1, we conclude that $V_s(\cdot)$ is an increasing function. ■

Proof of Proposition 3.1. When speedup dominates slowdown, or when $\mu(r)$ is U-shaped and $\mu(\bar{r}) \geq \mu(r)$ for all $r \in (0, \bar{r}]$, $D(\min\{y, b\}, \bar{r})$ first-order stochastically dominates $D(\min\{y, b\}, r)$ for $r < \bar{r}$. Because Lemma B.3.2 states that $V_s(\cdot)$ is an increasing function and thus future cost is increasing in the number of requests, the expected future costs for $r = \bar{r}$ is less than that of $r < \bar{r}$. Furthermore, staffing costs for the current shift is lower for $r = \bar{r}$ than that of $r < \bar{r}$. Therefore, we conclude that $r_s^* \geq \bar{r}$. ■

Proof of Proposition 3.2. When slowdown dominates speedup, the result automatically follows since $r_0 = 0$. For inverse U-shape, we know that, by Lemma B.3.2, future cost is increasing in the number of requests. When $\mu(r)$ is inverse U-shape as defined in Definition 3.3, increasing $r < r_0$ would decrease staffing costs while increasing the expected number of departures since $D(\min\{y, b\}, r_0)$ first-order stochastically dominates $D(\min\{y, b\}, r)$ for $r < r_0$. Thus, the expected future costs for $r = r_0$ is less than that of $r < r_0$, and we conclude that $r_s^* \geq r_0$. ■

Lemmas B.3.3 and B.3.4 are supporting lemmas for the case of shift-by-shift staffing with no capacity constraint presented in Subsection 3.4.3. We use Lemmas B.3.1, B.3.3, and B.3.4 to prove Proposition 3.3.

LEMMA B.3.3. Suppose that $\psi = 0$, $\phi = 1$, $b = \infty$, and $V_{\oplus(st)}(\cdot)$ is a linear function. Then r_{st}^* is independent of y.

Proof of Lemma B.3.3. We proceed by considering the finite horizon version of the problem with t cycles remaining presented in Appendix B.2. In cycle t, we determine $r_{st}^*(y)$ by minimizing $c_w \frac{y}{r} + c_v y + \alpha E\left[V_{\oplus(st)}\left(y + A_s - D(y, r)\right)\right]$. Because the

expectation operator is linear and we assumed $V_{\oplus(st)}(\cdot)$ is a linear function, this is equivalent to minimizing $c_w\frac{y}{r}+c_v y+\alpha V_{\oplus(st)}(y+E[A_s]-E[D(y,r)])$, which by applying the definition of A_s and D, is equivalent to minimizing $c_w\frac{y}{r}+c_v y+\alpha V_{\oplus(st)}\big(y+\lambda_s-y\mu(r)\big)$. Since $V_{\oplus(st)}$ is a linear function, there exist constant reals ξ_{st} and η_{st} such that for all $x\geq 0$, $V_{\oplus(st)}(x)=\xi_{st}x+\eta_{st}$. Using this fact, we can re-write

$$V_{st}(y)=c_w\frac{y}{r}+c_v y+\alpha\big[\,\xi_{st}\big(y+\lambda_s-y\mu(r)\big)+\eta_{st}\big]$$

$$=\left(\frac{c_w}{r}+c_v+\alpha\xi_{st}\big(1-\mu(r)\big)\right)y+\alpha(\xi_{st}\lambda_s+\eta_{st}) \qquad \text{(B.3.1)}$$

Because $r_{st}^*\in(0,\bar{r}]$ minimizes the above expression, it must either be equal to $\bar{r}$ or be the solution to the first order condition $0=c_w\left(-\frac{y}{r^2}\right)-\alpha\,\xi_{st}y\left(\frac{d\mu(r)}{dr}\right)$, or in other words, $0=c_w+\alpha\,\xi_{st}\frac{d\mu(r)}{dr}r^2$ (for the case where $y\neq 0$).

Note that neither $\bar{r}$ nor the solution(s) to $0=c_w+\alpha\,\xi\frac{d\mu(r)}{dr}r^2$ depend on y. If there are multiple solutions to $0=c_w+\alpha\,\xi\frac{d\mu(r)}{dr}r^2$, we see from Equation (B.3.1) that $r_{s,t}^*$ is equal to the solution that yields the lowest value of $\frac{c_w}{r}+c_v+\alpha\xi_{st}\big(1-\mu(r)\big)$, which is independent of y. When $y=0$, we have $V_{st}(0)=\alpha V_{\oplus(st)}(\lambda_s)$, which is a constant, and therefore we can set the ratio r arbitrarily to r_{st}^* and the result holds. ■

LEMMA B.3.4. Suppose that $\psi=0$, $\phi=1$, and $b=\infty$. If $V_{\oplus(st)}(\cdot)$ is a linear function, then so is $V_{st}(\cdot)$.

Proof of Lemma B.3.4. By definition, we have

$$V_{st}(y) = \min_{r_s \geq 0} \left\{ c_w \frac{y}{r_s} + c_v y + \alpha E\left[V_{\oplus(st)}\left(y + A_s - D(y, r_s) \right) \right] \right\}.$$

Because the expectation operator is linear and we assumed $V_{\oplus(st)}(\cdot)$ is a linear function, we have

$$V_{st}(y) = \min_{r_s \geq 0} \left\{ c_w \frac{y}{r_s} + c_v y + \alpha V_{\oplus(st)}\left(y + \lambda_s - y\mu(r_s) \right) \right\}.$$

Applying the min operator is equivalent to replacing r by $r_{st}^*(y)$, which is the function returning the optimal value of r_s for any given number of requests y. By Lemma B.3.3, there exists r_{st}^* such that $r_{st}^*(y) = r_{st}^*$. Making this substitution, we have

$$V_{st}(y) = c_w \frac{y}{r_{st}^*} + c_v y + \alpha V_{\oplus(st)}\left(y + \lambda_s - y\mu(r_{st}^*) \right).$$

Observe that the function $y + \lambda_s - y\mu(r_{st}^*)$ is linear in y because r_{st}^* does not depend on y. Since $V_{\oplus(st)}(\cdot)$ was assumed to be linear, we conclude that $V_{st}(y)$ is a linear function in y. ■

Proof of Proposition 3.3. We proceed by considering the finite horizon version of the problem with t cycles remaining presented in Appendix B.2.

First, Lemma B.3.3 states that for any s and t, if $V_{\oplus(st)}(\cdot)$ is a linear function, then there exists an optimal workload r_{st}^* for shift s in period t that is independent of the number of requests in the system. With that result in hand, it is sufficient to show that $V_{st}(\cdot)$ is a linear function for all $s = \{1, \ldots, k\}$ and $t \geq 0$. The base case is that $V_{k0}(\cdot)$ is a linear function in the number of requests. This result follows because $V_{k0}(y) = \min_{r_k \geq 0} \left\{ c_w \frac{y}{r_k} + c_v y \right\}$ and $r_k^* = \bar{r}$ for all y, thereby making $V_{k0}(\cdot)$ a linear function in y.

For the inductive step, we suppose that it has been shown that $V_{\oplus}(\cdot)$ is a linear function in the number of requests and Lemma B.3.4 demonstrates the linearity of $V_{st}(\cdot)$ by induction on $t \geq 0$ if $V_{\oplus(st)}(\cdot)$ is a linear function.

We conclude the proof of Proposition 3.3 by using Lemma B.3.1, which states $V_{st}(y) \to V_s(y)$ uniformly as $t \to \infty$. Therefore, since $V_s(y)$ is a linear function in y, by Lemma B.3.3, there exists r_s^* such that $r_s{}^*(y) = r_s^*$ for all $y > 0$. ■

Proof of Proposition 3.4. Since $V_{\oplus(st)}(y)$ is a linear function and the expectation of a linear function is equivalent to a linear function of expectations, $v_{st}(y, r) = \frac{c_w y}{r} + c_v y + \alpha\xi\big(y + \lambda_s - y\mu(r)\big) + \eta$, for some constants ξ and η. The first order condition would then be

$$\frac{d}{dr} v_{st}(y, r) = -\frac{c_w y}{r^2} - \alpha\xi y \frac{d}{dr}\mu(r) = 0.$$

Since we know from Lemma 3.1 that $\mu(r)$ is quasiconcave with global maximum r_0, $\frac{d}{dr}\mu(r)$ is non-negative and thus $\frac{d}{dr} v_{st}(y, r)$ is negative for $r \leq r_0$. For $r_0 \leq r \leq r_1$, $\frac{d}{dr} v_{st}(y, r)$ increases with r. If there exists $r_0 \leq r^* \leq r_1$ that is a solution to the first order condition above, it would a unique local minimum since there can only be one more local optimum, which would be a local maximum, at the most for $r \geq r_1$. If there does not exist a local minimum for $r_0 \leq r \leq r_1$, the local optimum for $r \geq r_1$ would be the unique local minimum. If there does not exist the solution for the first order condition, then we know that $v_{st}(y, r)$ is always decreasing and would be minimized at $\bar{r}$. ■

Appendix C Proofs for Chapter 4

Proof of Proposition 4.1. We first show the profit function under traditional policy is concave, and use first-order condition to solve for the optimal Q. We only need to consider the case in which Q is no less than n because $Q < n$ only represents idle time, and thus there is no benefit, for the doctor.

When $\gamma = 0$,

$$\Pi_T = rQ(1-p) - k_{ot}[(Q-n)^+]^2 - k_{gw}[(Q-n)^+]^2$$

$$\frac{\partial \Pi_T}{\partial Q} = r(1-p) - 2\left(k_{ot} + k_{gw}\right)(Q-n)$$

$$\frac{\partial^2 \Pi_T}{\partial Q^2} = -2\left(k_{ot} + k_{gw}\right) < 0$$

$$Q^* = n + \frac{r(1-p)}{2\left(k_{ot} + k_{gw}\right)}$$

$$\Pi_T^* = \frac{r(1-p)\left[4n\left(k_{ot} + k_{gw}\right) + r(1-p)\right]}{4\left(k_{ot} + k_{gw}\right)}$$

Thus, Proposition 4.1 follows. ■

Proof of Proposition 4.2. It is never optimal if $\mu_{OA} + l < n$ because increasing $\mu_{OA} + l$ up to n would only increase the revenue while reducing idle time incurred. Thus, we only need to consider two cases: $\mu_{OA} - l \le n \le \mu_{OA} + l$ and $n \le \mu_{OA} - l \le \mu_{OA} + l$

$$D_{OA} \sim Unif[\mu_{OA} - l, \mu_{OA} + l]$$

$$f(x) = \frac{1}{2l}$$

Case 1: $\mu_{OA} - l \leq n \leq \mu_{OA} + l$

$$\max_{\mu_{OA}} \Pi_{OA} = r\mu_{OA} - k_{ot} \int_{n}^{\infty} (y-n)^2 \phi(y) dy$$

$$\Pi_{OA} = r\mu_{OA} - k_{ot} \int_{n}^{\mu_{OA}+l} (y-n)^2 \frac{1}{2l} dy = r\mu_{OA} - \frac{k_{ot}}{6l} (\mu_{OA} + l - n)^3$$

$$\frac{\partial \Pi_{OA}}{\partial \mu_{OA}} = r - \frac{k_{ot}}{2l} (\mu_{OA} + l - n)^2$$

$$\frac{\partial^2 \Pi_{OA}}{\partial \mu_{OA}^2} = -\frac{k_{ot}}{l} (\mu_{OA} + l - n) \leq 0$$

Since Π_{OA} is concave, the first-order condition is necessary and sufficient to find the optimal μ_{OA}. The first derivative is zero at two different points, $n - l \pm \sqrt{\frac{2lr}{k_{ot}}}$. However, $n - l - \sqrt{\frac{2lr}{k_{ot}}}$ is not a feasible value for μ_{OA} for this case because $n \leq \mu_{OA} + l$. Therefore, the optimal average number of patients served $\mu_{OA}^* = n - l + \sqrt{\frac{2lr}{k_{ot}}}$.

$$\Pi_{OA}^* = r(n-l) + \frac{2r\sqrt{2lr}}{3\sqrt{k_{ot}}}$$

Case 2: $n < \mu_{OA} - l \leq \mu_{OA} + l$

$$\max_{\mu_{OA}} \Pi_{OA} = r\mu_{OA} - k_{ot} \int_{n}^{\infty} (y-n)^2 \phi(y) dy$$

$$\Pi_{OA} = r\mu_{OA} - k_{ot}\int_{\mu_{OA}-l}^{\mu_{OA}+l}(y-n)^2\frac{1}{2l}dy = r\mu_{OA} - \frac{k_{ot}}{3}[l^2 + 3(n-\mu_{OA})^2]$$

$$\frac{\partial \Pi_{OA}}{\partial \mu_{OA}} = r + 2k_{ot}(n-\mu_{OA})$$

$$\frac{\partial^2 \Pi_{OA}}{\partial \mu_{OA}^2} = -2k_{ot} \leq 0$$

$$\mu_{OA}^* = n + \frac{r}{2k_{ot}}$$

$$\Pi_{OA}^* = rn + \frac{r^2}{4k_{ot}} - \frac{k_{ot}l^2}{3}$$

Note that when the optimal average number of patients served is substituted to both cases, case 1 represents when $l \geq \frac{r}{2k_{ot}}$ and case 2 represents when $l < \frac{r}{2k_{ot}}$, thereby accounting for the entire range of l values. ■

Proof of Proposition 4.3. For consistent comparison of the two policies, we compare the average number of patients doctor sees in a day, $Q^*(1-p)$ and μ_{OA}^*.

Case 1: $\mu_{OA} - l \leq n \leq \mu_{OA} + l$

$$\Delta\mu = Q^*(1-p) - \mu_{OA}^* = l - np + \frac{r(1-p)^2}{2(k_{ot}+k_{gw})} - \sqrt{\frac{2lr}{k_{ot}}}$$

$$\frac{\partial \Delta\mu}{\partial l} = 1 - \sqrt{\frac{r}{2k_{ot}l}}$$

$$\frac{\partial^2 \Delta\mu}{\partial l^2} = \frac{1}{2l}\sqrt{\frac{r}{2k_{ot}l}} \geq 0$$

$\Delta\mu$ is zero, and thus the average number of patients served under the two policies are equal, when $l = np + \frac{r}{k_{ot}} - \frac{r(1-p)^2}{2(k_{ot}+k_{gw})} + \sqrt{\frac{r[2k_{ot}(k_{ot}+k_{gw})np + r(k_{gw}+2k_{ot}p-k_{ot}p^2)]}{k_{ot}^2(k_{ot}+k_{gw})}}$.

When $l = \frac{r}{2k_{ot}}$,

$$\Delta\mu = -2np - \frac{r[k_{gw} + k_{ot}(2-p)p]}{k_{ot}(k_{ot}+k_{gw})} \leq 0$$

Therefore, $Q^*(1-p) > \mu_{OA}^*$ for $l > np + \frac{r}{k_{ot}} - \frac{r(1-p)^2}{2(k_{ot}+k_{gw})} + \sqrt{\frac{r[2k_{ot}(k_{ot}+k_{gw})np + r(k_{gw}+2k_{ot}p-k_{ot}p^2)]}{k_{ot}^2(k_{ot}+k_{gw})}}$.

Case 2: $n < \mu_{OA} - l \leq \mu_{OA} + l$

$$\mu_{OA}^* = n + \frac{r}{2k_{ot}}$$

$$\Delta\mu = Q^*(1-p) - \mu_{OA}^* = -np - \frac{r[k_{gw} + k_{ot}p(2-p)]}{2k_{ot}(k_{ot}+k_{gw})} \leq 0$$

Therefore, in case 2, the optimal average number of patients served under open-access policy is always greater than under traditional policy. ■

Proof of Proposition 4.4. For consistent comparison of profits for the two policies, we compare the maximum profits under the two policies.

Case 1: $\mu_{OA} - l \le n \le \mu_{OA} + l$

$$\Pi_T^* = \frac{r(1-p)\left[4n\left(k_{ot}+k_{gw}\right)+r(1-p)\right]}{4\left(k_{ot}+k_{gw}\right)}$$

$$\Pi_{OA}^* = r(n-l) + \frac{2r\sqrt{2lr}}{3\sqrt{k_{ot}}}$$

$$\Delta\Pi = \Pi_T^* - \Pi_{OA}^* = rl - rnp + \frac{r^2(1-p)^2}{4\left(k_{ot}+k_{gw}\right)} - \frac{2r\sqrt{2lr}}{3\sqrt{k_{ot}}}$$

$$\frac{\partial\Delta\Pi}{\partial l} = r - \frac{r^2}{3}\sqrt{\frac{2}{k_{ot}lr}}$$

$$\frac{\partial^2\Delta\Pi}{\partial l^2} = \frac{r^2}{3l\sqrt{2k_{ot}lr}} \ge 0$$

$$\Delta\Pi = 0 \Leftrightarrow l^* = np + \frac{4r}{9k_{ot}} - \frac{r(1-p)^2}{4\left(k_{ot}+k_{gw}\right)}$$

$$+ \frac{1}{9k_{ot}}\sqrt{\frac{72k_{ot}rnp\left(k_{ot}+k_{gw}\right)+16r^2k_{gw}-2r^2k_{ot}[1-9p(2-p)]}{k_{ot}+k_{gw}}}$$

Case 2: $n < \mu_{OA} - l \le \mu_{OA} + l$

$$\Pi_{OA}^* = rn + \frac{r^2}{4k_{ot}} - \frac{k_{ot}l^2}{3}$$

$$\Delta\Pi = \Pi_T^* - \Pi_{OA}^* = \frac{k_{ot}l^2}{3} - rnp - \frac{r^2\left[k_{gw} + k_{ot}p(2-p)\right]}{4k_{ot}\left(k_{ot} + k_{gw}\right)}$$

$$\frac{\partial \Delta\Pi}{\partial l} = \frac{2k_{ot}l}{3}$$

$$\frac{\partial^2 \Delta\Pi}{\partial l^2} = \frac{2k_{ot}}{3} \geq 0$$

$$\Delta\Pi = 0 \Leftrightarrow l^* = \frac{1}{2k_{ot}}\sqrt{\frac{\left[12k_{ot}rnp\left(k_{ot} + k_{gw}\right) + 3r^2\left[k_{gw} + k_{ot}p(2-p)\right]\right.}{k_{ot} + k_{gw}}}$$

Since $\frac{\partial \Delta\Pi}{\partial l}\left(l = \frac{r}{2k_{ot}}\right) = \frac{r}{3} \geq 0$, if $\Delta\Pi$ at $l = \frac{r}{2k_{ot}}$ is negative, the threshold value for l where profits for both policies would be equal would fall under case 1. If it is positive, the threshold would be under case 2.

$$\Delta\Pi\left(l = \frac{r}{2k_{ot}}\right) = \frac{r^2\left[-2k_{gw} + k_{ot} - 3p(2-p)\right]}{12k_{ot}\left(k_{ot} + k_{gw}\right)} - rnp$$

$$= \frac{r\left[-12k_{ot}np\left(k_{ot} + k_{gw}\right) - 2k_{gw}r + k_{ot}r - 3k_{ot}rp(2-p)\right]}{12k_{ot}\left(k_{ot} + k_{gw}\right)}$$

The sign of the above expression is determined by $-12k_{ot}np\left(k_{ot} + k_{gw}\right) - 2k_{gw}r + k_{ot}r - 3k_{ot}rp(2-p)$. Thus, Proposition 4.4 follows. ■

References

Aiken, L. H., J. P. Cimiotti, D. M. Sloane, H. L. Smith, L. Flynn, D. F. Neff. 2011. The effects of nurse staffing and nurse education on patient deaths in hospitals with different nurse work environments. *Medical care*. **49**(12) 1047.

Aiken, L. H., S. P. Clarke, D. M. Sloane, E. T. Lake, T. Cheney. 2008. Effects of hospital care environment on patient mortality and nurse outcomes. *The Journal of nursing administration*. **38**(5) 223.

Aiken, L. H., S. P. Clarke, D. M. Sloane, J. Sochalski, J. H. Silber. 2002. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *JAMA*. **288**(16) 1987-1993.

Aiken, L. H., W. Sermeus, K. Van den Heede, D. M. Sloane, R. Busse, M. McKee, L. Bruyneel, A. M. Rafferty, P. Griffiths, M. T. Moreno-Casbas. 2012. Patient safety, satisfaction, and quality of hospital care: cross sectional surveys of nurses and patients in 12 countries in Europe and the United States. *BMJ*. **344** e1717.

Aiken, L. H., D. M. Sloane, L. Bruyneel, K. Van den Heede, P. Griffiths, R. Busse, M. Diomidous, J. Kinnunen, M. Kózka, E. Lesaffre. 2014. Nurse staffing and education and hospital mortality in nine European countries: a retrospective observational study. *The Lancet*. **383**(9931) 1824-1830.

Aiken, L. H., D. M. Sloane, J. P. Cimiotti, S. P. Clarke, L. Flynn, J. A. Seago, J. Spetz, H. L. Smith. 2010. Implications of the California nurse staffing mandate for other states. *Health services research*. **45**(4) 904-921.

Aiken, L. H., Y. Xue, S. P. Clarke, D. M. Sloane. 2007. Supplemental nurse staffing in hospitals and quality of care. *The Journal of nursing administration*. **37**(7-8) 335.

Anderson, D. J., K. B. Kirkland, K. S. Kaye, P. A. Thacker, Z. A. Kanafani, G. Auten, D. J. Sexton. 2007. Underresourced hospital infection control and prevention programs: penny wise, pound foolish? *Infection Control*. **28**(07) 767-773.

Armony, M., S. Israelit, A. Mandelbaum, Y. N. Marmor, Y. Tseytlin, G. B. Yom-Tov. 2015. Patient flow in hospitals: A data-based queueing-science perspective. *Forthcoming in Stochastic Systems*.

Association of American Medical Colleges. 2010. *Physician shortages to worsen without increases in residency training*.

Bae, S. H., B. Mark, B. Fried. 2010. Use of temporary nurses and nurse and patient safety outcomes in acute care hospital units. *Health Care Management Review*. **35**(4) 333.

Bailey, N. T. 1952. A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. *Journal of the Royal Statistical Society. Series B (Methodological)* 185-199.

Bard, J. F., H. W. Purnomo. 2005. Hospital-wide reactive scheduling of nurses with preference considerations. *IIE Transactions*. **37**(7) 589-608.

Barnhart, C., P. Belobaba, A. R. Odoni. 2003. Applications of operations research in the air transport industry. *Transportation science*. **37**(4) 368-391.

Batt, R. J., C. Terwiesch. 2012. Doctors under load: An empirical study of state-dependent service times in emergency care. *Working Paper, The Wharton School*.

Bodenheimer, T., H. H. Pham. 2010. Primary care: current problems and proposed solutions. *Health Affairs*. **29**(5) 799-805.

Bond, C. A., C. L. Raehl, M. E. Pitterle, T. Franke. 1999. Health care professional staffing, hospital characteristics, and hospital mortality rates. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*. **19**(2) 130-138.

Bretthauer, K. M., H. S. Heese, H. Pun, E. Coe. 2011. Blocking in healthcare operations: A new heuristic and an application. *Production and Operations Management*. **20**(3) 375-391.

Burke, E. K., P. De Causmaecker, G. V. Berghe, H. Van Landeghem. 2004. The state of the art of nurse rostering. *Journal of scheduling*. **7**(6) 441-499.

California Department of Industrial Relations. 2001. *Regulating wages, hours and working conditions in the personal service industry* (Industrial welfare commission order #2-2001) https://www.dir.ca.gov/IWC/IWCArticle2.pdf.

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

Cayirli, T., E. Veral. 2003. Outpatient scheduling in health care: A review of literature. *Production and Operations Management*. **12**(4) 519.

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

Chakraborty, S., K. Muthuraman, M. Lawley. 2010. Sequential clinical scheduling with patient no-shows and general service time distributions. *IIE Transactions*. **42**(5) 354-366.

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

Chhatwal, J., O. Alagoz, E. S. Burnside. 2010. Optimal breast biopsy decision-making based on mammographic features and demographic factors. *Operations research*. **58**(6) 1577.

Cho, S. H., S. Ketefian, V. H. Barkauskas, D. G. Smith. 2003. The effects of nurse staffing on adverse events, morbidity, mortality, and medical costs. *Nursing research*. **52**(2) 71.

Cimiotti, J. P., L. H. Aiken, D. M. Sloane, E. S. Wu. 2012. Nurse staffing, burnout, and health care–associated infection. *American journal of infection control*. **40**(6) 486-490.

Clark, A. R., H. Walker. 2011. Nurse rescheduling with shift preferences and minimal disruption. *Journal of Applied Operational Research*. **3**(3) 148-162.

Cook, A., M. Gaynor, M. Stephens Jr, L. Taylor. 2012. The effect of a hospital nurse staffing mandate on patient health outcomes: Evidence from California's minimum staffing regulation. *Journal of Health Economics*. **31**(2) 340-348.

Daskin, M. S., L. K. Dean. 2005. Location of health care facilities. *Operations research and health care* 43-76.

De Véricourt, F., O. B. Jennings. 2011. Nurse staffing in medical units: A queueing perspective. *Operations Research*. **59**(6) 1320-1331.

Denton, B., D. Gupta. 2003. A sequential bounding approach for optimal appointment scheduling. *Iie Transactions*. **35**(11) 1003-1016.

Denton, B. T., A. J. Miller, H. J. Balasubramanian, T. R. Huschka. 2010. Optimal allocation of surgery blocks to operating rooms under uncertainty. *Operations research*. **58**(4-Part-1) 802-816.

Dobson, G., S. Hasija, E. J. Pinker. 2011. Reserving capacity for urgent patients in primary care. *Production and Operations Management*. **20**(3) 456-473.

Dobson, G., H. H. Lee, E. Pinker. 2010. A model of ICU bumping. *Operations research*. **58**(6) 1564.

Dobson, G., E. Pinker, R. L. Van Horn. 2009. Division of Labor in Medical Office Practices. *Manufacturing & Service Operations Management*. **11**(3) 525-537.

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

Gallucci, G., W. Swartz, F. Hackerman. 2005. Brief reports: Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psychiatric Services*. **56**(3) 344-346.

Gnanlet, A., W. G. Gilland. 2009. Sequential and simultaneous decision making for optimizing health care resource flexibilities. *Decision Sciences*. **40**(2) 295-326.

Green, L. 2005. Capacity planning and management in hospitals. *Operations Research and Health Care* 15-41.

Green, L. V., S. Savin. 2008. Reducing delays for medical appointments: A queueing approach. *Operations Research*. **56**(6) 1526-1538.

Green, L. V., S. Savin, N. Savva. 2013. "Nursevendor problem": Personnel staffing in the presence of endogenous absenteeism. *Management Science*. **59**(10) 2237-2256.

Gulliford, M., M. Morgan. 2003. *Access to health care*. Psychology Press.

Gupta, D., B. Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE transactions*. **40**(9) 800-819.

Gurses, A. P., P. Carayon, M. Wall. 2009. Impact of performance obstacles on intensive care nurses' workload, perceived quality and safety of care, and quality of working life. *Health Services Research*. **44**(2p1) 422-443.

Halm, E. A., C. Lee, M. R. Chassin. 2002. Is Volume Related to Outcome in Health Care? A Systematic Review and Methodologic Critique of the Literature. *Annals of Internal Medicine*. **137**(6) 511.

Hillier, F. S., G. J. Lieberman. 2001. Introduction to Operations Research, McGraw Hill. *New York*.

Ho, C.-J., H.-S. Lau. 1992. Minimizing total cost in scheduling outpatient appointments. *Management science*. **38**(12) 1750-1764.

Hopp, W. J., E. Tekin, M. P. Van Oyen. 2004. Benefits of skill chaining in serial production lines with cross-trained workers. *Management Science* 83-98.

Hugonnet, S., J. Chevrolet, D. Pittet. 2007. The effect of workload on infection risk in critically ill patients. *Crit Care Med*. **35**(1) 76-81.

Hur, D., V. A. Mabert, K. M. Bretthauer. 2004. Real-Time Work Schedule Adjustment Decisions: An Investigation and Evaluation. *Production and Operations Management*. **13**(4) 322-339.

Institute of Medicine. 2001. *Crossing the quality chasm: A new health system for the 21st century*. National Academies Press, Washington D.C.

Iwashyna, T. J., A. A. Kramer, J. M. Kahn. 2009. Intensive care unit occupancy and patient outcomes*. *Critical Care Medicine*. **37**(5) 1545.

Jaeker, J. B., A. L. Tucker. 2013. An empirical study of the spillover effects of workload on patient length of stay. *Working Paper, Harvard Business School*.

Jaeker, J. B., A. L. Tucker. 2015. Hurry up and slow down: Workload saturation in hospitals. *Working Paper*.

Jones, C. B. 2005. The costs of nurse turnover, part 2: application of the Nursing Turnover Cost Calculation Methodology. *Journal of Nursing Administration*. **35**(1) 41-49.

Jordan, W. C., R. R. Inman, D. E. Blumenfeld. 2004. Chained cross-training of workers for robust performance. *IIE Transaction*. **36**(10) 953-967.

Juraschek, S. P., X. Zhang, V. Ranganathan, V. W. Lin. 2012. United States registered nurse workforce report card and shortage forecast. *American Journal of Medical Quality*. **27**(3) 241-249.

Kahn, J. M., C. H. Goss, P. J. Heagerty, A. A. Kramer, C. R. O'Brien, G. D. Rubenfeld. 2006. Hospital volume and the outcomes of mechanical ventilation. *New England Journal of Medicine*. **355**(1) 41-50.

Kane, R. L., T. Shamliyan, C. Mueller, S. Duval, T. J. Wilt. 2007a. *Nurse staffing and quality of patient care*. Agency for Healthcare Research and Quality Rockville, MD.

Kane, R. L., T. A. Shamliyan, C. Mueller, S. Duval, T. J. Wilt. 2007b. The association of registered nurse staffing levels and patient outcomes: systematic review and meta-analysis. *Medical Care*. **45**(12) 1195-1204.

KC, D. S. 2013. Does multitasking improve performance? Evidence from the emergency department. *Manufacturing & Service Operations Management*. **16**(2) 168-183.

KC, D. S., C. Terwiesch. 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science*. **55**(9) 1486-1498.

KC, D. S., C. Terwiesch. 2012. An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management*. **14**(1) 50-65.

Kopach, R., P.-C. DeLaurentis, M. Lawley, K. Muthuraman, L. Ozsen, R. Rardin, H. Wan, P. Intrevado, X. Qu, D. Willis. 2007. Effects of clinical characteristics on successful open access scheduling. *Health Care Management Science*. **10**(2) 111-124.

Kosel, K. C., T. Olivo. 2002. The business case for workforce stability. *Voluntary Hospitals of America*.

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.

Lankshear, A. J., T. A. Sheldon, A. Maynard. 2005. Nurse staffing and healthcare outcomes: a systematic review of the international research evidence. *Advances in Nursing Science*. **28**(2) 163-174.

Lau, H.-S., A. H.-L. Lau. 2000. A fast procedure for computing the total system cost of an appointment schedule for medical and kindred facilities. *IIE Transactions*. **32**(9) 833-839.

Lee, D. K. K., S. A. Zenios. 2009. Optimal capacity overbooking for the regular treatment of chronic conditions. *Operations research*. **57**(4) 852-865.

Lin, H. 2014. Revisiting the relationship between nurse staffing and quality of care in nursing homes: An instrumental variables approach. *Journal of health economics*. **37** 13-24.

Lindley, D. V. 1952. *The theory of queues with a single server*. Cambridge Univ Press.

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

Mandelbaum, A., P. Momcilovic, Y. Tseytlin. 2012. On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers. *Management Science* Forthcoming.

Martin, A. B., D. Lassman, B. Washington, A. Catlin, N. H. E. A. Team. 2012. Growth in US health spending remained slow in 2010; health share of gross domestic product was unchanged from 2009. *Health Affairs*. **31**(1) 208-219.

May, J. H., W. E. Spangler, D. P. Strum, L. G. Vargas. 2011. The surgical scheduling problem: Current research and future opportunities. *Production and Operations Management*. **20**(3) 392-405.

McCue, M., B. A. Mark, D. W. Harless. 2003. Nurse staffing, quality, and financial performance. *Journal of health care finance*. **29**(4) 54-76.

McHugh, M. D., J. Berez, D. S. Small. 2013. Hospitals with higher nurse staffing had lower odds of readmissions penalties than hospitals with lower staffing. *Health Affairs*. **32**(10) 1740-1747.

Mehrotra, V., O. Ozlük, R. Saltzman. 2010. Intelligent Procedures for Intra-Day Updating of Call Center Agent Schedules. *Production and Operations Management*. **19**(3) 353-367.

Moz, M., M. V. Pato. 2003. An integer multicommodity flow model applied to the rerostering of nurse schedules. *Annals of Operations Research*. **119**(1-4) 285-301.

Moz, M., M. V. Pato. 2004. Solving the problem of rerostering nurse schedules with hard constraints: new multicommodity flow models. *Annals of Operations Research*. **128**(1-4) 179-197.

Moz, M., M. V. Pato. 2007. A genetic algorithm approach to a nurse rerostering problem. *Computers & Operations Research*. **34**(3) 667-691.

Murray, M., C. Tantau. 1999. Redefining open access to primary care. *Managed Care Quarterly*. **7** 45-55.

Murray, M. M., C. Tantau. 2000. Same-day appointments: exploding the access paradigm. *Family Practice Management*. **7**(8) 45-45.

Muthuraman, K., M. Lawley. 2008. A stochastic overbooking model for outpatient clinical scheduling with no-shows. *Iie Transactions*. **40**(9) 820-837.

Needleman, J., P. Buerhaus, S. Mattke, M. Stewart, K. Zelevinsky. 2002. Nurse-staffing levels and the quality of care in hospitals. *New England Journal of Medicine*. **346**(22) 1715-1722.

Needleman, J., P. Buerhaus, V. S. Pankratz, C. L. Leibson, S. R. Stevens, M. Harris. 2011. Nurse staffing and inpatient hospital mortality. *New England Journal of Medicine*. **364**(11) 1037-1045.

Needleman, J., P. I. Buerhaus, M. Stewart, K. Zelevinsky, S. Mattke. 2006. Nurse staffing in hospitals: Is there a business case for quality? *Health Affairs*. **25**(1) 204-211.

Newhouse, R. P., M. Johantgen, P. J. Pronovost, E. Johnson. 2005. Perioperative nurses and patient outcomes—mortality, complications, and length of stay. *AORN*. **81**(3) 508-528.

OECD/European Union. 2014. *Health at a glance: Europe 2014*. OECD Publishing. http://dx.doi.org/10.1787/health_glance_eur-2014-en.

Olivares, M., C. Terwiesch, L. Cassorla. 2008. Structural estimation of the newsvendor model: an application to reserving operating room time. *Management Science*. **54**(1) 41-55.

Patrick, J. 2012. A Markov decision model for determining optimal outpatient scheduling. *Health Care Management Science*. **15**(2) 91-102.

Peelen, L., N. F. de Keizer, N. Peek, G. J. Scheffer, P. Van der Voort, E. de Jonge. 2007. The influence of volume and intensive care unit organization on hospital mortality in patients admitted with severe sepsis: a retrospective multicentre cohort study. *Crit Care*. **11**(2) R40.

Phibbs, C., A. Bartel, B. Giovannetti, S. Schmitt, P. Stone. 2009. *The Impact of Nurse Staffing and Contract Nurses on Patient Outcomes: New Evidence from Longitudinal Data*. Working Paper, Columbia Business School.

Phibbs, C. S., L. C. Baker, A. B. Caughey, B. Danielsen, S. K. Schmitt, R. H. Phibbs. 2007. Level and volume of neonatal intensive care and mortality in very-low-birth-weight infants. *New England Journal of Medicine*. **356**(21) 2165-2175.

Pinker, E. J., R. A. Shumsky. 2000. The efficiency-quality trade-off of cross-trained workers. *Manufacturing & Service Operations Management*. **2**(1) 32-48.

Pitts, S. R., J. M. Pines, M. T. Handrigan, A. L. Kellermann. 2012. National trends in emergency department occupancy, 2001 to 2008: effect of inpatient admissions versus emergency department practice intensity. *Annals of emergency medicine*. **60**(6) 679-686. e673.

Pronovost, P. J., D. Dang, T. Dorman, P. A. Lipsett, E. Garrett, M. Jenckes, E. B. Bass. 2001. Intensive care unit nurse staffing and the risk for complications after abdominal aortic surgery. *Effective Clinical Practice*. **4**(5) 199-206.

Punnakitikashem, P., J. M. Rosenberger, D. B. Behan. 2008. Stochastic programming for nurse assignment. *Computational Optimization and Applications*. **40**(3) 321-349.

Rauner, M. S., W. J. Gutjahr, K. Heidenberger, J. Wagner, J. Pasia. 2010. Dynamic Policy Modeling for Chronic Diseases: Metaheuristic-Based Identification of Pareto-Optimal Screening Strategies. *Operations research*. **58**(5) 1269-1286.

Robinson, L. W., R. R. Chen. 2003. Scheduling doctors' appointments: optimal and empirically-based heuristic policies. *IIE Transactions*. **35**(3) 295-307.

Robinson, L. W., R. R. Chen. 2010. A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing & Service Operations Management*. **12**(2) 330-346.

Ross, S. M. 1983. *Introduction to stochastic dynamic programming*. Academic press.

Rothberg, M. B., I. Abraham, P. K. Lindenauer, D. N. Rose. 2005. Improving nurse-to-patient staffing ratios as a cost-effective safety intervention. *Medical care*. **43**(8) 785.

Samorani, M., L. R. LaGanga. 2015. Outpatient appointment scheduling given individual day-dependent no-show predictions. *European Journal of Operational Research*. **240**(1) 245-257.

Schultz, K. L., D. C. Juran, J. W. Boudreau. 1999. The effects of low inventory on the development of productivity norms. *Management Science*. **45**(12) 1664-1678.

Schultz, K. L., D. C. Juran, J. W. Boudreau, J. O. McClain, L. J. Thomas. 1998. Modeling and worker motivation in JIT production systems. *Management Science*. **44**(12-part-1) 1595-1607.

Scott, R. D. 2009. *The direct medical costs of healthcare-associated infections in U.S. hospitals and the benefits of prevention*. Division of Healthcare Quality Promotion, National Center for Preparedness, Detection, and Control of Infectious Diseases, Coordinating Center for Infectious Diseases, Centers for Disease Control and Prevention.

Shamliyan, T. A., R. L. Kane, C. Mueller, S. Duval, T. J. Wilt. 2009. Cost savings associated with increased RN staffing in acute care hospitals: Simulation exercise. *Nursing Economic*. **27**(5) 302-331.

Simchi-Levi, D., Y. Wei. 2012. Understanding the performance of the long chain and sparse designs in process flexibility. *Operations research*. **60**(5) 1125-1141.

Sims, C. E. 2003. Increasing clinical, satisfaction, and financial performance through nurse-driven process improvement. *Journal of Nursing Administration*. **33**(2) 68-75.

Spence Laschinger, H. K., M. P. Leiter. 2006. The impact of nursing work environments on patient safety outcomes: The mediating role of burnout engagement. *Journal of Nursing Administration*. **36**(5) 259.

Spetz, J. 2004. California's Minimum Nurse-to-Patient Ratios: The First Few Months. *Journal of Nursing Administration*. **34**(12) 571-578.

Stanton, M. W., M. K. Rutherford. 2004. *Hospital nurse staffing and quality of care*. Agency for Healthcare Research and Quality Rockville.

Tan, T. F., S. Netessine. 2014. When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Science*. **60**(6) 1574-1593.

Theokary, C., Z. J. Ren. 2011. An Empirical Study of the Relations Between Hospital Volume, Teaching Status, and Service Quality. *Production and Operations Management*. **20**(3) 303-318.

Thompson, S., M. Nunez, R. Garfinkel, M. D. Dean. 2009. OR Practice---Efficient Short-Term Allocation and Reallocation of Patients to Floors of a Hospital During Demand Surges. *Operations research*. **57**(2) 261-273.

Tourangeau, A. E., D. M. Doran, L. M. G. Hall, L. O'Brien Pallas, D. Pringle, J. V. Tu, L. A. Cranley. 2007. Impact of hospital nursing care on 30-day mortality for acute medical patients. *Journal of Advanced Nursing*. **57**(1) 32-44.

Truong, V.-A. 2015. Optimal Advance Scheduling. *Management Science*. **Articles in Advance** 1-14.

Tsai, P.-F. J., G.-Y. Teng. 2014. A stochastic appointment scheduling system on multiple resources with dynamic call-in sequence and patient no-shows for an outpatient clinic. *European Journal of Operational Research*. **239**(2) 427-436.

U.S. General Accounting Office (GAO). 2009. *Hospital emergency departments: Crowding continues to occur, and some patients wait longer than recommended time frames*. Report GAO-09-347.

Van Ryzin, G. J., K. T. Talluri. 2003. *Revenue management*. Springer, Boston, MA.

Wang, W. Y., D. Gupta. 2011. Adaptive appointment systems with patient preferences. *Manufacturing and Service Operations Management*. **13**(3) 373.

White, D. L., C. M. Froehle, K. J. Klassen. 2011. The effect of integrated scheduling and capacity policies on clinical efficiency. *Production and Operations Management*. **20**(3) 442-455.

Wright, P. D., K. M. Bretthauer. 2010. Strategies for addressing the nursing shortage: Coordinated decision making and workforce flexibility. *Decision Sciences*. **41**(2) 373-401.

Wright, P. D., K. M. Bretthauer, M. J. Côté. 2006. Reexamining the nurse scheduling problem: Staffing ratios and nursing shortages. *Decision Sciences*. **37**(1) 39-70.

Yankovic, N., L. V. Green. 2011. Identifying Good Nursing Levels: A Queuing Approach. *Operations research*. **59**(4) 942-955.

Curriculum Vitae

NAME: David Dong Won Cho

BORN: Seoul, Korea, 1983

DEGREES:

	B.S.	University of California, Berkeley, 2005
	M.B.A.	University of Southern California, 2009
	M.B.	Indiana University, 2012
	Ph.D.	Indiana University, 2015