NORTHWESTERN UNIVERSITY

Essays on Improving System's Throughput, Customer's Wait-time and Customer's Satisfaction in Service Operations

A Prospectus Exam Report

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Industrial Engineering and Management Sciences

By

Sina Ansari

EVANSTON, ILLINOIS

December 2018

ABSTRACT

The vast majority of interactions between customers and service providers are experiences that extend over time. Service systems that deliver excellent customer experience achieve greater customer satisfaction and therefore customer loyalty, and eventually raise revenue. The temporal aspects of service delivery have not yet been analyzed as carefully as its monetary aspects, while knowing how to control the timing of the service delivery and consequently enhancing the customers' experience can give a competitive advantage to the service providers in many industries from call centers to hospitals. By developing analytical data-driven models and conducting empirical studies, we address the gap between operations management models and the-state-of-the-art in marketing and psychology literature, using quantitative methods such as Markov Decision Process (MDP), queueing theory, and predictive analytics as well as qualitative methods such as surveys and interviews. This line of research has led to the development of practical policies to control and optimize system's throughput, customer's wait-time and customer's satisfaction in different service operations, which are presented in three chapters of this thesis.

Acknowledgements

I would like to express my sincere appreciation to my committee members: Dr. Seyed M.R. Iravani, Dr. Laurens Debo, Dr. Sanjeev Malik, M.D., and Dr. Ohad Perry for their continuous and valuable guidance and support without which this work would not have been completed. Particularly, I am thankful to my advisor, Dr. Iravani and co-advisor, Dr. Debo for sharing with me their knowledge, experience, and enthusiasm throughout my doctoral program.

This effort would not have been possible without very special individuals in the Northwestern Memorial Hospital, Eileen Brassil, Heather Kiernan, Nicholas Christensen, Dana Perry, Denise Butera and Daniel Cruz. I also would like to thank my dear friends and students, Caleb Han, Yasemin Dogruol, Jeremy Joseph, Yilmazcan Ozyurt, and Yun Choi who work with me on different projects at Northwestern University. I would also acknowledge all faculty, staff, friends and peers during the years of my graduate studies for enhancing my academic and personal experience at Industrial Engineering and Management Sciences Department at Northwestern University.

Most of all, I am grateful to my parents and brother for their love, support, and confidence in me through all these years.

Preface

Most existing models in multi-stage service systems assume full information on the state of downstream stages. In Chapter 1, motivated by mortgage application process, we investigate how much the lack of such information impacts the task's waiting times in a two-stage system with two types of tasks. The goal is to find the optimal control policy for a server to switch between type-1 and type-2 tasks while minimizing the average number of tasks in the system. First, we discuss how and when the server can make the decision of working on the type-1 tasks or type-2 without knowing full information about the number of tasks at a downstream stage. Second, we analyze how this lack of information affects this decision and under what conditions the server can capture the benefit of the full information. We develop heuristic policies for the server to make this decision without knowing full information about the number of tasks at a downstream stage and discuss their implementation in practice.

Design and control of service systems with impatient customers have been extensively studied in queueing literature. In an extensive literature review, we synthesize recent advancements and identify current research gaps. One example of such research gaps is addressed in Chapter 2, where we study a multi-class queuing system with a single server and customer abandonment. Customer abandonment as a performance measure is of a great importance in many service systems, especially in Infomercial call centers. However, minimizing the loss of revenue due to the abandonment of impatient customers is rarely

studied in the literature. We characterize the structure of the server's optimal scheduling policy that minimizes the total average customer abandonment cost. We find that the optimal service policy is a static priority policy, which is easy-to-implement in practice. We derive sufficient conditions under which the so-called $b\mu$–rule is optimal. Under the $b\mu$-rule, it is optimal to give priority to the customer type that has higher service rate (μ) and higher abandonment cost (b), i.e., higher index $b\mu$. When those conditions are not met, we introduce $b\mu\theta$–rule as a heuristic policy that performs well. Our numerical analysis shows that the optimal scheduling policy results in an average cost saving of 80% compared to the commonly used first-come-first-served policy.

Excessive wait-time is the most common reason patients become unsatisfied and leave the emergency department (ED) before being treated. In Chapter 3, we aim to determine the impact of announcing patient's wait-times on patients satisfaction considering the loss-averse behavior of patients. For that purpose, using predictive analytics and two years of hospital data, a institution specific application is developed to predict patient's wait-times. In a field experiment in an urban emergency department, we observe that providing patients with their estimated wait-times improves their self-report satisfaction significantly. We also find that even though overestimating the announced delay increases the average wait-time satisfaction, overestimating too much may have negative impacts on patients wait-time experience. We discuss how to engineer the delay announced to maximize the average wait-time satisfaction.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Optimal Control Policies in Service Systems with Limited Information on the Downstream Stage

1.1. Introduction

One of the main goals of any multi-stage congested system is to minimize jobs waiting time to improve the efficiency of process flow Fitzsimmons et al. (2006) and customer satisfaction Davis and Heineke (1998). One common situation that is observed in multi-stage service systems is that jobs might be returned for reprocessing. In this case, the server at upstream stage receives two types of jobs at her stage and faces the decision of whether to work on a new arriving job or on a returned job.

One example of a two-stage service system in which a server at upstream stage faces a decision of working on a new job or on a returned job is the *Mortgage Application Process*. As described by Freddie Mac mortgage guide[1], the application process starts with a *Loan Officer* completing the mortgage loan application, see Figure 1.1- *Left*. The completed application is sent to a *Loan Processor* for review and for preparing the application for presentation to the *Underwriter*.[2] If the application is complete, the Loan Processor sends it to the Underwriter. If the application needs more documents or revisions, the Loan Processor sends it back to the Loan Officer for reprocessing. After reprocessing, the Loan

[1] `http://www.freddiemac.com/singlefamily/docs/Step_by_Step_Mortgage_Guide_English.pdf`

[2] The professional authorized to assess if the application is eligible for the mortgage loan he or she is applying for.

Officer then sends the completed application directly to the Underwriter herself. In this process, the Loan Officer receives both new applications and applications that require reprocessing and therefore, faces the decision of whether to work on a new arriving application or on a returned application. This administrative procedure is also common in other service systems such as law offices or human resource offices. In these systems, both the report owner and the reviewer have the permission to submit the report to the downstream stage and the reviewer's comments, which are usually simple and straightforward (e.g, asking for adding/removing a document), need not to be sent back to the report owner for re-checking. For example, as it is represented in the administrative process flow charts of Federal Aviation Administration (FAA)[3], whenever a *Secretary* submits a request (e.g., Procurements Request) to her *Manager* for approval, the Manager can either approve the request or send it back to the Secretary for reprocessing and submission. While our problem was motivated by different administrative processes, the queueing dynamics with feedback is also prevalent in Manufacturing. For example, Stage 1 can be a production/assembly (station or department) and Stage 2 can be a quality control/inspection (station or department). If quality inspection reveals that jobs require rework, they are sent to the first stage for rework to fix the quality issues. When the quality is fixed, the job is sent directly to Finished Goods Inventory (FGI) or to another stage of production.

The decision of which type of application to work on next will directly affect the number of applications in the system, which affects other operational performance measures such as flow time and throughput. For example, let us assume that mortgage applications are coming every 5 hours and suppose it takes an average of 3 hours for a Loan

[3] `www.tc.faa.gov/ota/FlowChartofProcesses_ver6b.ppt`

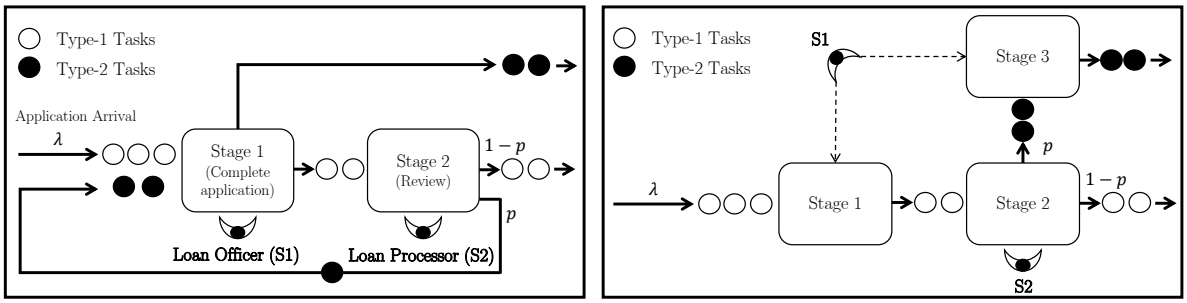


Figure 1.1. *Left*: Two-stage interconnect queue with feedback, *Right*: Corresponding three-stage tandem queue with flexible server.

Officer to process a new application and 2 hours to reprocess a returned application (i.e., assume Poisson arrivals and exponential service times). Further, suppose a Loan Processor's average review time is 4 hours and 90% of the applications are sent back to the Loan Officer. Two commonly used policies used by for Loan Officer are: (i) the new and returning applications are processed first-come-first-served (FCFS); (ii) priority is given to the returning applications. These two policies, however, perform poorly compared to the optimal policy (we characterize the optimal policy in Section 4). In this example, FCFS can result in 46% higher waiting time and priority policy can result in 7% higher waiting time than that under the optimal policy. These numbers can be even higher depending on the processing times and arrival rates.

Motivated by this problem as well as by the opportunity for significant improvement, this study investigates how a server should make a decision about which job to work on next when there are two types of jobs (i.e., new and returning jobs) in her stage. Making this decision optimally requires full information on the state of all stages in the system. However, it is often difficult (if not impossible) for a server to monitor all stages of the process. The system may also not allow sharing of such information for privacy reasons. In mortgage application process, for example, the Loan Officer does not necessarily know

how busy the Loan Processor (her boss) might be (i.e., how many application is waiting in her boss's queue).

To gain insight into the dynamics of such a problem, we consider a two-stage tandem system with returning jobs at the second stage, as shown in Figure 1.1- *Left*. All finished jobs at the Stage 1 are transferred to Stage 2. Jobs completed at Stage 2 leave the system with probability $1-p$ (type-1 jobs) or are sent back to Stage 1 for further processing with probability p (type-2 jobs). Therefore, type-1 jobs require work at Stage 1 and 2, while type-2 jobs require work at Stage 1, 2 and then 1 again.

There are two servers in the system. Server S1 works at Stage 1 and Server S2 works at Stage 2. Server S1 faces two queues, one for type-1 jobs and one for type-2 jobs. This system is equivalent to a three-stage system where Server S1, who is referred to as "flexible server" in the literature, works at Stages 1 and 3 and Server S2 works at Stage 2, see Figure 1.1- *Right*. Stage 1 corresponds to type-1 jobs and Stage 3 corresponds to type-2 jobs. Let N denote the number of jobs in the system. The goal is to find the optimal control policy for Server S1 to switch between Stage 1 and Stage 3 that minimizes the long-run average number of jobs in the system, which we refer to $E[N]$. Minimizing $E[N]$ is one of the main goals of service systems and it is important since for a given Throughput, minimizing $E[N]$ minimizes average waiting time in the system.

There are two main considerations for Server S1 when she makes the decision of which stage to work at. Finishing one type-2 job at Stage 3 can immediately reduce N by 1. From this perspective, Server S1 should prioritize Stage 3. On the other hand, not working on type-1 jobs at Stage 1 can temporarily reduce the Throughput of Stage 1, resulting in starvation and thus idling of Stage 2. This implies that the server should prioritize Stage

1. Therefore, there is a trade off between reducing number of type-2 jobs at Stage 3 (to reduce the number of jobs in the system) and reducing number of type-1 jobs at Stage 1 (to prevent the starvation of Stage 2).

In some interconnected systems, servers do not know the number of jobs in the downstream stages due to system's physical limitations or access level restrictions. In case of the model we study, while information about Stage 1 and 3 is always known to Server S1, information about the number of jobs at Stage 2 may not be fully known to Server S1. Therefore, we investigate three different information scenarios about the number of jobs in Stage 2: Full Information, No Information, and Partial Information.

Under *full information* scenario, where the number of jobs in each stage is fully known by Server S1 all the time, we prove that the optimal control policy for Server S1 has a monotone threshold structure with respect to the number of jobs waiting at Stage 1 and 3. When *no information* about the number of jobs in Stage 2 is available, we propose a heuristic policy (called No-Information Threshold (NIT) policy) that works well in minimizing the total number of jobs in the system. The performance of this heuristic is compared to the optimal static priority policy (where priority is given to a particular stage) and to the optimal dynamic policy under full information, where the number of jobs in each stage is fully known all the time.

Even though the exact number of jobs at Stage 2 may not be known, the server may know whether that number is small or large (i.e., it is below or above a threshold or whether the server at Stage 2 is busy or not). Taking this *partial information* into account, we finally propose an easy-to-implement heuristic policy (called Partial-Information

Threshold (PIT) policy) that performs almost as good as the optimal dynamic policy under the full information scenario.

In a comprehensive numerical analysis, we compare the performances of the proposed heuristics with that of the optimal dynamic policy. Our numerical analysis shows that PIT policy performs very well and NIT policy works as well as the optimal dynamic policy under certain conditions. When Server S1 and Server S2 have low utilizations and the percentage of type-2 jobs (i.e., p) is small, NIT performs well. This suggests that for systems that a small fraction of jobs are sent back to Stage 1 for rework, NIT policy can be a good candidate to replace the complicated optimal dynamic policy. We also examine the robustness of our numerical analysis by checking the impact of errors in setting thresholds for NIT and PIT policies and also the impact of variability of service times on the performance of the proposed heuristic policies.

To summarize, the contribution of our study is along two dimensions. First, we characterize the structure of the optimal policy for the flexible server, when she has full information about the number of jobs in all three stages. Second, we develop easy-to-implement polices for cases where no information or partial information about the downstream stage is available. We also provide insights into situations where these heuristic policies perform as well as the complex optimal dynamic policy under full information. To the best of our knowledge, the literature on flexible workers scheduling assume full information of all states. Our study is the first study that studies the lack of such information on server's policy and system performance.

The remainder of this chapter is organized as follows. We review the related literature in Section 2. In Section 3, we introduce the model specification. Section 4 discusses

the control policy under full information and presents the corresponding Markov decision process (MDP) formulation. In Section 5, control policies under no or partial information are introduced. Finally, the numerical analysis is presented in Section 6. We conclude the chapter in Section 7.

1.2. Literature Review

In this chapter, we study the optimal control policy of a flexible server in an interconnected queueing system. Flexible servers, who can be assigned dynamically to work in different stages of a process, are the foundation of workforce agility. Workforce agility provides systems with the ability to achieve high level of efficiency while meeting objectives for quality, operational efficiency and customer service. Hopp and OYEN (2004) outline frameworks for assessing and classifying the use of flexible servers in manufacturing and service operations. In a literature review of server assignment problems in manufacturing systems, Ammar et al. (2013) also acknowledge the impact of servers flexibility on the system performance.

One of the most commonly used system in manufacturing and service systems is a system in which stages of the system are in sequence (e.g., production lines). These systems are modeled as tandem queues in queueing literature. In recent years, the literature on tandem queues with flexible servers are growing. In this section, we review the literature on tandem queues with different combination of servers and stages.

The simplest non-trivial but prevalent queueing networks are two-stage tandem queues with no feedback and with one flexible server. Rosberg et al. (1982) and Hajek (1984) consider a two-stage tandem queueing system with one flexible server and one dedicated

server at Stage 2. Farrar (1993) studies a two-stage tandem queues with a dedicated server to each stage and a single flexible server for the whole system. He shows that the optimal allocation policy of an additional server is *transition-monotone* when holding cost in Stage 2 is larger than that in Stage 1. A control policy is transition-monotone if after a service completion at stage i, the optimal service rate at that stage does not increase, and the optimal service rate at other stages $j \neq i$ does not decrease. Generalizing the classic tandem (make-to-order) queue to include finished goods inventory (make-to-stock), Veatch and Wein (1994) analyze a manufacturing facility consisting of two stations in tandem that operates in a make-to-stock mode. Considering a two-stage tandem queue attended by a single flexible server, Iravani et al. (1997) show that the optimal dynamic policy in the second stage is greedy and, if the holding cost rate in the second stage is greater than the rate in the first stage, then the optimal dynamic policy in the second stage is also exhaustive. Our study is different from the existing literature on two-stage tandem queues with one flexible servers in the following dimensions. First, while in the current literature the server has full information about the number of jobs in each queue when making decision, in our study, we consider cases where the server has no or has partial information about the number of jobs in downstream queues. Second, in our model jobs can return to the first stage for reprocessing, on the contrary to the papers discussed above.

There are also studies on two-stage tandem queues with feedback, or with departure from any stage. Pandelis and Teneketzis (1994) consider a two-stage queueing system where jobs that complete service in Queue 1 join Queue 2 with probability p, and leave the system with probability $1-p$. They assume pre-loaded jobs in the system with no external

arrivals, while the system in our study has external arrivals. Pandelis (2007) studies a two-stage tandem queueing system with the same setting as Pandelis and Teneketzis (1994), but with a dedicated server at each stage and an additional flexible server. They find conditions under which idling is optimal for the flexible server. In another study, Pandelis (2008) considers a similar model with external arrivals where there are two dedicated servers and a flexible server. All servers have exponentially distributed service times and the service rate (capacity) may change randomly. He shows that the switching policy for the flexible server is monotone with respect to the number of jobs in the system. Including customer impatience, Zayas-Cabán et al. (2016) extend the server scheduling problems for a two-stage tandem queueing system. As a result of customer impatience, uniformization is not possible since the transition rates are unbounded. To address this difference, they formulate the server scheduling problem as a continuous-time Markov decision process (CTMDP) and provide sufficient conditions for when it is optimal to prioritize Stage 1 or Stage 2 service. On the contrary to our model, the jobs in these papers cannot return to the first stage for reprocessing and therefore there is no feedback. Considering feedback, Tang and Zhao (2008) analyze a tandem queue, where customers may either leave upon completion of service at the second stage or return to the first stage with some probability. As opposed to our model, server at the first stage do not need to make any decision on which job to process next and therefore they do not address the server scheduling problem. They use the tandem queue with feedback to demonstrate how to deal with a block generating function of GI/M/1 type, and to illustrate of how the boundary behavior can affect the tail decay rate.

The optimal control policy of a two-stage tandem queueing system with multiple flexible servers are also widely studied in the literature. Considering two parallel flexible servers, Ahn et al. (1999) study stochastic scheduling of a two-stage tandem queueing system and characterize sufficient and necessary conditions under which it is optimal to allocate both servers to the upstream or downstream. Javidi et al. (2001) address a two-stage system with Poisson arrival and exponential service times, considering multiple flexible servers in each queue. A single job that completes service in Stage 1 creates k jobs in Stage 2 with probability p and the job leaves the system with probability $1 - p$. The authors give two sets of conditions under which giving priority to Stage 1 is optimal or giving priority to Stage 2 is optimal. Weichbold and Schiefermayr (2006) study a scheduling problem with two interconnected queues and two flexible servers. Considering waiting costs, they find a sufficient condition under which it is optimal to allocate both servers to Stage 1 for any number of jobs in Stage 1 and Stage 2. More recently, Baumann and Sandmann (2017) study multi-server tandem queues where both stations have a finite buffer and service times are assumed to follow phase-type distribution. Two-stage systems with multiple flexible servers are also studied by Andradóttir and Ayhan (2005), Schiefermayr and Weichbold (2005), Wu et al. (2006) and Andradóttir et al. (2012). In contrast to our study, returning jobs is not considered in the setting of these studies and authors assume that the state information is fully known to all servers when they make the decisions of which queue to serve next.

There are also some papers that analyze queueing systems with flexible servers with more than two stages (e.g., Ahn and Righter (2006), Sennott et al. (2006), Andradóttir et al. (2007), Hopp et al. (2005), Kırkızlar et al. (2010) and Kırkızlar et al. (2014)). Even

though our underlying model is a two-stage system, the proposed equivalent model that we analyze in this study can be considered a three-stage system with a flexible server who works in Stage 1 and Stage 3. Dobson et al. (2013) model a three-stage two-server queueing system in which Server 2's work in Stage 2 depends on the decision of the Server 1 in Stage 1. Server 1 (investigator) collects information from the customer and decides what work needs to be done in the second stage. In the third stage, Server 1 provides customers with a conclusion, solution or diagnosis based on additional information or analysis done by Server 2 (back office). The authors then analyze the impact of server 1's decision of working on new customers versus discharging customers on system throughput. They show that when interruptions are not an issue, Server 1 should prioritize new customers to maximize throughput, keeping the system as full as possible. If customers who have been in the system for a long time generate interruptions and thus additional work for Server 1, it is asymptotically optimal for Server 1 to keep the system occupancy low and prioritize discharging customers. Even though there are some similarities, our model is different from Dobson et al. (2013) in several ways. In their model, jobs have to return to Server 1 after being processed at Stage 2, while in our model with some probability, jobs may leave the system without returning to Server 1. Therefore, a structurally different model is addressed in Dobson et al. (2013). Moreover, the unavailability of information on the state of downstream stage is not discussed in their analysis.

In a multi-stage multi-server interconnected queueing setting, Campello et al. (2016) propose a stochastic model of a baseline case-manager system. They define a case manager as a server who is assigned multiple customers and has frequent, repeated interactions with each customer until the customer's service is completed (e.g., emergency department

physicians). They formulate models that provide performance bounds and stability conditions for the baseline system. They also develop two approximations, one of which is based on a two-time-scale approach. The model analyzed in this study is structurally different from our model. Also, the authors do not address the issue of limited information on the downstream stage, as we do in our study.

In conclusion, the impact of limited information about the downstream stages (Stage 2 in our model) on server's policy and on system performance has not been studied in the literature of tandem queues with flexible servers. In this chapter, we characterize the structure of the optimal control policy under full information about Stage 2. Given the insights gained from the optimal control policy, we propose close-to-optimal heuristic polices that perform well when there is no information or limited information available about the number of jobs at Stage 2.

1.3. Model

Consider a system with two stages. Jobs arrive at Stage 1 according to a Poisson process with rate λ. All completed jobs at Stage 1 are transferred to Stage 2. Jobs completed at Stage 2 either leave the system with probability $1-p$ (type-1) or transferred back to Stage 1 with probability p (type-2). Type-2 jobs leave the system after their work is completed at Stage 1. We assume that all stages have unlimited buffers (e.g., jobs are files or folders that do not need a large space to store).

From the modeling perspective, as we mentioned, this problem is equivalent to a three-stage system with one dedicated server working at Stage 2, and one flexible server attending Stage 1 (type-1 jobs) and Stage 3 (type-2 jobs), see Figure 1.1-*Right*. The

three-stage model is a generalization of the two-stage tandem queue where the service rate at the second stage (i.e. μ_2) is infinity. When service rate at the second stage (i.e. μ_2) is infinity, our model works as a tandem queue with a flexible server, where jobs leave the system with probability $1 - p$ and return to the first stage with probability p. We assume that the service time at Stage i is exponentially distributed with rate μ_i. We note here that the exponential assumption regarding the service times allows us to formulate the model as a MDP and characterize the structure the optimal dynamic policy. After our MDP reveals the structure of the optimal dynamic policy, it becomes clear that our main insights about the structure are not influenced by this assumption on service times. In Section 6.4, we present the additional numerical analysis to check the robustness of our findings for the case that the service times are not exponentially distributed. We further assume that the switchover time (i.e., switching between Stages 1 and 3) is negligible compared to process times. We also assume that preemption is allowed so Server S1, if needed, can interrupt her job (processing a job at a stage) and start another job (i.e. processing a job at another stage, or remain idle).

The goal is to find the best control policy for Server S1 that minimizes the long-run average number of jobs in the system. We pursue this goal under three different information scenarios for Server S1: Full Information, No Information and Partial Information about the number of jobs at Stage 2. We discuss the control policy under each of these scenarios in the following sections.

1.4. Control Policy Under Full Information

In this scenario, we assume that Server S1 knows the number of jobs at each stage of the three-stage tandem queue, when she makes the decision of working on a type-1 or a type-2 job. Server S1 has three decisions to make: stay idle, work at Stage 1, or work at Stage 3.

1.4.1. The MDP Formulation

We formulate the problem as a Markov Decision Process (MDP) as follows:

- *Decisions epochs* are job completion or job arrival epochs at any stage.
- *State Space* $\mathscr{U}$ consists of 3 dimensional vectors $\mathbf{n}$, where $\mathbf{n} = (n_1, n_2, n_3)$, and n_i is the number of jobs at Stage i, including the one in service, $n_i \geq 0$, $\forall i$.
- *Actions* include $a \in \{I, P1, P3\}$: Idling (I), process (or continue processing) a job at Stage 1 (P1) and process (or continue processing) a job at Stage 3 (P3).

The optimality equation of the MDP with the objective of minimizing $E[N]$ in the system can be expressed as

$$\begin{aligned} \frac{g}{\Lambda} + V(\mathbf{n}) = & \frac{1}{\Lambda}\Big\{ n_1 + n_2 + n_3 + \lambda V(\mathbf{n} + \mathbf{e}^1) + \mu_2 B(\mathbf{n}) \\ & + \min\Big\{ A_I(\mathbf{n}), A_{P1}(\mathbf{n}), A_{P3}(\mathbf{n}) \Big\} \Big\}, \end{aligned} \tag{1.1}$$

where g is the long-run average number of jobs in the system and $\Lambda = \lambda + \mu_1 + \mu_2 + \mu_3$ is the uniformization rate. We define $\mathbf{e}^i$ as a 3-dimensional vector with zero elements except for its i^{th} element, which is one. Also, we define

$$B(\mathbf{n}) = \begin{cases} V(\mathbf{n}) & : \text{ if } n_2 = 0 \\ pV(\mathbf{n} - \mathbf{e}^2 + \mathbf{e}^3) + (1-p)V(\mathbf{n} - \mathbf{e}^2) & : \text{ if } n_2 > 0 \end{cases} \tag{1.2}$$

For $A_I(\mathbf{n}), A_{P1}(\mathbf{n})$ and $A_{P3}(\mathbf{n})$, we have:

$$A_I(\mathbf{n}) = (\mu_1 + \mu_3)V(\mathbf{n})$$

$$A_{P1}(\mathbf{n}) = \begin{cases} \mu_1 V(\mathbf{n}) + \mu_3 V(\mathbf{n}) & : \text{ if } n_1 = 0 \\ \mu_1 V(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2) + \mu_3 V(\mathbf{n}) & : \text{ if } n_1 > 0 \end{cases} \tag{1.3}$$

$$A_{P3}(\mathbf{n}) = \begin{cases} \mu_3 V(\mathbf{n}) + \mu_1 V(\mathbf{n}) & : \text{ if } n_3 = 0 \\ \mu_3 V(\mathbf{n} - \mathbf{e}^3) + \mu_1 V(\mathbf{n}) & : \text{ if } n_3 > 0 \end{cases}$$

Operator $A_a(\mathbf{n})$ represents possible transitions at State $(\mathbf{n})$ if action a is chosen.

1.4.2. Structure of The Optimal Control Policy

In this section, we characterize the optimal control policy for Server S1. We first need to discuss the stability condition for the system in Proposition 1. The proof of the proposition and other analytical results are presented in the appendix.

Proposition 1.1. *The system is stable if:*

$$\mu_i > \lambda, \quad i = 1, 2$$

$$\mu_3 > p\lambda \tag{1.4}$$

$$\frac{\mu_1 \mu_3}{p\mu_1 + \mu_3} > \lambda$$

Theorem 1.1. *If conditions in Proposition 1.1 hold, then there exists an average-cost optimal stationary policy for the MDP which has a constant average cost. Moreover, the value iteration algorithm converges.*

Let Υ be the set of functions defined on $\mathscr{U}$ such that if function $v \in \Upsilon$, then v satisfies

C1: $v(\mathbf{n})$ is nondecreasing in $n_1 \geq 0$, $n_2 \geq 0$ and $n_3 \geq 0$.

C2: $v(\mathbf{n}) \geq v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2)$, for $n_1 > 0, n_2 \geq 0, n_3 \geq 0$.

M1: For $n_1 > 1, n_2 \geq 0, n_3 > 0$,

$$\mu_3[v(\mathbf{n} - \mathbf{e}^1 - \mathbf{e}^3) - v(\mathbf{n} - \mathbf{e}^1)] + \mu_1[v(\mathbf{n} - \mathbf{e}^1) - v(\mathbf{n} - 2\mathbf{e}^1 + \mathbf{e}^2)]$$

$$\leq \mu_3[v(\mathbf{n} - \mathbf{e}^3) - v(\mathbf{n})] + \mu_1[v(\mathbf{n}) - v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2)].$$

M2: For $n_1 > 0, n_2 \geq 0, n_3 > 0$,

$$\mu_3[v(\mathbf{n} + \mathbf{e}^2 - \mathbf{e}^3) - v(\mathbf{n} + \mathbf{e}^2)] + \mu_1[v(\mathbf{n} + \mathbf{e}^2) - v(n - \mathbf{e}^1 + 2\mathbf{e}^2)]$$

$$\leq \mu_3[v(\mathbf{n} - \mathbf{e}^3) - v(\mathbf{n})] + \mu_1[v(\mathbf{n}) - v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2)].$$

M3: For $n_1 > 0, n_2 \geq 0, n_3 > 0$,

$$\mu_3[v(\mathbf{n}) - v(\mathbf{n} + \mathbf{e}^3)] + \mu_1[v(\mathbf{n} + \mathbf{e}^3) - v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2 + \mathbf{e}^3)]$$

$$\leq \mu_3[v(\mathbf{n} - \mathbf{e}^3) - v(\mathbf{n})] + \mu_1[v(\mathbf{n}) - v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2)].$$

Define operator $\mathbf{T}$ on the set of real-valued functions defined on $\mathscr{U}$ by

$$\mathbf{T}v(\mathbf{n}) = \frac{1}{\Lambda}\Big\{n_1 + n_2 + n_3 + \lambda v(\mathbf{n} + \mathbf{e}^1) + \mu_2 B(\mathbf{n}) + \min\Big\{A_I(\mathbf{n}), A_{P1}(\mathbf{n}), A_{P3}(\mathbf{n})\Big\}\Big\}.$$

The following Proposition shows that properties **C1**, **C2**, and **M1-M3** are preserved under operator $\mathbf{T}$. Moreover, it shows $V \in \Upsilon$ and therefore value function $V(\mathbf{n})$ satisfies properties **C1**, **C2**, and **M1-M3**.

Proposition 1.2. *If $v \in \Upsilon$, then (a) $\mathbf{T}v \in \Upsilon$; and (b) $V \in \Upsilon$.*

From properties **C1** and **C2** in Proposition 1.2, we have the following results.

Theorem 1.2. *Idling is not optimal for Server S1 as long as there are jobs available at Stages 1 or 3.*

Theorem 2 implies that, when the goal is to minimize long-run average number of jobs in the system, working on a job and pushing it to a downstream stage or finishing the job so it leaves the system is preferred to idling.

Theorem 1.3. *If at state* $\boldsymbol{n}$*, for* $n_1 > 0, n_2 \geq 0, n_3 > 0$*, it is optimal for Server S1 to work at Stage 3, then it is also optimal for her to work at Stage 3:*

- *at state* $(\mathbf{n} - \mathbf{e}^1)$*,*
- *at state* $(\mathbf{n} + \mathbf{e}^2)$*,*
- *at state* $(\mathbf{n} + \mathbf{e}^3)$*.*

Similarly, If at state $\mathbf{n}$*, for* $n_1 > 0, n_2 \geq 0, n_3 > 0$*, it is optimal for Server S1 to work at Stage 1, then it is also optimal for her to work at Stage 1:*

- *at state* $(\mathbf{n} + \mathbf{e}^1)$*,*
- *at state* $(\mathbf{n} - \mathbf{e}^2)$*,*
- *at state* $(\mathbf{n} - \mathbf{e}^3)$*.*

Theorem 1.3 implies that it becomes optimal for Server S1 to work at Stage 3 as n_3 increases or n_1 decreases, and work at Stage 1 as n_1 increases or n_3 decreases. When n_2 increases, preventing Stage 2 from starvation is less important and it is expected that there will be more arrivals at Stage 3. Therefore, the optimal dynamic policy recommends working at Stage 3, when n_2 increases.

Figure 1.2 illustrates a typical structure of the optimal dynamic policy. As the figure shows, the optimal control policy for Server S1 under full information has a monotone threshold structure. If Server S1 knows the number of jobs at Stage 2 all the time, she is able to minimize long-run average number of jobs in the system by following the optimal

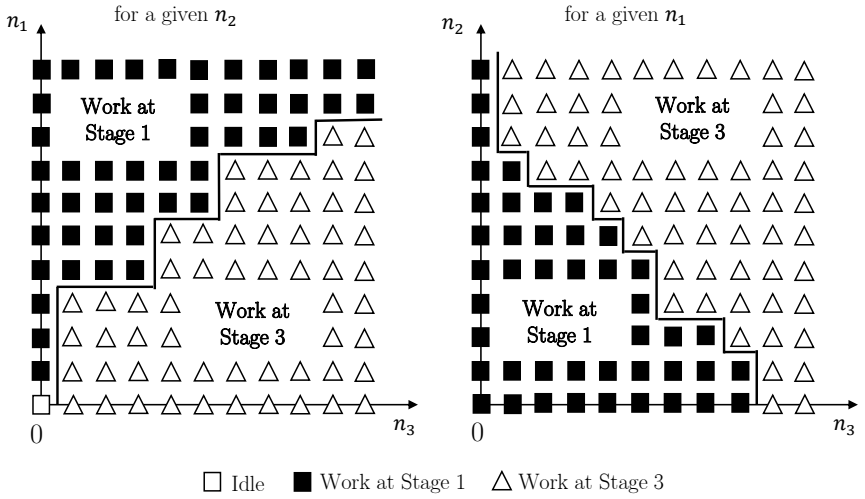


Figure 1.2. *Left*: A Typical structure of the optimal dynamic policy for given n_2, *Right*: A Typical structure of the optimal dynamic policy for given n_1. Depending on the values of n_1 and n_2 the monotone structure changes.

dynamic policy. However, she may not (or cannot) monitor the state of Stage 2 or the system may not allow sharing of such information for privacy reasons. Due to the fact that the lack of information on n_2 may be inevitable, we must design control policies that take this into account. In following sections, we introduce policies that can be used when the information about n_2 is not available or is partially available at Server S1. Before introducing these policies, we discuss an extension to the current MDP model.

Thus far, we assumed that Server S2 receives jobs only from Server S1. However, in practice, Server S2 may also receive jobs from another server independent of the jobs Server S2 receives from Server S1. For example, in mortgage application process, another Loan Officer may send a job to the Loan Processor independent of the Loan Officer at Stage 1 (i.e., independent of Server S1). We assume that the new arrivals are completely random and arrive independently according to a Poisson process with rate λ^e to Stage 2.

The question is whether the arrival of the new jobs to Stage 2 has any impact on the *structure* of the optimal control policy for Server S1? Do the structural properties in Theorem 2 and 3 still hold? In Proposition 3, we show that all the results discussed regarding the structure of the optimal control policy for Server S1, presented in Proposition 2, Theorem 2 and Theorem 3, still hold for $V^e(\mathbf{n})$, the value function with the new arrivals to Stage 2 defined in Equation (1.5). The proof is presented in Appendix for brevity.

Let $V^e(\mathbf{n})$ be the value function with the new arrivals to Stage 2. The optimality equation of the new MDP with the objective of minimizing average number of jobs in the system can be expressed as

$$\frac{g}{\Lambda^e} + V^e(\mathbf{n}) = \frac{1}{\Lambda^e}\Big\{n_1 + n_2 + n_3 + \lambda V^e(\mathbf{n}+\mathbf{e}^1) + \lambda^e V^e(\mathbf{n}+\mathbf{e}^2) + \mu_2 B^e(\mathbf{n}) + \min\Big\{A_I^e(\mathbf{n}), A_{P1}^e(\mathbf{n}), A_{P3}^e(\mathbf{n})\Big\}\Big\}, \tag{1.5}$$

where $\Lambda^e = \lambda + \lambda^e + \mu_1 + \mu_2 + \mu_3$ and $B^e(\mathbf{n})$, $A_I^e(\mathbf{n})$, $A_{P1}^e(\mathbf{n})$ and $A_{P3}^e(\mathbf{n})$ are defined below.

$$B^e(\mathbf{n}) = \begin{cases} V^e(\mathbf{n}) & : \text{ if } n_2 = 0 \\ pV^e(\mathbf{n}-\mathbf{e}^2+\mathbf{e}^3) + (1-p)V^e(\mathbf{n}-\mathbf{e}^2) & : \text{ if } n_2 > 0 \end{cases}$$

$$A_I^e(\mathbf{n}) = (\mu_1 + \mu_3)V^e(\mathbf{n})$$

$$A_{P1}^e(\mathbf{n}) = \begin{cases} \mu_1 V^e(\mathbf{n}) + \mu_3 V^e(\mathbf{n}) & : \text{ if } n_1 = 0 \\ \mu_1 V^e(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) + \mu_3 V^e(\mathbf{n}) & : \text{ if } n_1 > 0 \end{cases}$$

$$A_{P3}^e(\mathbf{n}) = \begin{cases} \mu_3 V^e(\mathbf{n}) + \mu_1 V^e(\mathbf{n}) & : \text{ if } n_3 = 0 \\ \mu_3 V^e(\mathbf{n}-\mathbf{e}^3) + \mu_1 V^e(\mathbf{n}) & : \text{ if } n_3 > 0 \end{cases}$$

Proposition 1.3. *All the results discussed regarding the structure of the optimal control policy for Server S1, presented in Proposition 2, Theorem 2 and Theorem 3, still hold for* $V^e(\mathbf{n})$*, defined in Equation (1.5).*

The intuition behind Proposition 3 is as follows. According to the optimal control policy, when n_2 is high enough (i.e., Server S2 is busy), Server S1 works at Stage 3 to complete more jobs. However, when n_2 is too low (i.e., Server S2 may become idle soon), Server S1 works at Stage 1 to avoid starvation of Stage 2. The addition of thenew arrivals to Stage 2 does not change the threshold-type *structure* of the optimal dynamic policy. It only changes the value of the thresholds on n_2. Therefore, the value function with additional arrivals satisfies the structural properties and results driven for Equation (1.1). Since the addition of the new arrivals do not provide any new insights, in the remainder of this chapter, we assume there is no additional arrival to Stage 2.

1.5. Control Policy Under No or Limited Information

As we mentioned, it is sometimes difficult (if not impossible) for Server S1 to monitor the number of jobs at Stage 2 due to the system's limitations. In the mortgage application process, the firm's access control procedure may restrict Server S1 to access Server S2's account information including the number of jobs at Stage 2. This is because Server S2 usually has a different access level due to higher organization rank or job description. Moreover, in some cases, the physical layout of the system and distance between servers may make it hard for the Server S1 to be able to estimate the number of jobs at Stage 2. The physical layout limitation is more prevalent in manufacturing systems. Thus, optimal control policy under full information would not always be applicable. Finally, even

when information about n_2 is available, as Figure 1.2 shows, the structure of the optimal dynamic policy is complex and hard-to-implement because it changes with the value of n_2. Therefore, there is a need for simple control policies that do not need information on n_2, while performing relatively *close-to-optimal*. In this section, we propose three policies to fulfill this need: Optimal Static policy, No-Information Threshold (NIT) policy and Partial-Information Threshold (PIT) policy.

1.5.1. Optimal Static Policy

Static priority policies, in general, give priority to a particular stage as long as there is a job at that stage. In the mortgage application process, there are two static priorities: (i) working on the new applications as long as there are new applications in the system, (ii) working on returning applications as long as there are returning applications in the system. To find the optimal static policy, we need to compare the long-run average number of jobs in the system when Server S1 gives priority to Stage 1 with that when the server gives priority to Stage 3. But one might ask when is it *optimal* to give priority to a particular stage?

To answer this question, we have the following theorem that provides a sufficient condition for static priority policy to be optimal.

Theorem 1.4. *If G_i, the processing time at each Stage i, has a general distribution and $P(G_2 < G_3 < G_1) = 1$, then it is optimal to give priority to Stage 3 when Stage 2 is not empty.*

If G_3 is always smaller than G_1, it is optimal for Server S1 to always work at Stage 3. This is because working at Stage 1 is less rewarding (it takes a long time for Server S1 to finish a job at Stage 1). When G_3 is always smaller than G_1, working at Stage 3 results in less total number of jobs in the system compared to working at Stage 1. In this case, it is optimal to give priority to Stage 3. Note that we need the processing time of Server S2 at Stage 2 to be smaller than that of Server S1 at Stage 3 to ensure that Stage 3 become empty faster than Stage 2. Otherwise, to prevent starvation of Stage 2, Server S1 has to work at Stage 1. Therefore, when Stage 2 is not empty and $P(G_2 < G_3 < G_1) = 1$, it is optimal to give priority to Stage 3.

Compared to optimal dynamic policy, a static priority policy does not use any information about the system state and is easier to implement. Server S1 continues to work at the prioritized stage as long as there is job in that stage. However, it is expected that static policy does not perform as well as the optimal dynamic policy. Even though a large gap is expected between the optimal static policy and the optimal dynamic policy, as we will show later, under some conditions beyond that in Theorem 4, the optimal static policy works relatively well.

1.5.2. No-Information Threshold (NIT) Policy

Even if the information on the number of jobs at Stage 2 is not available for Server S1, she could still observe the number of jobs at Stage 1 and 3 (n_1 and n_3). Therefore, one can use this information to construct a better control policy compared to the optimal static policy that does not use information about n_1 and n_3. Using the information on n_1 and n_3, No-Information Threshold (NIT) policy helps Server S1 to make the decision of

working at Stage 1 or Stage 3, when no information is available on n_2, as follows:

No-Information Threshold (NIT) Policy:

Under the NIT policy, Server S1 monitors n_1 in Stage 1 and n_3 in Stage 3 and makes her decisions based on two thresholds R_1 and R_3 as follows:

- As long as $n_3 < R_3$,
  - Server S1 works at Stage 1.
- Once $n_3 \geq R_3$ (i.e., the number of jobs at Stage 3 crosses threshold R_3)
  - If the number of jobs at Stage 1 is less than or equal to threshold R_1 (i.e., $n_1 < R_1$), Server S1 starts working at Stage 3.
  - Otherwise, she continues working at Stage 1.
- Server S1 idles only if there is no job at Stage 1 and Stage 3.

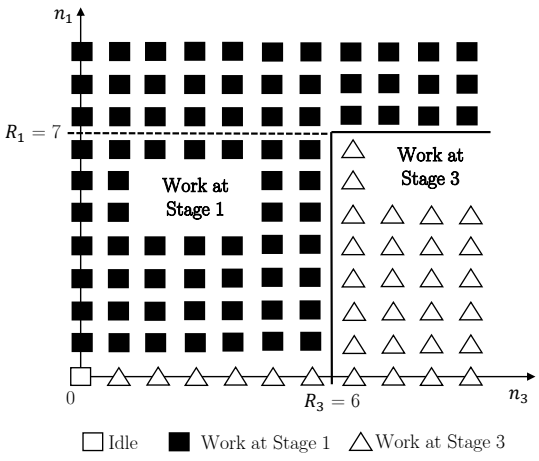


Figure 1.3. A Typical structure of NIT policy with $R_1 = 7$ and $R_3 = 6$.

Figure 1.3 shows a typical structure of NIT policy with $R_1 = 7$ and $R_3 = 6$. As shown in the figure, when $n_1 > R_1$, Server S1 works at Stage 1 and when $n_1 \leq R_1$, she works at Stage 1 only if $n_3 < R_3$. When $n_3 \geq R_3$, Server S1 works at Stage 3 under NIT policy.

Even though the lack of full information on n_2 is a possibility in the system under study, in some cases, Server S1 may be given access to some *partial information* about the number of jobs at Stage 2, namely, she may know if n_2 is large or small. In this case, using this additional information, she might be able to make a better decision of working at Stage 1 or Stage 3. Taking this into account, in the next section, we suggest another policy that can be used when partial information on n_2 is available.

1.5.3. Partial-Information Threshold (PIT) Policy

In this section, we propose a threshold policy under which Server S1 makes the decision of working at Stage 1 or Stage 3 using thresholds on the number of jobs at each stage. We call this policy *Partial-Information Threshold (PIT) policy*.

The intuition behind this policy is derived from the structure of the optimal dynamic policy under full information. The optimal dynamic policy under full information reveals that Server S1's decision of which stage to work at depends on n_2. When n_2 is small, the server is more likely to work at Stage 1 to feed Stage 2. When n_2 is large, the server is more likely to work at Stage 3 since it is expected that Stage 3 will receive a large number of jobs from Stage 2. PIT policy simplifies the optimal dynamic policy's complex threshold structure by defining a threshold on n_1, n_2 and n_3, separately. Under PIT policy, knowing if n_2 is small or large (i.e. whether it is larger or smaller than a threshold), Server S1 chooses to work at the stage with relatively larger number of jobs in

the system. More specifically, if n_1 exceeds a threshold, Server S1 works at Stage 1 and similarly if n_3 exceeds a different threshold, Server S1 works at Stage 3.

Based on this insight, PIT policy introduces threshold N_2 and two different sets of thresholds (Z_1, Z_3) and (S_1, S_3), respectively, depending on whether $n_2 \leq N_2$ or $n_2 > N_2$. The formal description of PIT policy is described below.

Partial-Information Threshold (PIT) Policy:

Under the PIT policy, Server S1 monitors n_1 and n_3, and makes her decision based on five thresholds Z_1, Z_3, S_1, S_3 and N_2 as follows:

- When $n_2 \leq N_2$,
  - If $n_3 \geq Z_3$, Server S1 works at Stage 3;
  - Otherwise, when $n_3 < Z_3$, if $n_1 > Z_1$, Server S1 works at Stage 1 or else (when $n_1 \leq Z_1$) Server S1 works at Stage 3.
  - Server S1 only idles when $n_1 = n_3 = 0$.
- Similarly, when $n_2 > N_2$,
  - If $n_3 \geq S_3$, Server S1 works at Stage 3;
  - Otherwise, when $n_3 < S_3$, if $n_1 > S_1$, Server S1 works at Stage 1 or else if $n_1 \leq S_1$ Server S1 works at Stage 3.
  - Server S1 only idles when $n_1 = n_3 = 0$.

Figure 1.4 shows a typical structure of PIT policy for set of thresholds $N_2 = 5$, $(Z_1, Z_3) = (1, 6)$ and $(S_1, S_3) = (2, 4)$. The graph on the left shows the structure of this policy when $n_2 \leq N_2$. In that graph, when $n_3 \geq Z_3$, Server S1 works at Stage 3 and when $n_3 < Z_3$, she works at Stage 3 only if $n_1 < Z_1$. When $n_3 \geq Z_3$, Server S1 works at

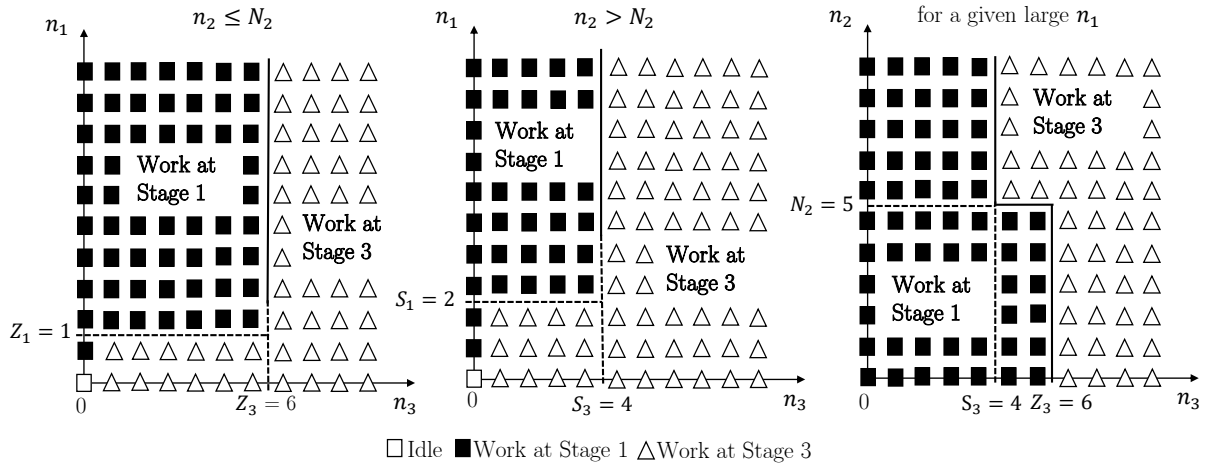


Figure 1.4. A Typical structure of the PIT policy (*Left*) for given n_2 when $n_2 \leq N_2$, (*Middle*) for given n_2 when $n_2 > N_2$, (*Right*) for given n_1.

Stage 3. If one of the stages is empty, she works at the other one, and if both stages are empty, she becomes idle. The graph in the middle shows the structure of PIT policy when $n_2 > N_2$. Similarly, when $n_3 \geq S_3$, Server S1 works at Stage 3 and when $n_3 < S_3$, she works at Stage 3 only if $n_1 < S_1$. When $n_3 \geq S_3$, Server S1 works at Stage 3. The graph on the right shows the structure of PIT policy when n_2 and n_3 are changing. When n_3 becomes larger than a threshold, PIT policy assigns Server S1 to Stage 3. The threshold is different for $n_2 \leq N_2$ and $n_2 > N_2$ and are Z_3 and S_3, respectively.

PIT policy is easier to implement than the optimal dynamic policy, which is a significant advantage in practice. One way to implement this policy in practice is *Kanban Board*. Similar to the idea of Visual Kanban in manufacturing systems, a Kanban board is a feature offered by process and document management softwares such as JIRA,[4] that let the servers track the status and the progress of their jobs in a service system.

[4] `https://www.atlassian.com/software/jira`

1.6. Numerical Analysis

To characterize the performance of static, NIT and PIT policies, we design a numerical study to answer the following questions for each policy:

(1) How does each proposed policy perform compared to the optimal dynamic policy under full information (i.e., MDP)?

(2) What is the impact of system parameters on the performance of the policy?

(3) How robust is the performance of the policy with respect to errors in setting their thresholds?

(4) How robust is the performance of the policy with respect to variability in the system?

In this section, we first explain how we systematically design an experiment to study these questions. To answer the first question, we perform a numerical study to compare the performance of all the proposed polices with that of the optimal dynamic policy. To answer the second question, we perform a sensitivity analysis on system parameters. Finally, we do robustness analysis with respect to thresholds used in NIT and PIT policies and service time distributions to further confirm our findings.

1.6.1. Design of Numerical Study

Our numerical study includes a total of 81 cases generated using the parameters in Table 2.1. We consider low ($p = 0.2$), medium ($p = 0.5$) and high ($p = 0.8$) percentage of type-2 jobs. We denote the utilization of server i by ρ_i. Since Server S1 works at both Stage 1 and 3, the utilization of Server S1 (i.e., ρ_1) is $\rho_1 = \lambda/\mu_1 + p\lambda/\mu_3$. The utilization of Server S2 (i.e., ρ_2) is $\rho_2 = \lambda/\mu_2$. In Table 2.1, we consider the case of low utilization

($\rho_i = 0.5$), medium utilization ($\rho_i = 0.75$) and high utilization ($\rho_i = 0.95$) for servers $i = 1, 2$. Finally, as shown in Table 2.1, there are 3 possible scenarios for μ_1 and μ_3 to consider: $\mu_1 = \mu_3$, $\mu_1 < \mu_3$ and $\mu_1 > \mu_3$.

Table 1.1. Parameters of the Experiment.

p	ρ_1	ρ_2	(μ_1, μ_3)
0.2	0.5	0.5	(10, 10)
0.5	0.75	0.75	(5, 15)
0.8	0.95	0.95	(15, 5)

To generate a case for our numerical study, we first choose a value for ρ_1 from three choices of: low utilization ($\rho_1 = 0.5$), medium utilization ($\rho_1 = 0.75$), and high utilization ($\rho_1 = 0.95$). We then choose a value for ρ_2 from three choices of: low utilization ($\rho_2 = 0.5$), medium utilization ($\rho_2 = 0.75$), and high utilization ($\rho_2 = 0.95$). Given the values of ρ_1 and ρ_2, for each combination of (μ_1, μ_3), p and ρ_1, we calculate λ as follows:

$$\lambda = \rho_1/(1/\mu_1 + p/\mu_3)$$

Finally, we calculate μ_2 based on λ and ρ_2 as follows:

$$\mu_2 = \lambda/\rho_2$$

For example, we choose $\rho_1 = 0.5$, $\rho_2 = 0.5$, $(\mu_1, \mu_3) = (10, 10)$ and $p = 0.5$. For this combination of (μ_1, μ_3), p and ρ_1, we calculate $\lambda = \frac{\rho_1}{1/\mu_1 + p/\mu_3} = \frac{0.5}{1/10 + 0.5/10} = 3.33$. Finally, using the values of λ and ρ_2, we calculate $\mu_2 = \lambda/\rho_2 = 3.33/0.5 = 6.67$. We compare the performance of each proposed policy with the optimal dynamic policy for all 81 generated cases.

1.6.2. Performance Analysis

In this section, we answer the question of how the proposed heuristic policies perform compared to the optimal dynamic policy under full information. To evaluate the performance of a policy, we define Performance Loss (PL) of the policy as follows:

$$\mathrm{PL}_{policy} = \frac{(E[N]_{policy} - E[N]_{MDP})}{E[N]_{MDP}}$$

Hence, PL represents how much (in percentage) the long-run average number of jobs under a policy is larger than that under the optimal dynamic policy (i.e., MDP), which corresponds to the case with full information about n_2. We use value iteration algorithm to find $E[N]_{MDP}$. The stopping criteria is set at $\epsilon = 0.001$ (i.e., the algorithm stops when the difference in $E[N]$ in two consecutive iterations is less than 0.1%).

To determine the optimal thresholds (R_1^*, R_3^*) for NIT policy, we enumerate all possible candidates for (R_1, R_3) and find the long-run average number of jobs in the system by forcing the value iteration algorithm to follow NIT policy for each candidate (R_1, R_3), instead of choosing the action that minimizes the value function. Optimal values (R_1^*, R_3^*) are the threshold numbers that result in the minimum long-run average number of jobs in the system among all possible candidates for (R_1, R_3). We call the NIT policy with the optimal thresholds (R_1^*, R_3^*), the *optimal NIT policy*.

To find the long-run average number of jobs in the system under the optimal static policy, we find the long-run average number of jobs in the system by forcing the value iteration algorithm to follow (1) static priority policy to Stage 1, and (2) static priority

policy to Stage 3. The minimum of the two values is the long-run average number of jobs in the system under the optimal static policy.

To determine the optimal thresholds $(N_2^*, Z_1^*, Z_3^*, S_1^*, S_3^*)$ for PIT policy, we enumerate all possible candidates for $(N_2, Z_1, Z_3, S_1, S_3)$ and find the long-run average number of jobs in the system by forcing the value iteration algorithm to follow PIT policy for each candidate $(N_2, Z_1, Z_3, S_1, S_3)$, instead of choosing the action that minimizes the value function. Optimal values $(N_2^*, Z_1^*, Z_3^*, S_1^*, S_3^*)$ are the threshold numbers that result in the minimum long-run average number of jobs in the system among all possible candidates for $(N_2, Z_1, Z_3, S_1, S_3)$. We call the PIT policy with the optimal thresholds $(N_2^*, Z_1^*, Z_3^*, S_1^*, S_3^*)$, the *optimal PIT policy.*

Note that when $n_2 \leq N_2^*$ (i.e., the number of jobs at Stage 2 is small), giving priority to Stage 1 decreases the long-run average number of jobs in the system more than giving priority to Stage 3, since it prevents the starvation of Stage 2. On the other hand, when $n_2 > N_2^*$ (i.e., the number of jobs at Stage 2 is large), giving priority to Stage 3 decreases the long-run average number of jobs in the system more than giving priority to Stage 1, since it immediately reduces the number of jobs in the system. As it is shown in Figure 5, for Server S1 to work more often at Stage 1, we need a smaller threshold for the number of jobs at Stage 1 when $n_2 \leq N_2^*$ compared to when $n_2 > N_2^*$. Therefore, the threshold chosen for the number of jobs at Stage 1 is smaller, when $n_2 \leq N_2^*$ compared to when $n_2 > N_2^*$, i.e., $Z_1^* \leq S_1^*$. Similarly, for Server S1 to work more often at Stage 3, we need a smaller threshold for the number of jobs at Stage 3 when $n_2 > N_2^*$ compared to when $n_2 \leq N_2^*$. Thus, the threshold chosen for the number of jobs at Stage 3 is expected to be smaller, when $n_2 > N_2^*$ compared to when $n_2 \leq N_2^*$, i.e., $Z_3^* \geq S_3^*$.

We evaluate a policy by the average PL value, the worst case PL value and by the percentage of cases with PL values less than 5% and 10%. Table 1.2 summarizes the performance of the optimal static policy, optimal NIT policy and optimal PIT policy compared to the optimal dynamic policy.

Table 1.2. Summary of the performance of the Optimal Static, optimal NIT and optimal PIT Policies

Item	Optimal Static Policy	Optimal NIT Policy	Optimal PIT Policy
Average PL	4.7%	3.1%	0.6%
Max PL	23.9%	12.7%	3.6%
% of cases with $PL < 10\%$	83%	95%	100%
% of cases with $PL < 5\%$	70%	77%	100%

These results demonstrate that in some cases there may be a large gap between the performance of the optimal static policy and optimal NIT policy and that of the optimal dynamic policy. The difference, however, is because of the nature of these policies that do not utilize the information about the number of jobs in Stage 2 (i.e., n_2). Nevertheless, as the average PL and the percentage of cases with the gap of less than 5% and 10% show, these policies perform reasonably well where there is a complete lack of information on the number of jobs at Stage 2 given that they are set optimally. Table 1.2 also shows that optimal PIT policy has an average PL_{PIT} value of only 0.6% and worst case PL_{PIT} value of only 3.6%. This implies that not having *full* information is not critical in making a near-optimal decision in this setting. By having *partial* information of whether n_2 is above or below a threshold, we can capture the most benefit of full information.

1.6.3. Sensitivity Analysis

In this section, we investigate the impact of system parameters on the performance of optimal static, NIT and PIT policies. Specifically, we are interested in finding system

characteristics under which our proposed policies perform well. Table 1.3 demonstrates the summary of our observations.

Table 1.3. Extended Numerical Analysis Summary.

		optimal static policy		optimal NIT policy		optimal PIT policy	
	Parameter	Avg PL	Max PL	Avg PL	Max PL	Avg PL	Max PL
Impact of ρ_1	0.5	2.5%	12.2%	2.3%	12.2%	0.1%	0.7%
	0.75	3.6 %	12.9%	3.3 %	10.1%	0.3%	0.9%
	0.95	8.1 %	23.9%	3.8 %	12.7%	1.2%	3.6%
Impact of ρ_2	0.5	2.5%	16.2%	1.2%	3.62%	0.4%	3.6%
	0.75	5.2%	23.7%	3.1%	9.4%	0.6%	3.6%
	0.95	6.6%	23.9%	5.1%	12.7%	0.7%	2.5%
Impact of μ_1, μ_3	$\mu_1 < \mu_3$	1.4%	7.1%	1.4%	7.1%	0.2%	1.1%
	$\mu_1 = \mu_3$	7.0%	23.9%	4.1%	11.3%	0.6%	2.9%
	$\mu_1 > \mu_3$	5.7%	12.0%	3.8%	12.7%	0.8%	3.6%
Impact of p	0.2	2.46%	12.20%	1.29%	4.07%	0.4%	2.9%
	0.5	4.94%	17.99%	3.32%	8.15%	0.6%	3.6%
	0.8	6.99%	23.91%	4.74%	12.69%	0.7%	3.6%

Observation 1 *The long-run average number of jobs under the optimal static, optimal NIT and optimal PIT policies are close to that under the optimal dynamic policy when Server S1 has low utilization (i.e., ρ_1 is small).*

As Table 1.3 shows, as ρ_1 becomes larger, we observe that PL_S, PL_{NIT} and PL_{PIT} increase, where PL_S, PL_{NIT} and PL_{PIT} are performance loss of the optimal static policy, optimal NIT policy and optimal PIT policy, respectively. The intuition behind this observation is that when ρ_1 is large, Server S1 is very busy and is the bottleneck. In this case, the benefit from optimally allocating Server S1 between Stages 1 and 3 becomes significant.

Observation 2 *The long-run average number of jobs under the optimal static, optimal NIT and optimal PIT policies are close to that under the optimal dynamic policy when Server S2 has low utilization (i.e., ρ_2 is small).*

As shown in Table 1.3, when ρ_2 is small, the average PL of all heuristic policies is small and therefore they work well. The intuition behind this observation is that when ρ_2 is small, the probability of having large number of jobs at Stage 2 (i.e., n_2 being large) is small and a new job from Stage 1 is expected to be processed soon after it arrives to Stage 2. In this situation, it is less important for Server S1 to know n_2 at Stage 2, when making her decision of which stage to work on. Therefore, even the optimal static and optimal NIT policies, which do not take into account any information about n_2, work relatively well.

Observation 3 *The long-run average number of jobs under the optimal static, optimal NIT and optimal PIT policies are close to that under the optimal dynamic policy when the processing of type-2 jobs at Stage 3 takes less time than the processing of type-1 jobs*

at Stage 1 (i.e., $\mu_1 < \mu_3$).

As shown in Table 1.3, when $\mu_1 < \mu_3$, all policies perform well. Their performance becomes worse when $\mu_1 > \mu_3$. The intuition behind this observation is as follows. When $\mu_1 < \mu_3$, all policies will give priority to Stage 3 in most cases. Similarly, since the processing rate at Stage 3 is higher than that at Stage 1, the optimal dynamic policy also gives priority to Stage 3, which leads to processing more jobs and hence reducing the number of jobs in the system at a faster rate. Hence, the performance of all policies is close to that of the optimal dynamic policy and thus all policies perform well.

Observation 4 *The long-run average number of jobs under the optimal static, optimal NIT and optimal PIT policies are close to that under the optimal dynamic policy when the percentage of type-2 jobs is small (i.e., p is small)*

As shown in Table 1.3, when p is small, all policies perform well. The intuition behind this observation is that when p is small, the number of jobs at Stage 3 (i.e. n_3) is small and there are not many decisions to make. This observation implies that optimal static and optimal NIT policies can be used for systems in which the fraction of returning jobs (i.e., p) is small.

Observations 1 to 4 show conditions under which policies with no or partial information on the number of jobs at Stage 2 work well. Under these conditions, with no information, Server S1 can capture the benefit of having full information. The small average (0.3%) and worst case (3.6%) PL of optimal PIT policy show that this policy performs close to

the optimal dynamic policy in all cases. Nonetheless, the long-run average number of jobs under PIT policy is the closest to that under optimal dynamic policy when (i) Server S1 has low utilization (i.e., ρ_1 is small), or (ii) Server S2 has low utilization (i.e., ρ_2 is small), or (iii) Processing of type-2 jobs at Stage 3 takes less time than the processing of type-1 jobs at Stage 1 (i.e., $\mu_1 < \mu_3$), or (iv) Percentage of type-2 jobs is small (i.e., p is small).

The optimal static and optimal NIT policies do not always perform well. The worst performance is observed when the utilization of Stage 1 and Stage 2 (i.e. ρ_1 and ρ_2) are both high and the percentage of type-2 jobs is large (i.e., p is big). As the system becomes more congested and has more type-2 jobs, it is more critical to have full information on the number of jobs at Stage 2 to be able to make a more informed decision of which stage to work at. Therefore, utilizing the optimal static policy and optimal NIT policy, when the information about the number of jobs at Stage 2 is not available and certain conditions are not met, lead to much worse performance compared to that of the optimal dynamic policy under full information.

1.6.4. Robustness Analysis

In this section, we discuss two robustness checks for our numerical analysis. First, we check the impact of errors in setting the threshold of NIT and PIT policies. It is possible that we make errors when estimating system parameters or when computing optimal thresholds for PIT and NIT polices. This leads us to setting sub-optimal values for the thresholds of NIT and PIT policies. We analyze how much we lose by using sub-optimal thresholds. Second, we check the impact of variability of service times on the performance of the proposed heuristic policies. In previous sections, we assumed that the service times

are exponentially distributed. In this section, we check the robustness of our results when service times have non-exponential distributions.

1.6.4.1. Robustness of Optimal NIT and Optimal PIT Policies. In this section, we run additional numerical studies to test the robustness of our findings with respect to the errors in setting the optimal thresholds for optimal NIT and optimal PIT policies. To implement optimal NIT and optimal PIT policies, we need to compute the optimal sets of thresholds (R_1^*, R_3^*) and $(N_2^*, Z_1^*, Z_3^*, S_1^*, S_3^*)$, respectively. The optimal values of the thresholds can be found by searching for the threshold numbers that result in the minimum long-run average number of jobs in the system. However, it is possible that errors are made when estimating system parameters (e.g., λ, μ_1, μ_2, μ_3, p) or when computing the threshold numbers. Thus, the thresholds would be sub-optimal. An interesting question that arises is: how much one loses by using sub-optimal thresholds? In other words, how sensitive is the optimal long-run average number of jobs in the system with respect to making errors in setting the thresholds?

To answer these questions, we check the robustness of the performance of optimal NIT and optimal PIT policies with respect to their thresholds. We recompute the long-run average number of jobs in the system when thresholds are set 10% and 20% below or above the optimal thresholds. We then compare the performance of the policy with sub-optimal thresholds with that of the policy with the optimal thresholds and with the optimal dynamic policy. Please see Appendix 2 for the details of our robustness analysis. We define Performance Loss (PL) of the policy with sub-optimal thresholds compared to

that with optimal thresholds and with the optimal dynamic policy as follows:

$$\mathrm{PL}_{Suboptimal} = \frac{E[N]_{\mathrm{Suboptimal}} - E[N]_{\mathrm{Optimal}}}{E[N]_{\mathrm{Optimal}}},\ \mathrm{PL}_{Sub-MDP} = \frac{E[N]_{\mathrm{Suboptimal}} - E[N]_{\mathrm{MDP}}}{E[N]_{\mathrm{MDP}}}$$

Hence, $\mathrm{PL}_{Suboptimal}$ represents how much (in percentage) the long-run average number of jobs under a policy with sub-optimal thresholds is larger than that under the smae policy with optimal thresholds. Similarly, $\mathrm{PL}_{Sub-MDP}$ represents how much (in percentage) the long-run average number of jobs under a policy with sub-optimal thresholds is larger than that under the optimal dynamic policy.

We observe that the performance of optimal NIT and optimal PIT policies were not significantly affected by using sub-optimal thresholds. More specifically, for optimal NIT policy, the average $\mathrm{PL}_{Suboptimal}$ of using sub-optimal thresholds is less than 0.5%. The average PL Sub-MDP is less than 3.7%. The worst performance is observed when the values of R_3 is 20% above or below the optimal value. This indicates that the optimal long-run average number of jobs in the system is more sensitive to the value of R_3 than R_1. Under NIT policy, when $n_3 \geq R_3$, Server S1 works at Stage 3 and by completing a job at that stage she reduces the number of jobs in the system. Therefore, setting R_3 sub-optimally affects the number of jobs in the system. This is the reason that the system is more sensitive to the value of the threshold R_3.

The average PL of using sub-optimal thresholds $(N_2^*, Z_1^*, Z_3^*, S_1^*, S_3^*)$ is less than 2% for optimal PIT policy,. The average PL Sub-MDP is less than 2.6%. The worst performance is observed when the values of N_2 is 10% or 20% below the optimal value and when the values of S_3 and Z_3 is 10% or 20% above the optimal value. As we discussed earlier, Server S1's decision of which stage to work depends on n_2. When n_2 is small, the server

is more likely to work at Stage 1 to feed Stage 2; and when n_2 is large, the server is more likely to work at Stage 3. Therefore, setting N_2 sub-optimally affects which stage Server S1 has to work at and this directly affects the number of jobs in the system. This is the reason that the system is more sensitive to the value of the threshold N_2. Similar to the argument for (R_1, R_3), the system is more sensitive to the value of threshold Z_3 and S_3 than that of thresholds Z_1 and S_1, since setting S_3 and Z_3 sub-optimally directly affects the number of jobs in the system. The worst PL is observed when utilization of Server S1 or Server S2 is high and return probability p is also high.

1.6.4.2. Non-exponential Service Times. In this section, we run additional numerical studies to test the performance of our proposed policies when service times have a non-exponential distribution. Without loss of generality and for tractability purposes, we only consider the case where service time at Stage 2 is non-exponential. Stage 2 is selected since the optimal dynamic policy under full information structure revealed that Server S1's decision of which stage to work at depends on Stage 2's state. Note that relaxing the exponential assumption for all stages makes the state space very large and thus makes the numerical method intractable. We chose Gamma distribution for service time at stage 2, since it has CV less, equal to or greater than one.

Consider the MDP model presented in the chapter, except than the service time at Stage 2 is not exponentially distributed. To analyze this system, we discretize the time horizon into equal, nonoverlapping infinitesimal intervals δt, where $\delta t \to 0$. We refer to Appendix 2 for the detail of numerical analysis on non-exponential service times, the MDP model descriptions and optimality equations.

For cases with CV = 0.5, we find that optimal PIT policy performs very well compared to the optimal dynamic policy with average Performance Loss (PL) of 0.9% and maximum PL of 5.6%. The average PL for the optimal static policy was 7.3% and for optimal NIT policy was 4.4%. For the cases with CV = 2, we find that optimal PIT policy performs well compared to the optimal dynamic policy with average PL of 2.0% and maximum PL of 9.2%. The average PL for the optimal static policy was 16.6% and for optimal NIT policy was 10.4%. Similar to the observations made for exponential service times, we observe that under similar conditions (i.e., low utilization of Server 1 and Server 2 and low return probability p) the optimal static, optimal NIT and optimal PIT policies work well compared with the optimal dynamic policy under full information.

1.7. Conclusion

In this chapter, we studied a two-stage system with a flexible server who processes two types of jobs. We characterized the optimal control policy for Server S1, that minimizes the average number of jobs in the system. The optimal dynamic policy has a monotone threshold structure with respect to the number of jobs at Stage 1 and 3. This structure recommends working at Stage 1 if there are more jobs at Stage 1, and working at Stage 3 if there are more jobs at Stages 2 or 3.

The structure of the optimal dynamic policy is very complex and, in most cases, is not suitable for practical use. Moreover, due to system's limitations, information on the state of downstream stage may not be fully available. To fill the need for more practical policies that can deliver good performances when there is lack of information on the state of the downstream stage, we proposed two heuristic policies (i.e. NIT and PIT). We then

compared the performances of these heuristics with that of the optimal dynamic policy. We found that PIT policy consistently performs well. We also found conditions under which the optimal static and NIT policies perform close to the optimal dynamic policy with full information. Specifically, when Server S1 and Server S2 have low utilization (i.e., ρ_1 and ρ_2 are small) and the percentage of type-2 jobs (i.e. p) is small, NIT and Static policy work well. This implies that these policies are good candidates for systems that a small fraction of jobs are sent back to Stage 1 (i.e., $p \leq 0.2$). In addition, the proposed heuristic policies are simpler and more easily implementable alternatives to the complex optimal dynamic policy.

There are two possible directions for future study. One is extending the current two-stage system to general multi-stage systems under different scenarios for the information about the downstream stages. The other is to study scenarios with multiple flexible servers. When there are multiple flexible servers, the optimal decision for one server will not only depend on the state of the system, but also depend on the decisions of other servers. One can also consider positive switchover time/cost as a more complex case for further research.

CHAPTER 2

Optimal Policy in Single-Server Multi-Class Queuing Systems with Abandonments

2.1. Introduction

The key objective of most service operations is to improve operational performance of a congested system Zeltyn and Mandelbaum (2005). In this paper, we focus on customer abandonment as a performance measure. This measure is of a great importance in many service systems, especially call centers (Gans et al. (2003), Brown et al. (2005) and Aksin et al. (2007)). Customers waiting in such systems may become impatient and abandon the system, which results in loss of revenue often modeled as cost of abandonment (Gans et al. (2003)). Marchex institute, a mobile advertising analytics company, recently published a report, titled *"America's Call Center Revealed"*(Busby and Wisehart (2016)). In this report, they reveal the first lesson for call centers that is to answer the phone calls as quickly as possible, since for call center industries, 11%- 14% of customers hang up during a call.

A good example of service systems that abandonment cost plays the main role is infomercial call centers. *Infomercials*, also called long-form television commercials, are 30-minute programs designed to motivate viewers to place an order by phone. The infomercial industry is worth over $200 billion in United States Bogle (2014). An infomercial

advertises items ranging from household, automotive, health, and beauty products, to fitness products, books, or to toys for children. These items may be promoted as *not sold in stores* and therefore, consumer can only purchase the items by calling a phone number Johnson (2013). Calls received during infomercials usually outsourced to call centers, in which a *Customer Service Representative* (CSR) may answer the calls typically in what is known in call center industry as a shared model. In the shared model, CSRs are shared among several brands, businesses or products. Ordertaking businesses like infomercials usually estimate the *opportunity cost of lost sales* due to abandonment as the main cost that is directly associated with system performance (for example, see Andrews and Parsons (1993) and Akşin and Harker (2003)). Hence, it is important for infomercial call centers to answer the calls in an order that minimizes the number of unanswered calls (i.e., abandoned calls), since every abandoned call is a loss of revenue.

Call centers that use toll-free services typically pay out-of-pocket for the time their customers spend waiting on hold (i.e., line occupancy) Gans et al. (2003). For infomercial call centers, however, due to the limited time available for callers to contact (usually 30 minutes), the line occupancy cost is negligible compared to the cost of losing customers due to abandonment. Reported by Marchex institute Busby and Wisehart (2016), an infomercial call center that receives 3,000 calls a day, may earn up to $150 million in revenue per year. Considering a line occupancy cost per minute per call of $0.05 for a typical call center Gans et al. (2003), the cost of losing a customer can be up to 50 times higher. For example, for a call center that operates 360 days per year/24 hours per day, the revenue is around $20,000 per hour, while the line occupancy expenses are 50

times smaller and around $400 per hour. Hence, the main effort in call centers, especially infomercial call centers, focuses on minimizing abandonment cost.

How should a CSR, in a shared model, decide which call to answer next to minimize abandonment cost? To answer this question, there are three factors to consider.

(i) ***Prioritizing more valuable customers***: We may prioritize customers who are more costly to abandon than others. For instance, the cost of losing a customer who wants to purchase a TV is much higher than that of a customer who wants to purchase an all-purpose cleaner.

(ii) ***Prioritizing less patient customers***: We may prioritize less patient customers to prevent them from leaving. For example, the abandonment probability of a customer who wants to purchase an item for which there are not that many similar outside options (e.g., a unique exercise equipment) is lower than that of a customer who wants to purchase an item with many similar outside options (e.g., an all-purpose cleaner).

(iii) ***Prioritizing customers with shorter processing times***: We may prioritize the customers who we can serve faster. Serving customers with shorter processing time results in less customer in the queue who may potentially abandon at any moment. This, in turn, reduces the abandonment cost. In our example, it takes longer to answer a call of a customer who is willing to purchase a TV than a customer who wants to buy an all-purpose cleaner, since there are more technical details to be explained.

All above factors impact the loss of revenue (i.e., abandonment cost). But what is the optimal scheduling decision that captures the trade-off among the above three factors?

Should we always prioritize one type of customers over the other (i.e., a static priority policy)? Or the answer depends on the number of customers of each type on hold (i.e., a dynamic priority policy)?

To gain insights into dynamics of this scheduling decision, we study a single-server multi-class queuing system with customer abandonment. There is an extensive literature on design and control of multi-class queuing system with impatient customers in call centers (see for example Garnett et al. (2002), Atar et al. (2004), Jouini et al. (2009) and Kim et al. (2016)). The focus of these papers is on staffing decisions or deriving system's performance measures, while we focus on server's scheduling decision minimizing the abandonment costs. There are several papers such as Atar et al. (2010), Atar et al. (2011), Kim and Ward (2013) and Ata and Tongarlak (2013), which focus on server's scheduling decisions, considering customer abandonment as a modeling feature but all focus on minimizing the holding cost solely. More recently, Salch et al. Salch et al. (2013) consider a stochastic scheduling problem with impatient jobs and develop optimal policies for a single machine to minimize the expected weighted number of late jobs. In all these papers, abandonment does not incur costs. In our paper, as mentioned above, we consider the abandonment cost as our main objective and the goal is to minimize the total expected abandonment cost per unit time.

Our work is most closely related to Atar et al. (2010). Analyzing a multi-class queuing system with multiple servers and customer abandonment, Atar et al. Atar et al. (2010) introduce a server-scheduling policy that assigns priority to classes according to the index $c\mu/\theta$ when minimizing holding cost, where θ is the abandonment rate, c is the holding cost and μ is the service rate. Considering b as the penalty incurred whenever a customer

abandons the queue, they *conjecture* that under a *many-server* fluid scaling and *overload* conditions, minimizing only the total abandonment cost leads to the $b\mu$ policy, which is independent of abandonment rate θ. But this conjecture does not hold outside the many-server queuing system under fluid scaling. For those systems, the optimal scheduling policy does depend on abandonment rate θ. For example, in a simple two-class example where $b_1 = b_2$ and $\mu_1 = \mu_2$, one can intuitively argue that when the abandonment rate of one customer class is much smaller than the other (e.g., $\theta_1 << \theta_2$), the optimal policy that minimizes the abandonment cost depends on the abandonment rate and serves the customer class with larger abandonment rate (i.e., type-2 customers), since less type-1 customers leave the system.

In this paper, we show that the server's optimal scheduling policy is not independent of θ. Formulating the problem as a Markov Decision Process (MDP), we show that the optimal servers scheduling policy is a static priority policy and derive sufficient conditions under which the $b\mu$–rule is optimal. Performing a numerical study, we also show that, when these conditions do not hold, in a system with low/medium utilization, the $b\mu\theta$–rule performs as good as the optimal policy.

This paper proceeds as follows. In Section 2, we introduce the model specification and in Section 3, we present the corresponding MDP formulation. Section 4 discusses the general characteristics of the optimal policy and Section 5 summarizes the scheduling policies under special settings. In Section 6, in a numerical study, we study the performance of the scheduling policy when sufficient conditions do not hold. Moreover, we compare the performance of the optimal scheduling policy with a First-Come First-Served (FCFS) policy, which is often used in call centers. We conclude the paper in Section 7.

2.2. The Model Formulation

To get insight into the structure of the optimal scheduling policy, we consider a single server queue with M classes of arriving customers. Customers of type i arrive at queue i according to a Poisson process with rate λ_i. We assume that the service time for the type-i customer is exponentially distributed with rate μ_i and customers may abandon the system while waiting for service if their waiting time is too long as well as while receiving service, similar to Phung-Duc and Kawanishi (2014), Righter (2000) and Ward and Glynn (2003). We assume that the time until a type-i customer abandons the system follows an exponential distribution with rate θ_i. The assumption of customers abandonment during service is not unreasonable. Reports have shown that customers are often put on hold during service for some time, resulting in some customers hanging up (ICMI[1] reports).

We also assume that customer balking cost as well as the customer holding cost is negligible. This is the case in infomercial call centers, where callers do not hear busy signals due to the large number of lines (i.e., large buffer size). Strategic Contact Inc.[2], leading company in contact center strategy, operations, and technology, reports that the balking (blocking) rate is generally kept very low (under 1%) in such call centers.

Considering the long-run average abandonment cost, the server's scheduling problem is to determine which type of customer to serve next upon a service completion or upon a new arrival. The abandonment cost per unit time for type-i customer is a constant b_i. We assume that the preemption is allowed, i.e., CSRs may interrupt the service of a customer

[1]The Incoming Call Management Institute (ICMI), a highly reputable industry association, regularly tracks published industry statistics from several sources Aksin et al. (2007).
[2]`www.strategiccontact.com`

and put her on hold to serve another customer. There have been practices reported by ICMI that states caller were put on hold after they had started talking to a CSR.

Finally, we note here that the exponential assumption regarding the service times and abandonment times allows us to formulate the server's scheduling problem as a Markov Decision Process (MDP) and characterize the optimal policy. After our MDP reveals the structure of the optimal scheduling policy, we observe that our main insights are not influenced by these assumptions.

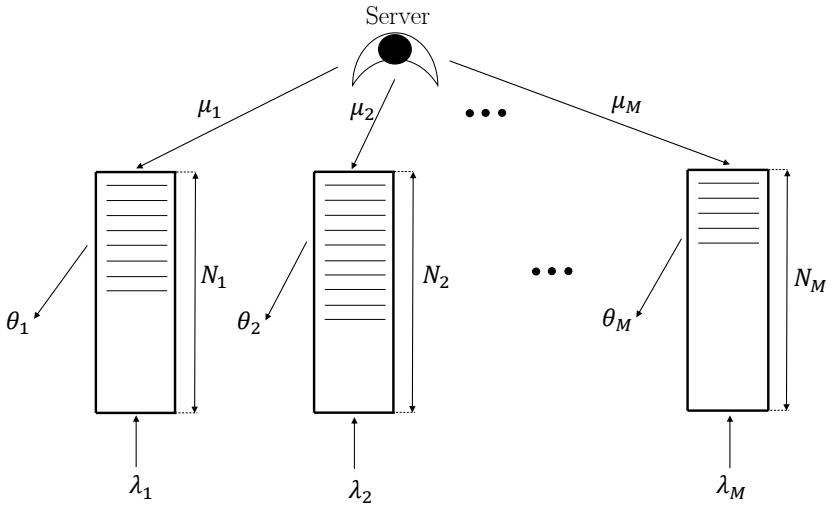
2.3. Markov Decision Process

To model this problem as a Markov Decision Process, we assume a maximum limit for the number of type-i customers allowed in the system, denoted by N_i. In other words, we assume customers of type i who find the system full (with N_i customers) do not enter the system (see Figure 2.1). This serves two purposes: (i) It prevents the uniformization rate (i.e., total transition rate) approaches infinity and thus allows us to model the queuing dynamics of the system (see optimality equation (2.1)); (ii) The systems with no limit on queue size are a special case of the more general systems with finite limit on queue size N_i, when $N_i \to \infty$.

With the objective of minimizing the total average abandonment cost per unit time, we may characterize server's scheduling decision using MDP as follows:

- *Decision epochs* are customer arrivals, service completion times and abandonment epochs.
- *State Space* $\mathcal{N}$ is a set of M-dimensional vectors $\mathbf{n} = (n_1, n_2, \ldots, n_M)$ where $0 \leq n_i \leq N_i$ is the number of customer of type $i \in \{1, 2, \ldots, M\}$ in the system.

Figure 2.1. A single-server multi-class queue with M classes of customers



- *Actions*: Considering $A_{\mathbf{n}}$ as the set of allowable actions at state $\mathbf{n}$, the action set $\mathcal{A}$ is therefore $\mathcal{A} = \cup_{\mathbf{n}\in\mathcal{N}} A_{\mathbf{n}}$. For any $\mathbf{n} \in \mathcal{N}$, allowable actions for the server are: *Serving* customer of type $j \in \{1, 2, \ldots, M\}$, if $n_j > 0$; and *Idling*.

Let $\mathbf{I}^i_{\{\mathbf{R}\}}$ be defined as follows:

$$\mathbf{I}^i_{\{\mathbf{R}\}} = \begin{cases} \mathbf{e}^i & \text{if R is true,} \\ \mathbf{0} & \text{otherwise,} \end{cases}$$

where $\mathbf{0}$ is an M-dimensional zero row vector, and $\mathbf{e}^i$ is an M-dimensional row vector with 1 on its i^{th} entry and 0 elsewhere. Let $\mathcal{J}_{\mathbf{n}}$ be the set that includes the indices of nonempty queues at state $\mathbf{n}$; then, for example, for state $\mathbf{n} = (0, 1, 2, 0, 8)$ we have $\mathcal{J}_{\mathbf{n}} = \{2, 3, 5\}$. Using the Lippmann's uniformization approach Lippman (1975), the optimality equations

for the MDP with the objective of minimizing abandonment cost is:

$$\begin{aligned}\frac{g}{\Lambda}+V(\mathbf{n}) =& \frac{1}{\Lambda}\Big\{\sum_{i=1}^{M} b_i\theta_i n_i+\sum_{i=1}^{M}\lambda_i V(\mathbf{n}+\mathbf{I}^i_{\{n_i<N_i\}})\\ &+\sum_{i=1}^{M} n_i\theta_i V(\mathbf{n}-\mathbf{I}^i_{\{n_i>0\}})+\sum_{i=1}^{M}(N_i-n_i)\theta_i V(\mathbf{n})+f(\mathbf{n})\Big\},\end{aligned} \tag{2.1}$$

where

$$f(\mathbf{n})=\min\begin{cases}\sum_{i=1}^{M}\mu_i V(\mathbf{n}) & \text{Idle}\\ \min_{j\in\mathcal{J}_{\mathbf{n}}}\{\mu_j V(\mathbf{n}-\mathbf{e}^j)+\sum_{i=1,i\neq j}^{M}\mu_i V(\mathbf{n})\} & \text{Serve type-}j\text{ customer}\end{cases}$$

where $\Lambda=\sum_{i=1}^{M}(\lambda_i+\mu_i+N_i\theta_i)$ is the uniformization rate and g is the total average cost per unit time.

2.4. Characteristics of the optimal policy

In this section, first, we discuss the existence of a stationary optimal policy of MDP problem (2.1).

Theorem 2.1. *There exists a stationary average-cost optimal policy for MDP problem (2.1).*

All proofs are in the On-line Appendix. Define difference operator $\mathbf{D}_x$ as

$$\mathbf{D}_j V(\mathbf{n})=V(\mathbf{n})-V(\mathbf{n}-\mathbf{e}^j),\quad j=1,2,\ldots,M.$$

Now, we can formulate the following proposition:

Proposition 2.2. *The optimality equation (2.1) has the following property:*

Property **P1:** $0 < \frac{b_j\theta_j}{\Lambda} \leq \mathbf{D}_j V(\mathbf{n}) \leq b_j$, *for all* $\mathbf{n} \in \mathcal{N}$ *and* $j \in \mathcal{J}_{\mathbf{n}}$.

Based on the definition of the difference operator $\mathbf{D}_x$, Proposition 2.2 implies that $V(\mathbf{n}) - V(\mathbf{n} - \mathbf{e}^j) > 0$ for all j and $\mathbf{n} \in \mathcal{N}$. Thus, Proposition 2.2 states that the value function is increasing in n_j for all j and $\mathbf{n} \in \mathcal{N}$. Proposition 2.3 describes another property of the value function (2.1).

Proposition 2.3. *The optimality equation (2.1) has the following property:*

Property **P2:** $V^k(\mathbf{n}) - V^j(\mathbf{n}) \geq 0$ is non-decreasing in n_z, *for all* $z \neq j, k$ *and* $j, k, z \in \mathcal{J}_{\mathbf{n}}$.

Where we define $V^j(\mathbf{n})$ as the value function if system is at state $\mathbf{n}$ and the server serves a type-j customer. In other words, $V^j(\mathbf{n})$ satisfies the following equation:

$$\begin{aligned}\frac{g}{\Lambda} + V^j(\mathbf{n}) = &\frac{1}{\Lambda}\Big\{ \sum_{i=1}^{M} b_i\theta_i n_i + \sum_{i=1}^{M} \lambda_i V(\mathbf{n} + \mathbf{I}^i_{\{n_i<N_i\}}) + \sum_{i=1}^{M} n_i\theta_i V(\mathbf{n} - \mathbf{I}^i_{\{n_i>0\}}) \\ &+ \sum_{i=1}^{M}(N_i - n_i)\theta_i V(\mathbf{n}) + \mu_j V(\mathbf{n} - \mathbf{e}^j) + \sum_{i=1, i\neq j}^{M} \mu_i V(\mathbf{n})\Big\},\end{aligned}$$

Proposition 2.3 states that if at any given state $\mathbf{n}$, it is less costly to serve customer type j over type k (i.e., $V^k(\mathbf{n}) > V^j(\mathbf{n})$), then it would be less costly to serve customer type j over customer type k at state $(\mathbf{n} + \mathbf{e}^z)$, $\forall z \neq j, k$ and $j, k, z \in \mathcal{J}_{\mathbf{n}}$. In other words, if at state $\mathbf{n}$, serving a type-j customer has lower value than serving a type-k customer, then serving a type-j customer still has lower value than serving a type-k customer, when more type-z ($z \neq k, j$) customers arrive to the system.

Properties **P1** and **P2** lead us to the following result:

Theorem 2.4. *The optimal scheduling policy for MDP problem (2.1) is characterized as follows:*

(i) Idling is not optimal in a nonempty system, i.e., when $\mathbf{n} \neq \mathbf{0}$.

(ii) Server's optimal scheduling policy is static priority policy.

Theorem 2.4 states that under the optimal scheduling policy, the server gives a customer type static priority over another type, regardless of the number of all each type of customer in the system. This means that, as long as there is a customer of higher priority in the system, that customer type is served, regardless of how large the queue of low priority customers are.

Now that we showed the optimal scheduling policy is a static priority policy, the question is how should the customers of different types be prioritized? To answer this question, we define Conditions **C1** and **C2** for two different customer of types j and k as follows:

Condition **C1**: $b_j\theta_j\mu_j \geq b_k\theta_k\mu_k$

Condition **C2**: $b_j\theta_j\mu_j < b_k\theta_k\mu_k$, $b_j\theta_j\mu_j \geq b_k\theta_k\mu_k\left[1-\frac{\theta_k-\theta_j}{\Lambda}\right]$ *and* $\theta_j \leq \theta_k$.

As it was mentioned in the introduction, there are several factors to consider when minimizing the total expected abandonments cost: the cost of abandonment (b), customer abandonment rate (θ) and service rate (μ). The index $b\mu\theta$ captures the impact of these factors altogether. Condition **C1** demonstrates the case where the index $b\mu\theta$ of type-j customer is larger than that of type-k customer. Condition **C2** demonstrates the case where the index $b\mu\theta$ of type-j customer is smaller than that of type-k customer, but

sufficiently close to it. As the difference between abandonment rates becomes larger (i.e., θ_k become larger than θ_j), the range covered by Condition **C2** becomes smaller.

Proposition 2.5. *Suppose at state* $\mathbf{n}$*, Condition* **C1** *holds for* $j \neq k$*, then optimality equation (2.1) has the following property:*

Property **P3:** $\mu_j \mathbf{D}_j V(\mathbf{n}) \geq \mu_k \mathbf{D}_k V(\mathbf{n})$*, for all* $k \neq j$ *and* $k, j \in \mathcal{J}_{\mathbf{n}}$, $\mathbf{n} \in \mathcal{N}$.

Property **P3** is sufficient to show how to prioritize customer types:

Theorem 2.6. *Suppose at state* $\mathbf{n}$*,* $b_j \mu_j \geq b_k \mu_k$*, for* $k \neq j$ *and* $k, j \in \mathcal{J}_{\mathbf{n}}$*. If either Condition* **C1** *or* **C2** *holds for* $j \neq k$*, then type-*j *customer has a higher static priority over type-*k *customer.*

Theorem 2.6 demonstrates the direction of the optimal static priority policy. If the index $b\mu\theta$ of type-j customers is larger than that of type-k customers (i.e., Conditions **C1** holds) or if the index $b\mu\theta$ of type-j customers is smaller than that of type-k customers but sufficiently close (i.e., Conditions **C2** holds), the optimal scheduling policy is to give priority to type-j customers, when the index $b\mu$ of type-j customers is larger than that of type-k customers. Conditions **C1** and **C2** show that the difference between θ's of different customer types plays an important role in specifying server's optimal scheduling policy and index $b\mu$ index solely is not sufficient to determine the optimal scheduling policy.

In the following section, we introduce scheduling rules for special cases of system parameters.

2.5. Scheduling Policies for Special Cases

Condition **C1** leads to some scheduling rules for some special cases, summarized in the following corollaries. These scheduling policies for special cases make the implementation of the optimal policy easier in practice and give insights about the underlying parameters and their impact on dynamics of such a server scheduling problem.

Corollary 2.7 ($b\mu$–**rule**). *If customer types can be renumbered such that* $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_M$ *and* $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_M$*, and we have* $b_1\mu_1 \geq b_2\mu_2 \geq \cdots \geq b_M\mu_M$*, then type-*$j$ *customers have higher priority than type-*$(j+1)$ *customers for* $j = 1, 2, \cdots, M-1$.

As stated by $b\mu$–rule, all classes can be ranked according to $b\mu$ index if customer types are renumbered as suggested by Corollary 2.7. To illustrate, without loss of generality, consider a type-j customer and a type-k customer, where $b_j\mu_j \geq b_k\mu_k$. Notice that since $\theta_j \geq \theta_k$, Condition **C1** holds and, using Theorem 2.6, it is clear that the type-j customers have higher priority than type-k customers. The intuition behind this corollary is that, when type-j customers are less patient (i.e., $\theta_j \geq \theta_k$) and it takes less time to serve type-j customers compared to type-k customers (i.e., $\mu_j \geq \mu_k$), serving type-j customers leads to lower number of abandonments and thus lower total abandonment cost.

Corollary 2.8 ($b\mu\theta$–**rule**). *If customer types can be renumbered such that* $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_M$ *and* $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_M$*, but we have* $b_1\mu_1\theta_1 \geq b_2\mu_2\theta_2 \geq \cdots \geq b_M\mu_M\theta_M$*, then type-*$j$ *customers have higher priority than type-*$(j+1)$ *customers for* $j = 1, 2, \ldots, M-1$.

As stated by $b\mu\theta$–rule, all classes can be ranked according to $b\mu\theta$ index if customer types are renumbered as suggested by Corollary 2.8. Consider a type-j customer and

a type-k customer, where $b_j\mu_j\theta_j \geq b_k\mu_k\theta_k$. Since Condition **C1** holds, using Theorem 2.6, type-j customers have higher priority than type-k customers. The intuition behind this corollary is that, when type-j customers are less patient (i.e., $\theta_j \geq \theta_k$) and it takes more time to service type-j customers compared to type-k customers (i.e., $\mu_j \leq \mu_k$) but the overall impact of marginal cost, customer patience and service rate is higher for type-j customers (i.e., $b_j\mu_j\theta_j \geq b_k\mu_k\theta_k$), serving type-$j$ customers leads to lower total abandonment cost.

Corollary 2.9. *If $b_j\theta_j = b\theta$ for all $j = 1, \ldots, M$, then Shortest Expected Processing Time first rule (SEPT–rule) is optimal.*

Therefore, for example, if all customer types have the same marginal abandonment cost and abandonment rate (i.e., $b_j = b$ and $\theta_j = \theta$ for $j = 1, 2, \cdots, M$), then serving the customers with shortest expected process time is optimal.

Corollary 2.10. *If $\mu_j\theta_j = \mu\theta$ for all $j = 1, 2, \cdots, M$, then giving priority to customers with highest abandonment cost is optimal.*

Therefore, for example, if all customer types have the same service rate and abandonment rate (i.e., $\mu_j = \mu$ and $\theta_j = \theta$ for $j = 1, 2, \ldots, M$), then serving the customers with highest abandonment cost is optimal.

Corollary 2.11. *If $b\mu_j = b\mu$ for all $j = 1, 2, \ldots, M$, then giving priority to customers with highest abandonment rate (i.e., least patient customers) is optimal.*

Therefore, for example, if all customer types have the same service rate and marginal abandonment cost (i.e., $\mu_j = \mu$ and $b_j = b$ for $j = 1, 2, \ldots, M$), then serving the customers with highest abandonment rate is optimal.

Corollary 2.12. *If customer types can be renumbered such that* $b_1\mu_1 \geq b_2\mu_2 \geq \cdots \geq b_M\mu_M$ *and we have* $b_1\mu_1\theta_1 \geq b_2\mu_2\theta_2 \geq \cdots \geq b_M\mu_M\theta_M$*, then type-*$j$ *customers have higher priority than type-*$(j+1)$ *customers for* $j = 1, 2, \ldots, M-1$.

When both $b\mu$ and $b\mu\theta$ indexes are higher for type-j customers, two cases may happen. Either $\theta_j \geq \theta_k$ (i.e., type-j customers are less patient than type-k customers) and thus, giving priority to type-j customers results in lower number of abandonments and therefore lower abandonment cost. Or $\theta_j < \theta_k$ (i.e., type-j customers are more patient). In this case, higher $b\mu$ and $b\mu\theta$ indexes for type-j customers implies that the impact of θ is much less than the impact of b or μ (i.e., type-j customers have lower abandonment cost or type-j customers have higher service rate). Hence, giving priority to the type-j customer results in lower abandonment cost.

2.6. Numerical Study

So far, we have determined the optimal static priority policy for the server under conditions in Theorem 2.6 when the goal is to minimize the abandonment cost. However, when conditions in Theorem 2.6 do not hold, it is not clear which customer type has higher priority over the others. Theorem 3 addresses the important role of $b\mu$–index in determining the optimal scheduling policy. Specifically, it shows that Conditions **C1** and **C2** are not sufficient to determine the optimal scheduling policy and $b\mu$–index is the key driver of prioritizing customers. Furthermore, under a many-server fluid scaling and

overloaded conditions, $b\mu$–rule is conjectured to be the optimal server scheduling policy (Atar et al. (2010)). When the overloaded conditions are not met, however, $b\mu$–rule may or may not work well. Nevertheless, $b\mu$–rule can certainly be considered as a candidate server scheduling policy, when conditions in Theorem 2.6 do not hold.

In addition, because of the role θ plays in prioritizing customers in our setting, we also consider $b\mu\theta$–index as a candidate server scheduling policy. This $b\mu\theta$–index is a key drivers in both Conditions **C1** and **C2**. Thus, to evaluate the performance of $b\mu$–rule or $b\mu\theta$–rule as heuristic policies, when conditions in Theorem 2.6 do not hold, we study the answer to the following question:

Question 1: How well $b\mu$–rule or $b\mu\theta$–rule perform when conditions in Theorem 2.6 do not hold?

On the other hand, typically, in infomercial call centers for example, calls are answered in the order they are received. Hence, it is important to compare the performance of the proposed scheduling policy with the typical first-come-first-served policy. Thus, we also need to investigate the answer to the following question:

Question 2: How does FCFS policy performs compared to the optimal scheduling policy?

6.1 Analysis of the Uncovered Region

To find an answer to Question 1, we give a closer examination of cases for which neither conditions in Theorem 2.6 holds. We call the region, that is formed by different values of

b, μ and θ for a customer of type j and type k for which neither conditions in Theorem 2.6 holds, *Uncovered Region*. Using a little bit of algebra, the Uncovered Region is found to be the region that satisfies inequalities in (2.2).

(2.2)

$$\textbf{Uncovered Region:} \quad b_k\theta_k\mu_k\left[1-\frac{\theta_k-\theta_j}{\Lambda}\right] > b_j\theta_j\mu_j, \quad b_j\mu_j > b_k\mu_k \quad \text{and} \quad \theta_j \leq \theta_k.$$

First, we study how large the Uncovered Region is. The answer to this question emphasizes on how broad are the conditions in Theorem 2.6. To estimate the size of the Uncovered Region, we use Monte Carlo simulation to bombard the solution region (i.e., region of all possible combination of b, λ, μ and θ for a customer of type j and type k) with random values for system parameters. This is done by uniformly generating different combination of b, λ, μ and θ for a customer of type j and type k. All values generated within the interval [0,100] and total of 10^8 cases were generated.

The number of cases out of 10^8 gives a proxy for the size of the uncovered region. Our procedure reveals that the ratio of the Uncovered Region (i.e., the number of generated cases which were in Uncovered Region) to the Solution Region (i.e., the number of all generated values whether or not conditions in Theorem 2.6 hold) is less than 10%. This shows that the likelihood that the parameters of a system of interest is inside the Uncovered Region (i.e., outside conditions in Theorem 2.6) is relatively small. Therefore, $b\mu$–rule is applicable as the optimal scheduling policy more than 90% of the times. However, since there is still 10% of cases for which neither conditions in Theorem 2.6 holds, it is still valuable to find a rule that captures the most benefit of the optimal policy in the Uncovered Region.

Now, we focus on a two-class queuing system and design a numerical study to answer Question 1. Specifically, we want to know what would be the optimal static priority policy if a type-j customer is more patient than a type-k customer (i.e., $\theta_j \leq \theta_k$) but it takes longer to serve the type-k customer (i.e., $\mu_j \geq \mu_k$), such that neither conditions in Theorem 2.6 holds. We use parameters in Table 2.1 to generate different problem instances with low utilization (i.e., $\rho = 0.5$), medium utilization (i.e., $\rho = 0.75$) and high utilization (i.e., $\rho = 0.95$). To generate a case for our numerical study, first, we choose pairs of (μ_1, μ_2), (θ_1, θ_2) and $(b_1,\ b_2)$ from Table 2.1. We then calculate λ's to obtain the set of traffic density $\rho = (0.5, 0.75.0.95)$, where $\rho = \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}$. To eliminate the impact of finite queue limit N_i, N_1 and N_2 are set to 120 in our numerical study to make sure that the probability of queues reaching their limit N_1 and N_2 is less than 1%, similar to infomercial call centers setting where buffer size is large and balking rate is small. Since we are interested in the performance of the proposed rules when neither conditions in Theorem 2.6 holds (i.e., Uncovered Region), we only keep instances that are inside the Uncovered Region.

Table 2.1. Parameters of the Experiment.

(μ_1, μ_2)	(θ_1, θ_2)	(b_1, b_2)
(10,1),(1,10)	(1,0.1), (0.1,1)	(5, 10)
(2,1),(1,2)	(0.1,0.2),(0.2,0.1)	(10, 5)
(1, 1)	(1,1), (0.1,0.1)	(1, 1)

We evaluate the performance of $b\mu$–rule and $b\mu\theta$–rule as an alternative to the optimal static priority policy. Recall that, $b\mu$–rule gives priority to customer type with the highest $b\mu$ index and $b\mu\theta$–rule gives priority to customer type with the highest $b\mu\theta$ index. To

evaluate the performance of policy π, we define Performance Loss (PL) of policy π as follows:

$$\mathrm{PL}(\pi) = \frac{B_\pi - B^*}{B^*}$$

$\mathrm{PL}(\pi)$ represents how much (in percentage) the total average abandonment cost under policy $\pi \in \{b\mu - rule, b\mu\theta - rule\}$ is larger than that under the optimal static priority policy. We use value iteration algorithm to find B^*, the optimal average cost per unit time. The stopping criteria is set at $\epsilon = 0.001$ (i.e., the algorithm stops when the difference between average abandonment cost in two consecutive iterations is less than 0.1%). Table 2.2 summarize the performance of $b\mu$–rule and $b\mu\theta$–rule. Notice that we only consider cases in the Uncovered Region. This includes 48 cases out of 270 previously designed cases.

Table 2.2. Summary of performance of $b\mu\theta$–rule and $b\mu\theta$–rule

Item	$b\mu\theta$–rule			$b\mu$–rule		
	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.95$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.95$
Average PL	0.0%	1.0%	4.9%	14.1%	15.8%	13.7%
Worst Case	0.0%	4.0%	27%	37%	42%	46%
% of cases has PL $< 5\%$	100%	100%	75%	38%	38%	38%

As shown in the Table 2.2, $b\mu\theta$–rule results in small PL compared to the optimal scheduling policy in systems with low utilization (i.e., $\rho = 0.5$) or medium utilization (i.e., $\rho = 0.75$). The average and maximum PL in these cases are (0.0%, 0.0%) and (1.0%, 4.0%), respectively. When the utilization is high (i.e., $\rho = 0.95$), the performance of $b\mu\theta$–rule is on average better than $b\mu$–rule and $b\mu$–rule does not perform well, on the contrary to what Theorem 3 and systems with overloaded condition predicted. The intuition is that, in a system with low or medium utilization, the number of customers in

the system rarely gets large. In this case, the server can control the number of abandonments by considering θ in determining the optimal scheduling policy. This can be done by using the $b\mu\theta$–rule, which captures the effect of θ directly and this is why $b\mu\theta$–rule performs very well. However, for a highly utilized system, it is more likely to have a large number of customers in the system. In this case, the role of θ becomes less important in optimal server scheduling, even though it is still important. In such a system, since the number of waiting customer is expected to be high, more customers are likely to abandon. Therefore, to minimize the number of abandoned customers, it is more important to serve customers with shorter processing time than customers with higher abandonment rate, since some customers will abandon regardless. This is why $b\mu\theta$–rule does not perform well (i.e., results in 27% PL), when $\rho = 0.95$. Nonetheless, $b\mu\theta$–rule is still on average as good as the optimal scheduling policy. It results in PL of less than 5% for 75% of the cases.

6.2 Comparison with First-Come First-Served Policy

In this section, we discuss the answer to Question 2. First-Come First-Served (FCFS) policy is the most common service policy used in practice because of the fairness. In call centers for instances, calls are typically answered in the order they are received Gans et al. (2003). Comparing the performance of the server's optimal scheduling policy with that of the FCFS policy reveals how much the optimal scheduling policy can save the long-run abandonment cost per unit time. We use parameters in Table 2.1 to generate a total of 270 cases to compare the long-run abandonment cost per unit time under optimal scheduling policy and under FCFS policy. To compute the long-run abandonment cost per unit time when FCFS policy is used, we develop a simulation model. To evaluate the

long-run abandonment cost saving per unit time when server uses the optimal scheduling policy instead of FCFS policy, we define $\mathrm{CS}_{\mathrm{FCFS}}$ as follows:

$$\mathrm{CS}_{\mathrm{FCFS}} = \frac{B_{\mathrm{FCFS}} - B^*}{B^*}.$$

Where B^* is the optimal average cost per unit time and B_{FCFS} is the average cost per unit time when server uses FCFS policy. We find that, on average, we save 80% on long-run abandonment cost per unit time when server uses the optimal scheduling policy instead of FCFS policy. The highest cost saving in our set of experiment is 98%.

When a customer type with shorter service time (i.e., higher μ) has relatively shorter time between arrivals (i.e., higher λ) and low patience tolerance (i.e., high θ), system can save considerably on the abandonment cost using optimal scheduling policy instead of FCFS policy. In this case, giving priority to the customer type with shorter service time and low patience tolerance become more critical to minimize the abandonment cost. When serving customers with shorter service time and low patience tolerance is delayed by serving customer types with longer service times and high patience tolerance, more customers of the former type leave the system and due to their higher arrival rate, system incurred larger abandonment cost. Note that the cost saving will be even more when the customer type with shorter service time, shorter time between arrivals, and low patience tolerance, also has higher cost of abandonment (i.e., higher b). In this case, the system under FCFS policy incurs even much higher abandonment cost than before, since delaying serving the customer type with higher b is more costly.

2.7. Conclusion

In this paper, we studied the optimal server scheduling policy in a multi-class queuing system with a single server when minimizing the abandonment cost. We showed that optimal scheduling policy is a static priority policy. We derived conditions under which we can determine the optimal scheduling policy among customers of different types. The optimal scheduling policy gives priority to the customer type with higher service rate (μ) and higher abandonment cost (b), i.e., higher index $b\mu$, only if either condition **C1** or **C2** holds. We also numerically observed that when these conditions do not hold $b\mu\theta$–rule performs well. In our numerical analysis, we also observed that using the optimal scheduling policy results in a significant cost saving compared to the FCFS service policy, which is commonly used in call centers.

There are still several possible directions for further research. One possible direction for future research is extending the current model to multi-server multi-class system. In this case, the optimal scheduling policy for each server would also be dependent on other servers' attributes and scheduling policies. Another possible direction for future research is server scheduling policy for systems where customers' abandonment cost and holding cost need to be considered simultaneously, e.g., other types of call centers where holding cost is not negligible or the delivery of health care where holding cost of patients is considerable. The trade-off between these two costs may make the optimal scheduling policy more complicated and different.

CHAPTER 3

Engineering the Delay Announcement to Improve Patient Satisfaction

3.1. Introduction

In 2010, Centers for Medicare and Medicaid Services (CMS) initiated the Hospital Value-Based Purchasing (VBP) Program to adjust payments to hospitals based on the quality of care they deliver (CMSb 2017). The VBP program includes clinical quality measures (e.g., 30-Day Mortality Rate) as well as *patient satisfaction* with care measure, i.e., Hospital Consumer Assessment of Health care Providers and Systems (HCAHPS) survey. The clinical measures account for 75 percent of a hospital's VBP score and the HCAHPS survey accounts for 25 percent. The total score is used to determine the amount of incentive payment each hospital receives (AHA 2013). In addition, HCAHPS score are shared with the public on Medicare website and impact hospitals' reputation and standing in the community they serve. Hence, *patient satisfaction* with care has become a financial priority for hospitals.

Financial impact of patient satisfaction. Accenture[1] analyzed hospital profit margin data reported to the CMS and survey results from HCAHPS to examine the relationship between patient satisfaction and hospital financial performance. They reported

[1] https://www.accenture.com/t20180111T085228Z__w__/us-en/_acnmedia/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Industries_17/Accenture-Happy-Patients-Healthy-Margins.pdf

that profit margin increase per 10% increase in HCAHPS score is from 0.4% for rural hospitals to 3.4% for urban hospitals. This means, for example, a hospital system earning $2B in revenue would have to cut 460 jobs (assuming a salary of $100K) to achieve the 2.3% profit margin that improving the consumer satisfaction might bring through revenue growth. Considering the relationship between patient satisfaction and hospital financial performance, hospital leaders are looking for strategies to improve patient satisfaction and boost their HCAHPS scores.

How to improve patient satisfaction? *Wait Time Before Seen By Doctor* has a significant impact on patient satisfaction in EDs (Boudreaux and O'Hea 2004). Recent internal patient satisfaction report[2] at the ED under study shows that the survey question asking about patients' *Wait Time Before Seen By Doctor* on Press Ganey[3] survey accounts for more than 46% increase in patients' *Likelihood-To-Recommend* (i.e., a metric for patient satisfaction assessment). Therefore, any improvement in patients' wait-time before seeing the doctor would have a considerable impact on patient satisfaction. More recently, patients' wait-times become even more important since hospitals across the country struggle with overcrowded EDs and long wait-times for patients who seek emergency care (GAO 2011). Increase in patient satisfaction consequently boosts hospitals HCAHPS scores, which in turn leads to an increase in the revenue and Medicare reimbursements.

To improve patient satisfaction, one can reduce the wait-times. Often, the only way to reduce wait-times without sacrificing quality is to add beds or staff, which is expensive.[4]

[2]Patient Satisfaction report FY2012-FY2014- N=4,668 paper surveys

[3]Press Ganey is the largest HCAHPS administrator in the country, partnering with more than 8,000 clients.

[4]In the United States, it is estimated that it costs approximately $1,000,000 to build a hospital bed, and $600,000 to $800,000 to staff that same bed Salway et al. (2017).

One effective, inexpensive, way to improve patient satisfaction is providing delay information, so called *delay announcement*. Case studies from service systems such as call centers (Antonides et al. 2002), restaurants, supermarkets (Tom and Lucey 1995, 1997) or airlines (Forbes et al. 2017), show that delay announcement can improve customers' waiting experience. In this study, we use delay announcement to improve patient satisfaction.

Satisfaction vs. Abandonments. The delay announcement literature chiefly focuses on the effect of announcing delay information on customers' behavioral intentions during a service encounter, in particular the customer abandonment from system caused by the incurred holding (waiting) cost (Armony et al. 2009, Yu et al. 2016, Akşin et al. 2016). The effect of delay announcement on customers' self-reported satisfaction, especially in the context of EDs, has not been carefully studied in the operations management literature. Yet as discussed above, self-reported satisfaction is important for ED managers considering the direct impact of patient satisfaction on hospital's reimbursement and revenue. Hence, we study the impact of delay announcement on patient satisfaction in EDs.

EDs vs. Call centers. Delay announcement have been chiefly studied in the context of call centers. However, there are key operational differences between EDs and call centers, especially when it comes to delay announcement. First, *Queues and arrivals are observables* in EDs, while queues are unobservable in call centers. Therefore, patients may form their own perception about the wait-times and progress in line by observing the flow. Second, EDs work as a *priority system*, while call centers typically work as a first-come-first-served (FCFS) system. Even if a call center does work as a priority system, due to the unobservability of the queue, customers will not necessarily realize

that system does not work FCFS. However, in EDs, patients occasionally realize that the system does not work FCFS by seeing other patients who come after them receive service before them. Third, patients are *different in the severity of their condition* and they may have different sensitivity to the wait-times and announced wait-times. Bolandifar et al. (2014), observed that patients with sever conditions are less sensitive to wait-times. Announcing delay to patients with sever conditions with their estimated wait-time may make them more sensitive to their wait-time and have some negative effects on their satisfaction. These differences make EDs operationally different than call centers, where delay announcement's impact on satisfaction may not be the same.

Delay announcement in EDs. Providing patients with an estimate of their wait-time is challenging. Uncertainty of patients arrival pattern and treatment times along with the complexity of process flows in EDs make accurately estimating wait-times difficult. On the other hand, patients may take the announced wait-time as their reference point for their wait-time and become dissatisfied if they have to wait beyond the announced wait-time. Hence, considering the potential inaccuracy of estimated wait-times, it is important to carefully study what wait-time to announce to maximize patient satisfaction. In the literature of delay announcement, the mean (or median) of the estimated wait-time distribution is usually selected to be announced (e.g., call centers (Armony et al. 2009, Jouini et al. 2011, Yu et al. 2017) and health care (Mowen et al. 1993, Shah et al. 2015)). However, by announcing the average wait-time, we underestimate the wait-time of a considerable number of customers. Waiting beyond the announced wait-time is a time loss for a customer and this may make her dissatisfied. If an overestimate of wait-time is announced instead (i.e., announcing a larger delay than what is estimated), the customer

experiences a time gain. However, it is still not clear how much larger than estimated wait-time the announced wait-time needs to be, considering the potential negative consequences of announcing large wait-times, which we discuss below. The question of what wait-time to announce to maximize satisfaction, especially EDs, has not been studied in the literature. Hence, when designing a delay announcement process in EDs, the *first* question (**Q1**) that ED managers face is whether they should announce the mean (or median) of wait-times in EDs similar to common practice in some service settings such as call centers, or announce an overestimate of wait-times instead?

Wait-time Gap, Actual wait-time and Satisfaction. Patients evaluate their waiting experience based on the difference between expected wait-time and actual wait-time (Maister 1984), which we call *wait-time gap*. If actual wait-time exceeds the expected wait-time, the patient experiences a time loss and may become dissatisfied. If actual wait-time does not exceed the expected wait-time, the patient experiences a time gain. Patients derive utility from gains and losses, as stated by Prospect Theory (Kahneman and Tversky 1979). With delay announcement, we tend to set an expectation (i.e., a reference point) for patients' wait-time. On the other hand, even though people may derive utility from gains and losses, the actual wait-time may play an important role in their evaluations as well (Barberis 2013). The *second* question (**Q2**) is how actual wait-time and wait-time gap contributes to patient satisfaction, in presence of delay announcement?

Announcing long wait-times. Since patients usually experience long wait-time in urban EDs (e.g., 5-6 hours), the *third* important question (**Q3**) that ED managers face is whether they should tell patients that their wait-time to see the doctor would be long? Announcing long delays may have a negative initial impact on patient wait-time

satisfaction (Carmon and Kahneman 1996). On the other hand, not knowing the wait-times may make patients anxious and dissatisfied (Hui and Tse 1996). Announcing long wait-times may also turn patients away, which is a health risk for patients and revenue loss for hospitals. Thus, the effect of announcing long wait-times on patient satisfaction is not clear and need to be carefully investigated.

Engineering the delay announcement. We define *engineering the delay announcement* as to study what wait-time to announce to maximize the ED's average patient wait-time satisfaction. To engineer the delay announcement and to answer questions **Q1-Q3**, we conduct a field experiment in a urban ED in which all patients who get triaged and have to wait to see a doctor are provided with their estimated personalized wait-times with no updates, computed using a prediction model developed based on the ED's historical data. In this field experiment, we study the effect of different ranges of wait-time gaps on patient wait-time satisfaction. To create different ranges of wait-time gaps, we cannot simply add some specific minutes to the estimated wait-times, since the wait-times are difficult to estimate accurately and therefore unavailable. One way to create different ranges of wait-time gaps is to announce different prediction upper-bounds on patients' wait-time using the developed prediction model. By announcing different prediction upper-bounds on wait-times, we expect to create different ranges of overestimations and underestimations of wait-time. With this experiment design, we, first, study how patients form their utility in the presence of delay announcement, and how they evaluate their wait-time satisfaction retrospectively. Considering the potential positive impact of delay announcement in reducing patients uncertainty and the possible negative effects of delay announcement when announcing long delays that may cause an initial negative impact, we explore how

these positives and negatives together impact patient wait-time satisfaction. We then, discuss how to engineer the delay announcement to maximize the total average patient wait-time satisfaction.

This chapter contributes to the literature on patient satisfaction and delay announcement in EDs. The first goal of this chapter is to demonstrate the impact of time gains and time losses, the actual wait-time and the announced wait-time on patient satisfaction with their wait-time in an ED setting. EDs perform operationally different from other service settings and these differences lead us to different implications of current available finding in the literature. In particular, we show that while patients drive utility from time gains and time losses relative to the announced wait-time as well as their actual wait-time, the time gains effect on wait-time satisfaction is not always positive. With a better understanding of patients' evaluation process, we then turn our attention to the second goal of this chapter, which is developing a model for patient wait-time satisfaction based on our observations. We also explore how we can maximize the positive impact of delay announcement on patient satisfaction. As a by-product of this research, we also study factors that affect wait-times and develop an institution specific and accurate wait-time predictor application.

Our results enable ED managers to get a better understanding of the relationship among time perceptions, delay announcement and patient satisfaction. This helps managers design a delay announcement process that improves patient satisfaction in their ED. We formulate this relationship as a function of wait-time gap and actual wait-time. Our field experiment delivers a number of key findings. First, considering the positive and negative impacts of delay announcement in an ED, where patients observe the queue

and arrivals and typically have to wait for long hours, we find that delay announcement can improve patients' wait-time satisfaction. While patients' overall satisfaction is also increased as a result of delay announcement in the ED under study, the increase in the number of patients who left after learning their long announced wait-times was not found to be significant in our study. Second, we find that announcing an overestimate of the wait-time improve patient satisfaction more than the current practice of announcing the mean (or median) of the wait-time. This is because patients are loss averse with respect to their wait-time and by announcing delays, we can set their expectations and make them satisfied by exceeding it. Finally, we find evidence that for sufficiently large positive wait-time gaps, the average wait-time satisfaction is not increasing in the gap, which may be because of the negative impact of announcing long wait-times on patient satisfaction.

This chapter proceeds as follows. In Section 2, we review related literature and in Section 3, we develop our hypotheses. The experiment design is presented in Section 4. We introduce our empirical models and our comprehensive statistical analysis results in Section 5. In Section 6, we explore how to engineer the delay announcement to maximize ED's average wait-time satisfaction. We explore the robustness of our findings in Section 7. We conclude the chapter and discuss our findings in Section 8.

3.2. Literature Review

In this study, we explore the impact of delay announcement on patient wait-time satisfaction in EDs. We, first, review literature related to customer wait-time satisfaction in service operations and the literature on the impact of delay announcement. We then, develop the hypotheses considering the literature review.

3.2.1. Customers Wait-time Satisfaction

For customers waiting in a line, the concept of disconfirmation (i.e., gap) between actual wait and expectations and how this disconfirmation affects both cognitive and affective components with respect to the appraisal of the wait has been discussed in the literature (e.g., Hornik (1984), Pruyn and Smidts (1993), Taylor and Fullerton (1999)). Research on the customers wait-time experience focuses on managing customers perception of wait-time by occupying periods of idle time (Carmon et al. 1995), increasing the feeling of progress (Soman and Shi 2003), managing anxiety and uncertainty (Osuna 1985), setting accurate expectations and improving perceptions of fairness (Maister 1984), managing sequence and duration effects (Chase and Dasu 2001), shaping memories of the experience (Norman 2009), and operational transparency (Buell and Norton 2011). In ED setting, there are several studies that propose strategies for improving the experience of waiting patients. These strategies include expressing empathy for patients and making them feel occupied while waiting (Cohen et al. 2013) and keeping patients informed about the treatment process (Krishel and Baraff 1993, White et al. 2005) and the time (Göransson and von Rosen 2010, Johnson et al. 2012, Shah et al. 2015). Even though, the importance of providing delay information is recognized in these papers, it is unclear how patients form their satisfaction with respect to their wait-time, especially when delay information is provided. We study the effect of delay announcement on patient satisfaction.

3.2.2. Delay Announcement

Delay information (i.e., estimated wait-time in queue or queue length) impacts service evaluations (Hassin 1986) and the system throughput (Hassin 2007). Customers use the

delay information to estimate the distribution of delay and then to determine their expected waiting costs. Comparing the costs to the reward they anticipate from receiving service, they decide whether to stay or leave. In call center setting, Whitt (1999) analytically shows that if the service provider does communicate anticipated delays, the customers are more likely to balk when all servers are busy (leave immediately upon arrival) than renege (leave after waiting for some time) and this reduces the system's average waiting time. Guo and Zipkin (2007) extend Whitt (1999) and explore how delay announcements with different levels of precision impact customers' balking behavior. They show that different levels of information lead to different delay distributions in the expected waiting cost calculations, and exact delay information may improve or hurt system performance. Hu et al. (2017) investigate how information heterogeneity among customers impacts the throughput and social welfare. They discuss how delay information helps system capacity to be more efficiently matched with customer demand, while selfish joining behavior of informed customers may overload the system. Armony et al. (2009) extend Hu's paper by considering customer abandonment in their model. Delay announcement is not always credible or treated as such by customers. Allon et al. (2011) address this concern by considering a model in which both the firm and the customers act strategically. Yu et al. (2015) extend Allon et al. (2011) by incorporating customer heterogeneity and allowing the firm to prioritize its customers. All the papers above are analytical in nature and their results are not examined in real systems.

There are a few papers that empirically explore the impact of delay announcement on system performance. Hui and Tse (1996) conduct an experiment to study the impact of waiting duration information and queuing information on customers' service evaluations

(i.e., the extent of customer's preference for the provided service). They show that acceptability of the wait and affective response[5] to the wait have a significant mediating effect on the relationship between waiting information and service evaluation. In call center setting, Yu et al. (2016), Akşin et al. (2016), and Yu et al. (2017) empirically study the impact of delay announcement on customers' beliefs about their wait-time and as a consequence customer abandonment behavior. Akşin et al. (2016) assumes that delay announcements only impact customers' beliefs about their wait-time but do not directly impact customers per unit waiting cost. Yu et al. (2016) relax this assumption by showing that the delay announcements also directly impact customers' per unit waiting cost. They assume customers' per unit waiting cost is constant over time. Relaxing this assumption, Yu et al. (2017) explore the reference effect of delay announcements in a field experiment and allow the customers' per unit waiting cost before the reference point to be different than that after the reference point. Since the value of time is context-dependent (Schmitt and Leclerc 2002), insights derived from these papers, even though valuable, are limited to the context of their study and the performance measure of interest. It is still important to explore the impact of delay announcement on patient satisfaction in EDs.

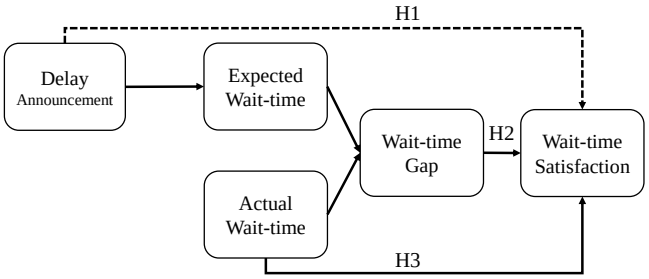
There are also several papers in ED setting that study the effects of publishing delay information on patients hospital selection (Xie and Youash 2011) and coordination within hospital networks (Dong et al. 2015). However, the focus of our study is on the effect of delay announcement on patients wait-time satisfaction and how to engineer the announced wait-time to maximize the patient satisfaction.

[5]The general psychological state of an individual, including but not limited to emotions and mood, within a given situation.

3.3. Framework and Hypothesis Development

In this study, we use *Delay Announcement*, providing patients with their estimated wait-time, to improve patient satisfaction with their wait-time. As mentioned in introduction, patient satisfaction is important for hospitals due to its direct impact on hospital's reimbursement and revenue. Figure 3.1 shows the conceptual model of delay announcement impact on wait-time satisfaction.

Figure 3.1. The Conceptual Model of Delay Announcement Impact on Wait-time Satisfaction



As shown in Figure 3.1, patients evaluate their wait-time satisfaction based on (1) the difference between their expected wait-time and actual wait-time (i.e., wait-time gap) and (2) the actual wait-time. With delay announcement, we tend to set an expectation for patients' wait-time (i.e., provide patients with a reference point on their wait-time). In the following subsections, we further explain our conceptual framework shown in Figure 3.1 and discuss the answer to questions **Q1-Q3** raised in the introduction. First, we explore the treatment effect of delay announcement on wait-time satisfaction.

Impact of Delay Announcement on Wait-time Satisfaction. In a field experiment, Shah et al. (2015) find evidence that providing expected wait-time information increases patients' *overall satisfaction*, but they observe no significant improvement with patients' *wait-time satisfaction*. No explanation is provided by Shah et al. (2015) about why no

improvement is observed with patient wait-time satisfaction in their study. One would, however, expect that wait-time information improves patient satisfaction: Uncertainty about the wait-times may cause anxiety and dissatisfaction (Hui and Tse 1996). In a crowded ED, information about the estimated wait-time can reduce the uncertainty of an anxious patient and make her less dissatisfied. The reasons behind Shah et al. (2015) not observing any improvement in patient wait-time satisfaction may be as follows. First, they used a table of *expected* wait-times for different acuity levels and shifts of day to communicate the wait-times. Using expected wait-times, may increase the number of patients who wait longer than what it is announced to them. Having to waiting longer than the announced wait-time (which happens 50% of the time when the wait-time distribution is symmetric around the *expected* wait time) may cause disutility. Second, patients are asked to fill out an in-house patient satisfaction questionnaire at the ED discharge desk, i.e., after their treatment process is over. This may cause patients' evaluations to be influenced by the quality of the treatment process rather than patient wait-time satisfaction. In our study, we survey patient right after they are seen by the doctor to exclude the effect of quality of treatment.

Our first hypothesis test the overall impact of delay announcements on patient wait-time satisfaction, see Figure 3.1. While Shah et al. (2015) find that delay announcement has no impact on wait-time satisfaction, we hypothesize that delay announcement can also have a positive impact on patient wait-time satisfaction.

Hypothesis 1A. Delay announcement increases wait-time satisfaction.

Hypothesis 1B. Delay announcement has no impact on wait-time satisfaction.

If Hypothesis 1A is true, ED managers are encouraged to use delay announcement to effectively improve patient wait-time satisfaction. Below, we decompose the determinants of wait-time satisfaction into two components; the wait-time gap (Hypothesis 2) and the actual wait-time (Hypothesis 3).

3.3.1. Impact of Wait-time Gap on Wait-time Satisfaction.

In a goal-oriented activity such as waiting-in-line, people's evaluation of the waiting experience is dominated by the end of their experience (Fredrickson and Kahneman 1993, Carmon and Kahneman 1996), so called *end effect*. Experiencing a time gain or loss once being seen by a doctor relative to the expected wait-time can be considered an *end effect* moment in the patient's waiting experience. Therefore, we expect that the *wait-time gap*, i.e., the difference between the expected and the actual wait time, to be a strong predictor of wait-time satisfaction. In this section, we explore the relationship between wait-time gap and wait-time satisfaction, as shown in Figure 3.1. Naturally, a negative (positive) wait-time gap reduces (increases) the satisfaction, as the patient perceives she had to wait more (less) than expected. From the perspective of Prospect Theory (Kahneman and Tversky 1979), the expected wait-time is a *reference point*. First, we hypothesize that the announced wait time acts as such reference point and that the satisfaction increases in the wait-time gap. In other words, compared with the satisfaction level of an actual wait-time that is the same as the announced wait-time (i.e., a zero wait-time gap), the increase (decrease) in satisfaction for a positive (negative) wait-time gap of 30 minutes is higher (lower) than the increase (decrease) in satisfaction for a positive (negative) wait-time gap of 15 minutes. Second, according to Prospect Theory, people evaluate

losses more than the same amount of gains, which is referred to as *loss aversion*. Thus, compared with the satisfaction for a zero wait-time gap, the decrease in satisfaction of a negative wait-time gap of 30 minutes is higher than the increase in satisfaction of a positive wait-time gap of 30 minutes. As the preferences should hold for all levels of the wait-time gaps (positive or negative), we state the following hypotheses to check the properties of patients' wait-time satisfaction to be in accordance with Prospect Theory.

Hypothesis 2A. Wait-time satisfaction increases as wait-time gain increases, for all levels of positive and negative wait-time gaps.

Hypothesis 2B. The average wait-time satisfaction increases in wait-time gap is smaller when the wait-time gap is positive compared to when the wait-time gap is negative (i.e., loss aversion).

Hypothesis 2A help us understand the effect of wait-time gap on wait-time satisfaction as the wait-time gap size changes. In other words, we study if wait-time satisfaction is an increasing function of wait-time gap both in the regions of gains and losses. Hypothesis 2B explores if delay announcement provide patients with a reference point for their wait-time, whether patients are loss averse with respect to the reference point.

While there is an emerging interest in customers' reference-dependent and loss aversion behavior in the operations management literature, most of the works related to consumers' temporal decisions are analytical in nature, focusing on investigating the managerial implications of customers' loss aversion in time (Yang et al. 2013). There are several studies in the literature that explore customers' loss aversion behavior in the temporal domain using field data (Abdellaoui and Kemel 2013, Crawford and Meng 2011, Yu et al. 2017).

Crawford and Meng (2011) show that New York City taxi drivers' labor supply decisions are loss averse relative to both the targeted hour and income. In a series of lab experiment, Abdellaoui and Kemel (2013) show that loss aversion exists for both time and money and the magnitude of loss aversion is significantly lower for time. Yu et al. (2017) explore customers' loss aversion behavior in the temporal domain in a field experiment approach in a call center setting. They found that while delay information does not alter the nature that customers are loss averse, it does seem to impact the reference points customers use when the announcements are accurate. To the best of our knowledge, there has been no study to explore patients' loss aversion behavior in the temporal domain in an ED setting. Since EDs works operationally different from the other previously studied settings, we explore patients' loss aversion behavior when there is delay announcement in Hypothesis 2B.

3.3.2. Impact of Actual Wait-time on Wait-time Satisfaction.

In this section, we explore the relationship between actual wait-time and wait-time satisfaction, as shown in Figure 3.1. If patients A and B experience the same amount of positive wait-time gap of 2 hours, they should have the same evaluation of their service experience. This is because it is gain or loss relative to a reference point (i.e., can be the announced wait-time) that drives a customer's utility. What if patient A has actually waited only 1 hour and patient B's actual wait-time was 3 hours. This difference in the wait-time that patient A and patient B actually experienced should lead them to evaluate their wait-time satisfaction differently. Hence, it is necessary to explore the contribution

of patients' actual wait-time to their evaluation of their wait-time satisfaction, when we control for the wait-time gap.

Some studies in behavioral economics literature state that the duration of an experience could have an effect on the experience (Schreiber and Kahneman 2000). However, other studies (Fredrickson and Kahneman 1993, Varey and Kahneman 1992) have suggested that customer evaluations are uncorrelated with the duration of a experience. Therefore, further research is needed to test the implications of experience duration in different service settings. In our setting, the duration of the waiting-time experience is the actual wait-time of the patients, which may or may not have a direct impact on her wait-time satisfaction. There are studies in patient satisfaction literature that introduce actual wait-time as one of the most significant predictors of patients' dissatisfaction in EDs (Boudreaux and O'Hea 2004, Welch 2009). Recall from Hypothesis 2A that the actual duration already plays a role in the wait-time satisfaction via the wait-time gap. We hypothesize that the actual wait-time has a direct, negative effect on the wait-time satisfaction, controlling for the wait-time gap.

Hypothesis 3. Keeping the wait-time gap fixed, the actual wait-time has a negative impact on patient satisfaction.

Hypothesis 3 highlights the extend to which actual wait-time contributes to patients' evaluation process. This hypothesis is basically testing the relationship between actual wait-time and wait-time satisfaction as it is shown in Figure 3.1.

3.4. Experiment Design

In this field experiment, we study the effects of delay announcement on patient wait-time satisfaction in a urban ED. In our study, we announce the $(1-\alpha)\%$ prediction upper-bound on the wait-time to see a doctor, which we call $(1-\alpha)\%$ overestimation of wait-time, computed using a regression model developed based on historical data and $(1-\alpha)$ takes the values 0.5, 0.7 and 0.9. The ED under study serves up to 312 patients per day for 8 shifts daily in two tracks: main track and fast track. Our study focuses on the patients who are treated in the main track, which accounts for approximately 70% of patients. Patients assigned to the main track usually wait longer than those assigned to the fast track. Patients under influence, patient with mental problems and patients assigned to bed with no wait in the waiting room are excluded from our study.

3.4.1. Participants.

We surveyed 373 patients (66% female; $\text{Mean}_{age} = 59$, $\text{Standard Deviation}_{age} = 17$) who visited the ED during the period of analysis. All of these patients had to wait in the waiting room to see a doctor. No inducements were offered to patients in exchange for their participation.

3.4.2. Design and Procedure.

When a patient arrives at ED, first, she will be checked-in and wait in a waiting room to be called for triage (typically within 15 minutes). At triage, the triage nurse evaluates the patient's medical conditions and determines her severity of condition according to a 5-level severity index, called Emergency Severity Index (ESI). The patient will be asked

to wait in the waiting room for a bed to get empty. See the process flow in Figure 3.2. We redesigned this process by asking the triage nurses to guide all patient's to a desk

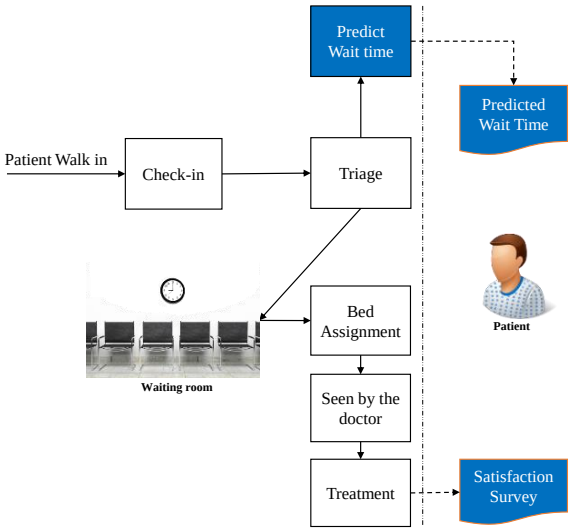


Figure 3.2. Process Flow Chart of Patients under study at Northwestern Medicine ED.

located in the waiting area, right after the triage process is over. A research assistant[6] uses the patient's characteristics and visit information to estimate the patient's wait-time and informs the patient of her personalized wait-time according to the planned intervention. With this manipulation, we are able to investigate the directional effects of delay announcement in an experiment design summarized in Table 3.1. After the patient is seen by the doctor, patients are surveyed about their wait-time satisfaction in the waiting room[7] (see Appendix 1).

Table 3.1. Experiment Design

Week	Delay Announcement	Condition
1	Off	No Intervention (Condition $C0$)
2	On	Announce 90% overestimation of wait-time (Condition $C1$)
3	On	Announce 70% overestimation of wait-time (Condition $C2$)
4	On	Announce 50% overestimation of wait-time (Condition $C3$)

[6]We used the same research assistant throughout the study to ensure consistency.
[7]No personally identifiable data were collected.

3.4.3. Wait-time Estimation

Using two years of de-identified data of 177,831 patients, we study patient arrival patterns and hospital factors that affect wait-times and develop an institution specific, accurate wait-time predictor model. See Appendix 2 for the details of the prediction model. We found that the most important predictors of patients wait-time are patient ESI level, patient age, the time of the day, the day of the week, the number of patients in the ED, and the number of patients waiting to be admitted.

In our study, we compare time-series methods (rolling averages (Dong et al. 2015) and Holt-Winters (Kalekar 2004)), regression-based methods (Generalized Linear Regression Model, Stepwise Regression Model, Quantile Regression (Sun et al. 2012) and Q-Lasso (Ang et al. 2015)) and machine learning methods (Boosted Regression (Bühlmann and Hothorn 2007)) to predict patients wait-time to see a doctor at the ED under study. We take triage-to-doctor time (i.e., the time between seeing the triage nurse and seeing the doctor for the first time) as a proxy of patients wait-time in the system. This definition of wait-time is due to medical and logistical considerations. Triage-to-doctor time is the largest portion of time a patient waits to be treated, and is shown to be a strong predictor of patient satisfaction (Boudreaux and O'Hea 2004, Shah et al. 2015). Moreover, only at the point of triage, the medical information needed to estimate an individualized wait-time will be available.

We compare the performance of the statistical learning models, using out-of-sample and in-sample errors. The boosted regression model produces the lowest out-of-sample MSE and also provides predictions with particularly low overestimates/underestimates of wait-time. This result can be explained by the model's ability to produce coefficients that

most accurately represented the effect of the variables in our data. This model is therefore used to estimate arriving patients wait-time. Using the final regression model, we compute the $(1-\alpha)\%$ prediction upper bound on the wait-time W for a given $X = (x_1, x_2, \cdots, x_p)$ as follows.

$$\hat{W} + t_{\alpha,n-p}\sqrt{MSE + [se(\hat{W})]^2}$$

where $\hat{W}$ is predicted value or fitted value of the response given X^*, and $t_{\alpha,n-p}$ is the t-multiplier with $n-p$ degrees of freedom (n is the sample size and p is the number of predictors). The term $\sqrt{MSE + [se(\hat{W})]^2}$ is the standard error of the prediction which consists of standard error of the fit and Mean Squared Error (MSE). By announcing the $(1-\alpha)\%$ prediction upper bound on the wait-time to a patient, with probability $(1-\alpha)\%$ the patient waits less than the announced wait-time and experiences a gain in time. For example, by announcing the 70% overestimation of wait-time, we expect to overestimate the wait-times for 70% of patients to whom we announced the wait-time. Note that the amount of overestimation depends on prediction model's standard error.

3.4.4. Data Collection

The data collection was performed on Monday, Tuesday and Wednesday of each week noon to 10 p.m. The reasoning behind our data collection schedule was that the distribution of patients' wait-times is almost the same on Mondays, Tuesdays and Wednesdays and patients wait to see a doctor more often if they arrive between noon to 10 p.m. One research assistant, who was stationed at the designated desk and interacted with patients, while a survey collector collected surveys from patients once they were seen by the doctor. After agreeing to participate, the survey collector asked the survey questions from the

patient (see Appendix 1 for details of survey). Due to infection concerns, no paper-based or I-pad based survey were allowed to be given to patients.

3.4.5. Dependent Variables

Our key dependent variable is: Wait-time satisfaction. We asked patient satisfaction with the wait-time experience between seeing the triage nurse and seeing the doctor ("On a scale of 1 to 7, how satisfied are you with the wait between seeing the triage nurse and seeing the doctor?") as a measure of quality (Edwards 1968, Garvin 1984). Also, we explored other dependent variables: Overall satisfaction and Perceived fairness. We asked patients about their overall experience in ED ("On a scale of 1 to 7, how satisfied are you with your overall experience at our Emergency Department?") We also investigated whether patient perceptions of fairness, regarding the order patients received care, are affected ("On a scale of 1 to 7, how did you feel about the order at which patients receive care in this Emergency Department?"). Table 3.2 summarizes the definitions of all dependent and independent variables.

Table 3.2. Summary of Dependent and Independent Variables Definitions

Variable	Description
	Dependent variables
Wait-time satisfaction	Response to question 3 on the survey: "On a scale of 1 to 7, how satisfied are you with your wait-time in the waiting room (after seeing the triage nurse and before seeing the doctor)?". An integer number from 1 to 7.
Overall satisfaction	Response to question 4 on the survey: "On a scale of 1 to 7, how satisfied are you with your overall experience at our Emergency Department?". An integer number from 1 to 7.
Perceived fairness	Response to question 7 on the survey: "On a scale of 1 to 7, how did you feel about the order at which patients receive care in our Emergency Department?". An integer number from 1 to 7.
	Independent and control variables
C_{90}	Indicator variable which equals to 1 if the 90% overestimation of wait-time is announced and equals to 0 otherwise.
C_{70}	Indicator variable which equals to 1 if the 70% overestimation of wait-time is announced and equals to 0 otherwise.
C_{50}	Indicator variable which equals to 1 if the expected wait-time is announced and equals to 0 otherwise.
Actual wait-time	Patient's actual wait-time between seeing the triage nurse and seeing the doctor.
Perceived wait-time	Response to question 1 on the survey: "How long did you wait in the waiting room after seeing the triage nurse and before seeing the doctor? Provide the best estimate of your wait .....".
Announced wait-time	The estimated wait-time announced to the patients, estimated using the developed wait-time predictor.
Wait-time gap	The difference between the announced wait-time and the actual wait-time.
Wait-time gap sign	Indicator variable which equals to 1 if wait-time gap is positive and equals to 0 otherwise.
High Acuity Indicator	Indicator variable which equals to 1 if the patient has ESI level 2 or higher and equals to 0 otherwise.
First time visit Indicator	Indicator variable which equals to 1 if the patient visited the ED under study for the first time and equals to 0 otherwise.
Male Indicator	Indicator variable which equals to 1 if the patient is male and equals to 0 otherwise.
Age	Patient's age.

3.4.6. Control Variables

We account for additional factors in our analysis that varied over the period of our study. In modeling patient wait-time satisfaction, overall satisfaction and perceived fairness, we control for (I) patients demographics and medical condition, (II) the state of the ED, and (III) time trends. To account for patients demographics, we control for the age and the gender of survey respondents. In addition to these variable, we control for patient ESI level as a proxy for patients chief complaint and pain level. Finally, since treatments were assigned independent of staffing decisions, and we account for the shift of the day and the day of week to show that the primary effects of interest are not driven by outliers corresponding with any particular day or shift. Result patterns appear to fluctuate by experimental condition (see Table 3.2).

3.5. Empirical Models and Results

Our main analyses examine the relative changes in patient wait-time satisfaction when there is delay announcement and when there is no delay announcement. We use linear regression models with time and patient-related control variables to address the proposed hypotheses. In addition to the standard assumptions of linear regression models, to obtain unbiased standard errors, we use robust standard errors, clustered by shift of the day and day of the week.

3.5.1. Overview

Table 3(a) presents means and standard variations of all variables included in the experiment, stratified by the experiment conditions (i.e., C0, C1, C2 and C3). Summary

definition of all variables is presented in Appendix 3. Table 3(b) presents correlations between all continuous variables included in the experiment.

Table 3(a). Summary Statistics of All Variables Included in the Experiment

	No Announcement (C0)			90% overestimation of wait-time (C1)			70% overestimation of wait-time (C2)			50% overestimation of wait-time (C3)			All		
	n	mean	SD	n	mean	SD	n	mean	SD	n	mean	SD	n	mean	SD
Age	87	60	19	105	58	17	91	59	14	90	62	16	373	59	17
Female pct.	87	0.67	0.47	105	0.64	0.48	91	0.68	0.47	90	0.66	0.47	373	0.66	0.47
High Priority pct.	87	0.7	0.46	105	0.74	0.44	91	0.73	0.44	90	0.77	0.42	373	0.73	0.44
First time visit pct.	87	0.25	0.43	105	0.31	0.46	91	0.19	0.39	90	0.21	0.41	373	0.24	0.43
Actual Wait-time (min)	87	163	94	105	160	76	91	158	70	90	166	85	373	161	81
Perceived Wait-time (min)	87	181	101	105	194	96	91	178	85	90	189	103	373	186	96
Announced Wait-time (min)	87	NA	NA	105	246	49	91	206	29	90	185	51	373	214	51
Perceived Announced Wait-time (min)	87	NA	NA	105	208	57	91	166	42	90	165	42	373	181	52
Larger wait-time gap pct.	87	0.59	0.49	105	0.58	0.49	91	0.53	0.5	90	0.66	0.47	373	0.59	0.49
Wait-time Satisfaction (out of 7)	87	3.52	1.98	105	4.19	1.83	91	4.74	1.78	90	3.81	1.96	373	4.08	1.93
Overall Satisfaction (out of 7)	87	5.13	1.8	105	5.82	1.83	91	5.96	1.53	90	5.73	1.53	373	5.67	1.62
Perceived Fairness (out of 7)	87	4.73	1.89	105	5.53	1.82	91	5.67	1.8	90	5.93	1.63	373	5.41	1.85

Table 3(b). Correlations of Continuous Variables Included in the Experiment

Variable	1	2	3	4	5	6	7
1. Wait-time Satisfaction	1.00						
2. Age	-0.12*	1.00					
3. Actual Wait-time	-0.64*	0.07	1.00				
4. Perceived Wait-time	-0.58*	0.10	0.79*	1.00			
5. Announced Wait-time	-0.18*	-0.14*	0.29*	0.36*	1.00		
6. Overall Satisfaction	0.54*	-0.04	-0.39*	-0.46*	-0.17*	1.00	
7. Perceived Fairness	0.42*	0.03	-0.23*	-0.36*	-0.23*	0.57*	1.00

Note. N = 286. Excludes the observations from the control week.
*$p < 0.05$

Closer examination of Table 3(a) leads to some observations. All weeks are similar in terms of average age, female percentage, first time visit percentage and high priority percentage. The average actual wait-times of all patients surveyed under experiment conditions (i.e., actual wait-time under C0 = 163 min, under C1 = 160 min, under C2 = 158 min, and under C3 = 166 min) were close and there was no statistical different in wait-times under different conditions (i.e., the p-value corresponding to the t-test of equality of the means are large). This observation leads us to believe that the reason behind any possible difference in wait-time satisfaction under each condition is not simply the difference between the actual wait-times. None of the correlations between variables in the same regression model have level close to or higher than 0.8, minimizing the concerns about multicollinearity (see Table 3(b)). We also check for multicollinearity by calculating variance inflation factors (VIF), which will be reported later.

3.5.2. Impact of Delay Announcement on Wait-time Satisfaction (H1)

To test Hypothesis 1A-B, we model wait-time satisfaction, S, as a linear function of the treatment condition: (1) indicator variable DA, which equals to 1 if an estimate of wait-time is announced; (2) the actual wait-time, W; and (3) a vector of control variables, X.

$$E[S] = \alpha_0 + \alpha_1 DA + \alpha_2 W + \beta' X \tag{3.1}$$

This specification facilitates the direct interpretation of the coefficient corresponding to the treatment condition (i.e., delay announcement) as the performance difference relative to the baseline control condition (i.e., no delay announcement). Specifically, coefficient α_1 shows the performance difference of delay announcement relative to the baseline condition. The result of this comparison helps us to confirm or reject the Hypothesis 1A. Note that a patient wait-time satisfaction is the response to question 3 on the survey: "*On a scale of 1 to 7, how satisfied are you with your wait-time in the waiting room (after seeing the triage nurse and before seeing the doctor)*?", which is an integer number from 1 to 7. As shown in Table 3(a), we observe that the average wait-time satisfaction is higher when there is delay announcement (i.e., 4.25) than the no announcement scenario (i.e., 3.52). To confirm this observation, we estimate Equation (3.1) to access the impact of delay announcement on patient wait-time satisfaction. To do so, we combine all the data collected under Conditions C1-C3 and compare it with the baseline (i.e., Condition C0). Table 3.4 presents different models for wait-time satisfaction.

In Table 3.4, column 1, we compare the mean wait-time satisfaction under treatment condition. Delay announcement led to statistically large gains in wait-time satisfaction

Table 3.4. Models for Wait Satisfaction Survey Responses

	Dependent variable:			
	Wait-time Satisfaction			
	(1)	(2)	(3)	(4)
1. Delay Announcement Indicator	0.728***	0.701***	0.711***	0.704***
	(0.107)	(0.222)	(0.221)	(0.230)
2. Actual Wait-time		−0.016***	−0.016***	−0.016***
		(0.001)	(0.001)	(0.002)
3. High acuity Indicator			−0.232	−0.207
			(0.071)	(0.159)
4. Age				−0.007
				(0.006)
5. Male Indicator				0.175
				(0.425)
6. First Visit Indicator				−0.193
				(0.258)
Constant	3.517***	6.050***	6.285***	6.658***
	(0.236)	(0.548)	(0.473)	(0.322)
Observations	373	373	373	373
Adjusted R^2	0.023	0.451	0.452	0.454
Pred. difference relative to baseline (%)	20.7	11.59	11.31	10.57

Notes. Parentheses contain robust standard errors, clustered by shift of the day and day of the week. Predicted differences represent percent increase over baseline condition. Although we use OLS in our primary analysis to facilitate coefficient interpretation, we note that all reported results are similar when satisfaction is estimated with an ordinal logistic model. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively (two-tailed tests).

($\alpha_1 = 0.728$, $p < 0.01$). This supports Hypothesis 1A, while rejecting Hypothesis 1B. This result remain robust in the fully specified model, which controls for priority level, age and gender. In the fully specified model, column 4, wait-time satisfaction was more than 10% higher than the baseline when delays announced. To check for multicollinearity in the full model, we use VIFs. The largest VIF is 1.09, which falls well below the conventional threshold of 10, providing evidence that multicollinearity is not a concern (Wooldridge 2015).

3.5.3. Impact of Wait-time Gap on Wait-time Satisfaction (H2)

To test the impact of wait-time gap in wait-time satisfaction (i.e., Hypothesis 2A-B), we, first, model wait-time satisfaction, S, as a piece-wise linear function of wait-time gap Δ, as shown below. We denote $I_{(\Delta>0)}$ as an indicator variables that equals one if the

wait-time gap is positive.

$$E[S] = \alpha_0 + \alpha_1\Delta + \alpha_2\Delta \times I_{(\Delta>0)} \tag{3.2}$$

In this model, α_1 represents the mean wait-time satisfaction change when wait-time gap increases by 1 minute for patients with negative wait-time gap. Similarly, $\alpha_1 + \alpha_2$ represents the mean wait-time satisfaction change when wait-time gap increases by 1 minute for patients with positive wait-time gap. To check if patients are loss averse with respect to their wait-time, we estimate Equation (2). Note that in Equation (2), there is no $I_{(\Delta>0)}$ term, since the connection point is zero (i.e., the reference value for wait-time gap is zero). The equation is derived as shown below:

$$\begin{aligned} E[S] &= \alpha_0 + \alpha_1\Delta + \alpha_2(\Delta - 0) \times I_{(\Delta>0)} \\ &= \alpha_0 + \alpha_1\Delta + \alpha_2\Delta \times I_{(\Delta>0)} \end{aligned}$$

Table 3.5 presents the regression model summary.

Table 3.5. Model for Loss Aversion

	Dependent variable:
	Wait satisfaction
1. Wait Gap	0.026^{***}
	(0.003)
2. Wait Gap $\times I_{\text{Wait Gap}>0}$	-0.019^{*}
	(0.011)
Constant	4.096^{***}
	(0.002)
Observations	286
Adjusted R^2	0.291

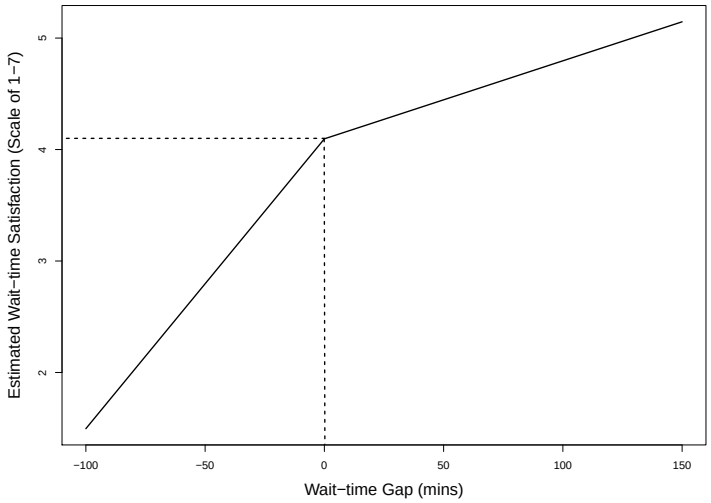
Notes. Parentheses contain robust standard errors. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively (two-tailed tests).

As shown in Table 3.5, we find that $\alpha_1 = 0.026$ is statistically significant ($p < 0.01$) and positive, while the coefficient $\alpha_2 = -0.019$ is statistically significant ($p < 0.01$) and negative. Therefore, the mean wait-time satisfaction change when wait-time gap increases by 1 minute for patients with negative wait-time gap is $\alpha_1 = 0.026$ and that for patients with positive wait-time gap $\alpha_1 + \alpha_2 = 0.007$. This means that 10 minutes increase in wait-time gap increases the mean wait-time satisfaction by 0.26 (in scale of 1-7) in the loss region and increases the mean wait-time satisfaction by 0.07 in the gain region.

Finding α_2 to be negative shows that patients weigh time losses more than time gains, i.e., they are loss averse. Thus, we find support for Hypothesis 2B. Figure 3.3 is used to better visualize the loss aversion. Assuming the wait-time gap equals to zero as the reference value, patient wait-time satisfaction increase in wait-time gap is presented in Figure 3.3. The slope in the negative side of wait-time gap (i.e., 0.026) is more three times than that in the positive side (i.e., 0.007). In other words, the one unit increase of wait-time gap in the negative side, increase the wait-time satisfaction more than that in the positive side.

Moreover, since both slope values are positive, this may lead us to confirm that the wait-time satisfaction is increasing in wait-time gap, for all levels of positive and negative wait-time gap (i.e., Hypothesis 2A is true). However, note that the coefficients in Equation (3.2) are only the average effect of positive and negative wait-time gaps on wait-time satisfaction. Looking at the different levels of positive and negative wait-time gaps may show us different effects. To test the impact of different levels of wait-time gap (Δ) on wait-time satisfaction (i.e., Hypothesis 2A), we classify the data based on how much the

Figure 3.3. The Visualization of Loss Aversion



wait-time was overestimated, using quartiles of wait-time gap and use several piecewise regression models, separately for patients with positive and negative wait-time gap.

For patients whose wait-time gap were positive (i.e., we overestimated their wait-time), the first, second and third quartiles of wait-time gaps in our data are approximately 40 minutes, 80 minutes and 120 minutes, respectively. Using these quartiles, we define Class 1 to be patients whose wait-time gap is positive and less than 40 minutes, Class 2 to patients whose wait-time gap is positive and between 40 minutes and 80 minutes, Class 3 to patients whose wait-time gap is positive and between 80 minutes and 120 minutes, and Class 4 to patients whose wait-time gap is positive and more than 120 minutes. The average wait-time satisfaction for Class 1 is 4.27, for Class 2 is 4.8, for Class 3 is 5.25 and for Class 4 is 5.02. As the wait-time gap increases, patients are on average more satisfied. Except than for Class 4, where the average wait-time gap decreases compared to Class 3. This may suggest more careful analysis of patients who experience large wait-time gaps is necessary. To test Hypothesis 2A, we use the model in Equation (3.3), where $I_{(\Delta>120)}$

is an indicator variables that equals one if the wait-time gap is more than 120 minutes.

$$E[S] = \alpha_0 + \alpha_1\Delta + \alpha_2(\Delta - 120) \times I_{(\Delta>120)} \tag{3.3}$$

In this model, α_1 represents the mean wait-time satisfaction change when wait-time gap increases by 1 minute for patients whose wait-time gap is less than 120 minutes (i.e., less than third quartile of positive wait-time gaps). Similarly, $\alpha_1 + \alpha_2$ represents the mean wait-time satisfaction change when wait-time gap increases by 1 minute for patients whose wait-time gap is more than 120 minutes. To test Hypothesis 2, we estimate Equation (3.3) to further confirm the impact of actual wait-time value on wait-time satisfaction. Table 3.6 summarizes the estimated regression model.

Table 3.6. Model for Large Wait-time Gaps

	Dependent variable:
	Wait-time satisfaction
1. Wait-time Gap	0.012^{***}
	(0.005)
2. Wait-time Gap (> 120 mins)	-0.020^{***}
	(0.011)
Constant	4.009^{***}
	(0.146)
Observations	211
Adjusted R^2	0.045

Notes. Parentheses contain robust standard errors. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively (two-tailed tests).

We find that the coefficient $\alpha_1 = 0.012$ is statistically significant ($p < 0.01$) and positive. This shows that as wait-time gap increases, wait-time satisfaction increases, for patients whose wait-time gap is less than 120 minutes. However, the coefficient $\alpha_2 = -0.02$ is statistically significant ($p < 0.01$) and negative and therefore $\alpha_1 + \alpha_2 = 0.012 - 0.2 = -0.008$. This shows that for patients whose wait-time gap is more than 120 minutes, as wait-time gap increases, the wait-time satisfaction decreases. We find evidence that for

large wait-time gaps, the wait-time satisfaction is not a increasing function of wait-time gap. Thus, we cannot find full support for Hypothesis 2A.

One explanation for this observation can be that a large wait-time gap is often the result of announcing a long wait-time and this long announced wait-time may have a negative initial impact on the patient satisfaction. Receiving a long announced wait-time from system may be a *peak* negative moment (Kahneman et al. 1993, Fredrickson and Kahneman 1993) in a patient's waiting experience that stays with her when evaluating her satisfaction and has a stronger effect on the satisfaction than the positive end effect of a time gain. In a laboratory experiment, Carmon and Kahneman (1996) observe a substantial effect of initial observation of the queue length on the affective response recorded at the end of the queuing episode. Subjects indicate more pleasure at the end of short queues, where they expected wait-time was shorter, than at the end of long ones. This suggests that patients, who take long announced wait-time as bad news delivered to them, may still remember it as an unpleasant moment when evaluating their waiting experience. The negative effect of the long announced wait-time may be one explanation for the dissatisfaction of patients with large positive wait-time gap.

For patients whose wait-time gap was negative (i.e., we underestimated their wait-time), the first, second and third quartiles of wait-time gaps in our data are approximately 20 minutes, 40 minutes and 60 minutes, respectively. We define Class 1 to be patients whose wait-time gap is negative and less than 20 minutes, Class 2 to patients whose wait-time gap is negative and between 20 minutes and 40 minutes, Class 3 to patients whose wait-time gap is negative and between 40 minutes and 60 minutes, and Class 4 to

patients whose wait-time gap is negative and more than 60 minutes. The average wait-time satisfaction for Class 1 is 3.37, for Class 2 is 3.1, for Class 3 is 2.18 and for Class 4 is 1.74.

Similar to Equation (3.3), we develop a piecewise linear regression model for all patients with positive wait-time gap, using positive wait-time gap quartiles as the connection points. To test Hypothesis 3A, we use the model in Equation (3.4), where $I_{(40<\Delta\leq 80)}$ is an indicator variables that equals one if the wait-time gap is between 40 minutes and 80 minutes, $I_{(80<\Delta\leq 120)}$ is an indicator variables that equals one if the wait-time gap is between 80 minutes and 120 minutes, and $I_{(\Delta>120)}$ is an indicator variables that equals one if the wait-time gap is more than 120 minutes.

$$E[S] = \alpha_0 + \alpha_1\Delta + \alpha_2(\Delta - 40) \times I_{(40<\Delta\leq 80)} + \alpha_3(\Delta - 80) \times I_{(80<\Delta\leq 120)} + \alpha_4(\Delta - 120) \times I_{(\Delta>120)} \tag{3.4}$$

Alternatively, we can write the piecewise model as:

$$E[S] = \alpha_0 + \alpha_1\Delta + \alpha_2\Delta_{40} + \alpha_2\Delta_{80} + \alpha_2\Delta_{120}$$

where $\Delta_{40} = (\Delta - 40) \times I_{(40<\Delta\leq 80)}$, $\Delta_{80} = (\Delta - 80) \times I_{(80<\Delta\leq 120)}$ and $\Delta_{120} = (\Delta - 120) \times I_{(\Delta>120)}$. In this model, α_1 represents the mean wait-time satisfaction change when wait-time gap increases by 1 minute for patients whose wait-time gap is less than 40 minutes (i.e., less than first quartile of positive wait-time gaps). Also, $\alpha_1 + \alpha_2$ represents the mean wait-time satisfaction change when wait-time gap increases by 1 minute for patients whose wait-time gap is between 40 minutes and 80 minutes. Similarly, $\alpha_1 + \alpha_3$ represents the mean wait-time satisfaction change when wait-time gap increases by 1 minute for patients

whose wait-time gap is between 80 minutes and 120 minutes. Similarly, $\alpha_1+\alpha_3$ represents the mean wait-time satisfaction change when wait-time gap increases by 1 minute for patients whose wait-time gap is more than 120 minutes. We find that the coefficient $\alpha_1 = 0.011$ is statistically significant ($p < 0.05$) and positive (see Table 3.7 for more details). This shows that as wait-time gap increases, wait-time satisfaction increases, for patients whose wait-time gap is less than 40 minutes. The coefficients α_1 and α_2 were not significant, suggesting that the slope of wait-time gap for Class 2 and Class 3 is positive and not statistically different. However, the coefficient $\alpha_3 = -0.018$ is statistically significant ($p < 0.05$) and negative and therefore $\alpha_1 + \alpha_3 = 0.011 - 0.018 = -0.006$. This shows that for patients whose wait-time gap is more than 120 minutes, as wait-time gap increases, the wait-time satisfaction decreases.

Table 3.7. Piecewise Model for Positive Wait-time Gaps

	Dependent variable:
	Wait-time satisfaction
1. Wait-time Gap	0.011^{**}
	(0.004)
2. Wait-time Gap (Class 2)	-0.0002
	(0.013)
3. Wait-time Gap (Class 3)	0.006
	(0.016)
4. Wait-time Gap (Class 4)	-0.018^{**}
	(0.009)
Constant	4.048^{***}
	(0.217)
Observations	211
Adjusted R^2	0.037

Notes. Parentheses contain robust standard errors. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively (two-tailed tests).

Similar to Equation (3.4), we develop a piecewise linear regression model for all patients with negative wait-time gap, using wait-time gap quartiles as the connection points.

To test Hypothesis 3A, we use the model in Equation (3.5), where $I_{(-20<\Delta\leq-40)}$ is an indicator variables that equals one if the negative wait-time gap is between 20 minutes and 40 minutes, $I_{(-40<\Delta\leq-60)}$ is an indicator variables that equals one if the negative wait-time gap is between 40 minutes and 60 minutes, and $I_{(\Delta>-60)}$ is an indicator variables that equals one if the negative wait-time gap is more than 60 minutes.

$$
\begin{aligned}
E[S] = \alpha_0 + \alpha_1\Delta + \alpha_2(\Delta - (-20)) \times I_{(-20<\Delta\leq-40)} \\
+ \alpha_3(\Delta - (-40)) \times I_{(-40<\Delta\leq-60)} + \alpha_4(\Delta - (-60)) \times I_{(\Delta>-60)}
\end{aligned} \tag{3.5}
$$

The interpretation of the coefficients is similar to Equation (3.4). Table 3.8 summarizes the estimated regression model.

Table 3.8. Piecewise Model for Negative Wait-time Gaps

	Dependent variable:
	Wait-time satisfaction
1. Wait-time Gap	0.023^{***}
	(0.006)
2. Wait-time Gap (Class 2)	-0.018
	(0.018)
3. Wait-time Gap (Class 3)	0.040
	(0.016)
4. Wait-time Gap (Class 4)	-0.011
	(0.021)
Constant	3.615^{***}
	(0.001)
Observations	75
Adjusted R^2	0.276

Notes. Parentheses contain robust standard errors. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively (two-tailed tests).

We find that the coefficient $\alpha_1 = 0.023$ is statistically significant ($p < 0.01$) and positive. This shows that as wait-time gap decreases, wait-time satisfaction decreases, for patients whose negative wait-time gap is less than 20 minutes. There is no significant

difference between the slopes of Class 2-4 and Class 1, since coefficients α_2, α_3 and α_4 are not significant.

Now, to combine all these findings, we develop a piecewise linear regression model for all patients with positive and negative wait-time gap, using only first quartile of negative gaps and third quartile of positive gaps as the connection points. We use the model in Equation (3.6), where $I_{(-20<\Delta\leq 0)}$ is an indicator variables that equals one if the wait-time gap is negative but below 20 minutes, $I_{(0<\Delta\leq 120)}$ is an indicator variables that equals one if the wait-time gap is positive but lower than 120 minutes, and $I_{(\Delta>120)}$ is an indicator variables that equals one if the wait-time gap is more than 120 minutes.

$$E[S] = \alpha_0 + \alpha_1\Delta + \alpha_2(\Delta - (-20)) \times I_{(-20<\Delta\leq 0)} + \alpha_3\Delta \times I_{(0<\Delta\leq 120)} + \alpha_4(\Delta - 120) \times I_{(\Delta>120)} \tag{3.6}$$

The interpretation of the coefficients is similar to Equation (3.4). Table 3.9 summarizes the estimated regression model.

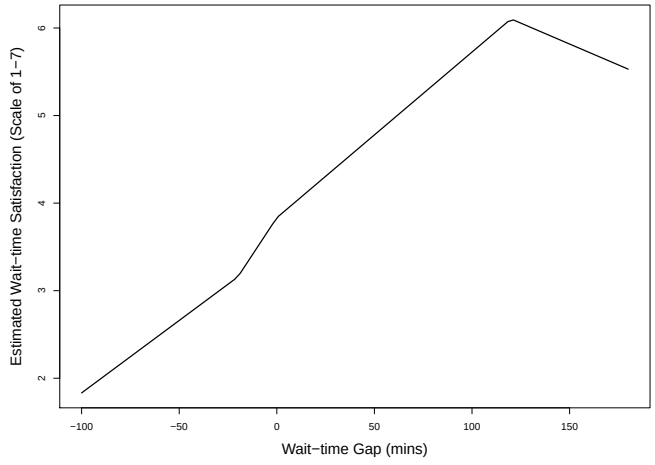
Table 3.9. Piecewise Model for All levels of Wait-time Gaps

	Dependent variable:
	Wait-time satisfaction
1. Wait-time Gap	0.0165***
	(0.006)
2. Wait-time Gap (Neg. and < 20 mins)	0.0173*
	(0.030)
3. Wait-time Gap (Pos. and < 120 mins)	0.0024*
	(0.005)
4. Wait-time Gap (> 120 mins)	−0.0261***
	(0.008)
Constant	3.487***
	(0.201)
Observations	286
Adjusted R^2	0.317

Notes. Parentheses contain robust standard errors. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively (two-tailed tests).

We find that the coefficient $\alpha_1 = 0.0165$ is statistically significant ($p < 0.01$) and positive. We also find that slope for negative wait-time gaps below 20 minutes to be $\alpha_1 + \alpha_2 = 0.0338$. This shows that as wait-time gap decreases, wait-time satisfaction decreases, for patients with negative wait-time gap and the decrease is more for patients whose wait-time gap is negative and smaller than 20 minutes. For patients with positive and below 120 minutes wait-time gap, the slope is $\alpha_1 + \alpha_3 = 0.0189$ and for patients with positive and above 120 minutes wait-time gap, the slope is $\alpha_1 + \alpha_4 = -0.0096$. Figure 3.4 summarizes the piecewise model for all levels of wait-time gaps. Note that adding more connection points may increases the fit accuracy but decreases the statistical significance of the results. Therefore, we used only 3 connection points for our piecewise linear model.

Figure 3.4. The Visualization of Impact of Wait-time Gap on Wait-time Satisfaction



Therefore, we find no significant difference between the effect of all 4 classes on wait-time satisfaction. As wait-time gap decreases in the loss region, the wait-time satisfaction also decreases. This supports Hypothesis 2A for the levels of negative wait-time gap.

3.5.4. Impact of Actual Wait-time on Wait-time Satisfaction (H3)

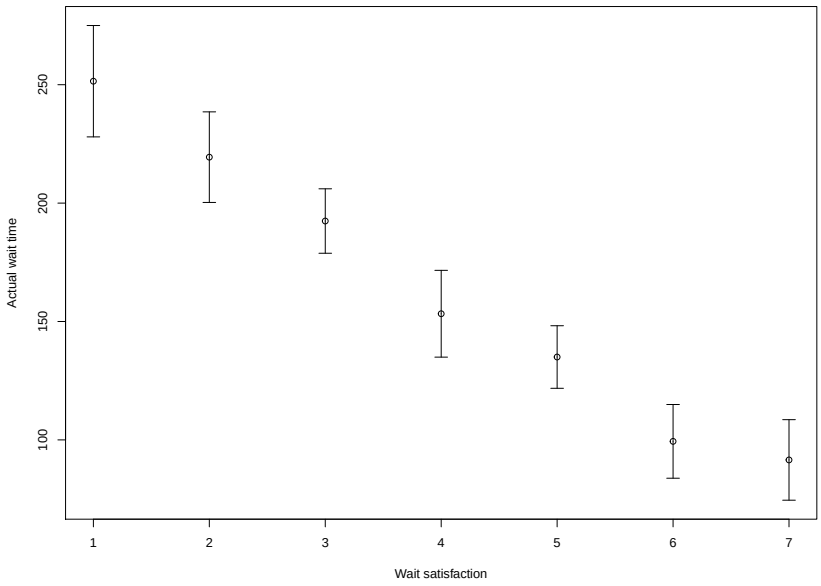
To test Hypothesis 3, we use model wait satisfaction, S, as a linear function of, W, the actual wait-time, Δ, wait-time gap, and $I_{(\Delta>0)}$, the indicator wait-time gap sign to be positive, as shown below. We control for the wait-time gap and the gap's sign, since Prospect Theory states that people derive utility from gains and losses measured relative to a reference point.

$$E[S] = \alpha_0 + \alpha_1 W + \alpha_2 \Delta + \alpha_3 \Delta \times I_{(\Delta>0)} \tag{3.7}$$

In this model, α_1 represents the actual wait-time of the patients between seeing the triage nurse and the first visit by a doctor (i.e., triage-to-doctor time). Coefficients α_2 and α_3 can be interpreted as it is explained in previous section. If the coefficient α_1 is statistically significant and negative Hypothesis 3 will be confirmed. This would confirm that we should not ignore the actual wait-time, when aiming for improving patient satisfaction.

To examine the impact of actual wait-time on wait-time satisfaction, first, we can use Equation (3.1) and Table 4. As shown in Table 3.4 column (4), the coefficient of the actual wait-time is statistically significant and negative ($\alpha_1 = -0.016$ and $p < 0.01$). As expected, as the actual wait-time increases, the wait-time satisfaction decreases. To further demonstrate this effect, Figure 3.5 shows the average actual wait-time with a 95% Confidence Interval for each wait-time satisfaction score. Patients who waited the longest evaluate their wait-time satisfaction to be the worst by choosing 1 or 2 on the survey.

Figure 3.5. Actual Wait-time and Wait-time Satisfaction



To test Hypothesis 3, we estimate Equation (4) to further confirm the impact of actual wait-time value on wait-time satisfaction. Table 3.10 summarizes the estimated regression model.

Table 3.10. Model for Impact of Actual Wait-time on Wait-time Satisfaction

	Dependent variable:
	Wait-time Satisfaction
1. Actual Wait-time	-0.015^{***}
	(0.004)
2. Wait-time Gap	0.014^{***}
	(0.003)
3. $I_{extWait-timeGap>0}$	-0.018^{***}
	(0.003)
4. Wait-time Gap $\times I_{\text{Wait-time Gap}>0}$	7.068^{***}
	(0.872)
Observations	286
Adjusted R^2	0.433

Notes. Parentheses contain robust standard errors. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively (two-tailed tests).

As shown in Table 3.10, the actual wait-time is a significant predictor of wait-time satisfaction, when we control for the wait-time gap. We find that the coefficient $\alpha_1 = -0.015$ is statistically significant ($p < 0.01$) and negative, while α_2 and α_3 are still statistically significant. As actual wait-time value increases, the mean wait-time satisfaction decreases. This supports Hypothesis 3.

3.5.5. Additional Hypotheses and Analyses.

To better understand our findings and consider potential alternative explanations and possible outcomes, we conduct several additional analyses. First, we explored if delay announcement changed patients' perceived wait-time, by making them more sensitive to their wait-time. Patients are not usually accurate in their estimation of actual waiting times and they tend to overestimate the amount of time from triage until initial examination by the emergency physician (Thompson et al. 1996a). Delay announcement may intensify this overestimation, by making patients more sensitive to their wait-time. We explore this impact in Hypothesis 4 below.

Hypothesis 4. *Delay announcement increases the perception of wait-time.*

Since the perception of waiting time better predicts satisfaction than the actual waiting time (Davis and Heineke 1998), it is important to take the effect of possible overestimation into account. If delay announcement causes patients to perceive their wait-time to be even longer than the actual, we should be more cautious how much we overestimate the wait-time.

To test Hypothesis 4, we model perceived wait-time, $\dot{W}$, as a linear function of each treatment condition: (1) indicator variable C_{90}, which equals to 1 if 90% overestimation

of wait-time of patients' wait-time to see a doctor is announced; (2) indicator variable C_{70}, which equals to 1 if 70% overestimation of wait-time of patients' wait-time to see a doctor is announced; (3) indicator variable C_{50}, which equals to 1 if the expected patients' wait-time to see a doctor is announced; (5) the actual wait-time, W; and (6) a vector of control variables, X.

$$E[\dot{W}] = \alpha_0 + \alpha_1 C_{90} + \alpha_2 C_{70} + \alpha_3 C_{50} + \alpha_4 W + \beta' X \tag{3.8}$$

Coefficients α_1, α_2 and α_3 show the difference in perceived wait-time of Conditions C1 (i.e., announcing 90% overestimation of wait-time), C2 (i.e., announcing 70% overestimation of wait-time), C3 (i.e., announcing 50% overestimation of wait-time) relative to the baseline condition, respectively. The result of this comparison helps us to confirm or reject the Hypothesis 4. Coefficient α_4 shows how much patients notice the passage of time and whether they are aware of how much they actually wait. Note that a patient's perceived wait-time is the response to question 1 on the survey: "*How long did you wait in the waiting room after seeing the triage nurse and before seeing the doctor? Provide the best estimate of your wait*".

As shown in Table 3a, even though patients average perceived wait-time patients is close for all weeks (i.e., average perceived wait-time under C0 = 181 min, under C1 = 194, under C2 = 178, and under C3 = 189), patients perceive their wait-time to be 20 to 30 minutes higher than the actual wait-time on average in all weeks (i.e., under C0: 181 min > 163 min, under C1: 194 min > 160 min, under C2: 178 min > 158 min, and under C3: 189 min > 166 min).

To check the relationship between actual wait-times and patients' perception of wait-time, we compare all survey patients actual wait-time and perceived wait-time using Welch two sample t-test. The t-test confirms that patients perceive their wait-time to be longer than the actual ($t(724) = -3.72,\ p < 0.01$), as it is also observed by Thompson et al. (1996b).

Table 3.11. Model for Impact of Delay Announcement on Perceived Wait-time

	Dependent variable:
	Perceived Wait-time
1. 90% overestimation of wait-time	15.728
	(20.024)
2. 70% overestimation of wait-time	2.512
	(10.525)
3. 50% overestimation of wait-time	4.343
	(12.464)
4. Actual wait-time	0.896***
	(0.035)
5. High acuity Indicator	−1.788
	(5.360)
6. Age	−1.657
	(5.406)
7. Male Indicator	0.196
	(0.409)
8. First Visit Indicator	6.706
	(7.885)
Constant	23.595
	(28.192)
Observations	373
Adjusted R^2	0.575

Notes. Parentheses contain robust standard errors. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively (two-tailed tests).

To test Hypothesis 4, we estimate Equation (3.8). Table 3.11 summarizes the estimated regression model. As shown in Table 3.11, we can further confirm that perceived wait-time is positively correlated with actual wait-time ($\alpha_4 = 0.896$ and $p < 0.01$). However, there is no significant difference between all weeks in how patients perceive wait-time ($\alpha_1 = 15.728,\ p = NS,\ \alpha_2 = 2.512,\ p = NS,\ \alpha_3 = 4.343,\ p = NS$). Thus, we cannot find any support for Hypothesis 4. Also, none of the control variables found to be a

significant predictor of the perceived wait-time. Thus, we found no significant difference between all weeks in how patients perceive wait-time. This suggests that patients are fairly aware of their wait and delay announcement did not make patients more sensitive to their wait-time.

Second, we explored how delay announcement changed the proportion of *very unsatisfied* patients (i.e., those who choose 1 or 2 as their wait-time satisfaction in the surveys) and *very satisfied* (i.e., those who choose 6 or 7 as their wait-time satisfaction). It is important to increase the number of customers who are very satisfied with our service and reduce the number of customers so unhappy that they speak out against us and discourage other potential customers from using our service. Leading service companies always try to quantify customer satisfaction to drive customer loyalty. For example, Heskett et al. (1994) reports that Xerox polled 480,000 customers per year regarding product and service satisfaction using a five-point scale. To achieve 100% satisfaction, which they defined to be that customers choose 4s and 5s, they analyzed the relationship between the scores and actual loyalty. They found that customers who gave Xerox 5, were six times more likely to repurchase Xerox equipment than those giving 4s (6 times more loyalty). This study shows how important is to increase the number of patients who are very satisfied with our service. Just as important for profitability is to avoid creating, so called *terrorist* customer, i.e., customers so unhappy that they speak out against a poorly delivered service at every opportunity and discourage other potential customers from using the service. These findings can be easily extended to ED setting, where we want to maximize the number of very satisfied patients while minimizing the number of dissatisfied patients for a scale of 1 to 7 in our survey. It is not clear how delay announcement may impact

the percentage of very satisfied and very unsatisfied patients. Delay announcement may only increase the percentage of very satisfied patients or decrease the percentage of very unsatisfied patients or both (or even none). We explore this impact in Hypothesis 5A below.

Hypothesis 5A. *Delay announcement increases the percentage of very satisfied patients.*

Hypothesis 5B. *Delay announcement decreases the percentage of very unsatisfied patients.*

This analysis helps us understand whether delay announcement reduce the number of very unsatisfied patients or it actually increase the number of very satisfied patients. ED managers may communicate more with the patients throughout the visit process to amplify the delay announcement's positive impacts.

To test Hypothesis 5A-B, we compare wait-time satisfaction scores under Conditions C1 (i.e., announcing 90% overestimation of wait-time), C2 (i.e., announcing 70% overestimation of wait-time) and C3 (i.e., announcing 50% overestimation of wait-time) to the baseline condition in terms of very satisfied and very unsatisfied patients percentage. As a reminder, We define very unsatisfied to be patients who choose 1 or 2 as their wait-time satisfaction in the surveys and very satisfied to be patients who choose 6 or 7 as their wait-time satisfaction. We compute the percentage of very satisfied and very unsatisfied patients under each conditions to clarify whether delay announcement under each condition increases the number of happy patients (i.e., patients who choose 6 or 7) or decreases the number of unhappy patients (i.e., patients who choose 1 or 2) or both. We, then,

use series of two-proportion z-test to compare the impact of delay announcement under different conditions.

Table 3.12 shows the very satisfied and very unsatisfied patients percentage under each conditions. The percentages provided in Table 3.12 represents the ratio of patients who choose 1 or 2 (i.e., the very unsatisfied patients percentage) and 6 or 7 (i.e., the very satisfied patients percentage) under each condition.

Table 3.12. The Very Satisfied and Very Unsatisfied Patients Percentage

Statistics	**Condition C0 Base**	**Condition C1 90% UB**	**Condition C2 70% UB**	**Condition C3 50% UB**
Very Satisfied Patients %	17%	26%	42%	21%
(95% CI)	(10%,24%)	(19%,33%)	(33%,51%)	(14%,28%)
Very Unsatisfied Patients %	33%	18%	12%	29%
(95% CI)	(25%,41%)	(12%,24%)	(6%,18%)	(21%,37%)

As shown in Table 3.12, with announcing the wait-times (i.e., under all conditions) the percentage of very satisfied patients (C1: 26%, C2: 42%, and C3: 21%) increased compared to the baseline (17%) and the very unsatisfied patients percentage (C1: 18%, C2: 12%, and C3: 29%) decreased compared to the baseline (33%). The most significant increase in the very satisfied patients percentage and concurrently decrease in the very unsatisfied patients percentage occurred under condition C2 (i.e., 70% overestimation of wait-time). This suggests that by carefully designing the delay announcement procedure, we not only can increase happy patients but also make the unhappy patients less unhappy.

To compare the very satisfied and very unsatisfied patients percentages among experiment conditions, we use series of two-proportion z-tests. The z-test also confirms that the percentage of both very satisfied patients ($z = 12.786,\ p < 0.01$) and very unsatisfied patients ($z = 11.523,\ p < 0.01$) are statistically different and higher for Condition C2 (i.e., 70% overestimation of wait-time) compare to the baseline. The percentage of very

satisfied and very unsatisfied patients did not find to be statistically different than the baseline (very satisfied patients, C1: $z = 1.998,\ p = NS$ C2: $z = 0.427,\ p = NS$, very unsatisfied patients, C1: $z = 5.892,\ p < 0.05$ C2: $z = 0.408,\ p = NS$). Thus, we found that the delay announcement not only can increase happy patients but also make the unhappy patients less unhappy.

Third, we explored whether delay announcement may make patients more sensitive to their surroundings or more attentive to the ordering system. ED works as a priority system and patients are not seen according to their order of arrival. Patients may notice that patients who arrived after them to be taken to a care space earlier and therefore perceive the system to be unfair. The ability to secure fairness or equity between patients might stand as a way to improve satisfaction in EDs. Announcing personalized delay may help patients build trust on the system's ordering mechanism and therefore accept its priority nature. On the other hand, delay announcement may make patients more sensitive to their surroundings or more attentive to the order of service. This potentially may lead them to feel that the system is unfair. Thus, we explore whether delay announcement have any impact on patients perceived fairness using the following hypotheses.

Hypothesis 6. *Delay announcement increases perceived fairness.*

If delay announcement increases perception of fairness (i.e., Hypothesis 6 is confirmed), this further encourages ED managers to use delay announcement to improve patient wait-time satisfaction.

To test Hypothesis 6, we model perceived fairness, F, as a linear function of each treatment condition: (1) indicator variable C_{90}, which equals to 1 if 90% overestimation

of wait-time of patients' wait-time to see a doctor is announced; (2) indicator variable C_{70}, which equals to 1 if 70% overestimation of wait-time of patients' wait-time to see a doctor is announced; (3) indicator variable C_{50}, which equals to 1 if the expected patients' wait-time to see a doctor is announced; (5) the actual wait-time, W; and (6) a vector of control variables, X.

$$E[F] = \alpha_0 + \alpha_1 C_{90} + \alpha_2 C_{70} + \alpha_3 C_{50} + \alpha_4 W + \beta' X \tag{3.9}$$

Coefficients α_1, α_2 and α_3 show the performance difference in perceived fairness of Conditions C1 (i.e., announcing 90% overestimation of wait-time), C2 (i.e., announcing 70% overestimation of wait-time), C3 (i.e., announcing 50% overestimation of wait-time) relative to the baseline condition, respectively. The result of this comparison helps us to confirm or reject the Hypothesis 6. Note that a patient's perceived fairness is the response to question 7 on the survey: "*On a scale of 1 to 7, how did you feel about the order at which patients receive care in our Emergency Department?*", which is an integer number from 1 to 7.

As shown in Table 3a, the average perceived fairness under all conditions with delay announcement increased compared to the control condition (i.e., perceived fairness under C0 = 4.73, under C1 = 5.53, under C2 = 5.67, and under C3 = 5.93), except than for condition C3 (i.e., announcing 50% overestimation of wait-time). Under condition C3, patients perceived system to be more fair on average compared to other conditions. We estimate an equation similar to 3.9 but for perceived fairness to access the impact of delay announcement on patients perceived fairness. Table 3.13 presents different models for perceived fairness.

Table 3.13. Models for Perceived Fairness Survey Responses

	Dependent variable:			
	Perceived Fairness			
	(1)	(2)	(3)	(4)
1. 90% overestimation of wait-time	0.794***	0.718***	0.719***	0.746***
	(0.220)	(0.144)	(0.149)	(0.182)
2. 70% overestimation of wait-time	0.941***	0.893**	0.894**	0.905**
	(0.316)	(0.369)	(0.424)	(0.409)
3. 50% overestimation of wait-time	1.200***	1.248***	1.248***	1.226***
	(0.100)	(0.216)	(0.237)	(0.209)
4. Actual wait-time		−0.007***	−0.007***	−0.007***
		(0.002)	(0.001)	(0.001)
5. High acuity Indicator			−0.022	−0.050
			(0.561)	(0.540)
6. Age				0.009
				(0.009)
7. Male Indicator				0.020
				(0.337)
8. First Visit Indicator				0.067
				(0.293)
Constant	4.732***	5.915***	5.938***	5.434***
	(0.209)	(0.608)	(0.200)	(0.508)
Observations	274	274	274	274
Adjusted R^2	0.053	0.158	0.155	0.152
Pred. difference relative to baseline (90% overestimation of wait-time) (%)	16.78	12.14	12.11	13.73
Pred. difference relative to baseline (70% overestimation of wait-time) (%)	19.89	15.1	15.06	16.65
Pred. difference relative to baseline (50% overestimation of wait-time) (%)	25.36	21.10	21.02	22.56

Notes. Parentheses contain robust standard errors, clustered by shift of the day and day of the week. Predicted differences represent percent increase over baseline condition. Although we use OLS in our primary analysis to facilitate coefficient interpretation, we note that all reported results are similar when satisfaction is estimated with an ordinal logistic model. Some patients did not respond to the perceived fairness questions and are excluded from this model. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively (two-tailed tests).

In Table 3.13, column 1, we compare the mean perceived fairness under each experimental condition. All conditions with delay announcement led to statistically large gains in perceived fairness ($\alpha_1 = 0.794$, $p < 0.01$, $\alpha_2 = 0.941$, $p < 0.01$ and $\alpha_3 = 1.2$, $p < 0.01$). This result remain robust in the fully specified model, which controls for priority level, age and gender. Perceived fairness was 13.73% higher than the baseline when 70% overestimation of wait-time, 16.65% above the baseline when announcing 90% overestimation of wait-time and 22.56% above the baseline when announcing 50% overestimation of wait-time. This results suggests that delay announcement helped improving perceived fairness of ED and the maximum increase achieved under C3. The reason behind this increase in perceived fairness might be that delay announcement helped patients' acceptance of the

ordering mechanism or reduced patient's sensitivity to being "jumped" in the line. Thus, we observed that delay announcement helps improving perceived fairness of ED.

Fourth, in addition to wait-time satisfaction, we need to study the impact of delay announcement on the overall satisfaction with the ED visit, since any increase in overall satisfaction can directly translate to an increase in HCAHPS scores, which means more revenue and government funding for the hospital. We discussed earlier why also delay announcement may or may not increase the wait-time satisfaction. Since overall satisfaction is a function of the wait-time satisfaction, we also expect all those discussions to be relevant for overall satisfaction. We explore the impact of delay announcement on overall satisfaction in Hypothesis 7-B.

Hypothesis 7. *Delay announcement increases overall satisfaction.*

If delay announcement increases overall satisfaction (i.e., Hypothesis 7 is confirmed), this further encourages ED managers to use delay announcement to improve patients evaluation of the service experience, which as already discussed have direct financial benefits for the hospitals and the society.

To test Hypothesis 7, we model overall satisfaction, $\mathcal{S}$, as a linear function of each treatment condition: (1) indicator variable C_{90}, which equals to 1 if 90% overestimation of wait-time of patients' wait-time to see a doctor is announced; (2) indicator variable C_{70}, which equals to 1 if 70% overestimation of wait-time of patients' wait-time to see a doctor is announced; (3) indicator variable C_{50}, which equals to 1 if the expected patients' wait-time to see a doctor is announced; (5) the actual wait-time, W; and (6) a vector of

control variables, X.

$$E[\mathcal{S}] = \alpha_0 + \alpha_1 C_{90} + \alpha_2 C_{70} + \alpha_3 C_{50} + \alpha_4 W + \beta' X \tag{3.10}$$

Coefficients α_1, α_2 and α_3 show the performance difference in overall satisfaction of Conditions C1 (i.e., announcing 90% overestimation of wait-time), C2 (i.e., announcing 70% overestimation of wait-time), C3 (i.e., announcing 50% overestimation of wait-time) relative to the baseline condition, respectively. The result of this comparison helps us to confirm or reject the Hypothesis 7. Note that a patient's overall satisfaction is the response to question 4 on the survey: "*On a scale of 1 to 7, how satisfied are you with your overall experience at our Emergency Department?*", which is an integer number from 1 to 7.

Moreover, Boudreaux and O'Hea (2004) classify the statistically significant predictors of patient overall satisfaction, studied in the literature of patient satisfaction, to be: interpersonal interaction with providers, perceived technical skills of providers, perceived waiting times, actual waiting times, patient characteristics, visit characteristics, and facility characteristics. Considering the available data, we use relevant regression models to determine to what extend wait-time satisfaction predicts the overall satisfaction.

As shown in Table 3a, the average overall satisfaction under all conditions with delay announcement increased compared to the control condition (i.e., overall satisfaction under C0 = 5.13, under C1 = 5.82, under C2 = 5.96, and under C3 = 5.73). We estimate an equation similar to 3.10 but for overall satisfaction to access the impact of delay announcement on patients overall satisfaction. Table 3.14 presents different models for overall satisfaction.

Table 3.14. Models for Overall Satisfaction Survey Responses

	Dependent variable:			
	overall satisfaction			
	(1)	(2)	(3)	(4)
1. 90% overestimation of wait-time	0.693***	0.665***	0.658***	0.641***
	(0.182)	(0.141)	(0.109)	(0.091)
2. 70% overestimation of wait-time	0.830***	0.784***	0.782***	0.786***
	(0.124)	(0.090)	(0.150)	(0.169)
3. 50% overestimation of wait-time	0.607***	0.639***	0.624***	0.637***
	(0.204)	(0.180)	(0.153)	(0.193)
4. Actual wait-time		−0.009***	−0.008***	−0.008***
		(0.002)	(0.002)	(0.002)
5. High acuity Indicator			0.197	0.216
			(0.266)	(0.283)
6. Age				−0.004
				(0.007)
7. Male Indicator				0.054
				(0.288)
8. First Visit Indicator				0.134
				(0.366)
Constant	5.126***	6.533***	6.333***	6.510***
	(0.148)	(0.156)	(0.102)	(0.370)
Observations	373	373	373	373
Adjusted R^2	0.029	0.217	0.217	0.214
Pred. difference relative to baseline (90% overestimation of wait-time) (%)	13.52	10.18	10.39	9.85
Pred. difference relative to baseline (70% overestimation of wait-time) (%)	16.19	12.00	12.35	12.07
Pred. difference relative to baseline (50% overestimation of wait-time) (%)	11.84	9.78	9.85	9.78

Notes. Parentheses contain robust standard errors, clustered by shift of the day and day of the week. Predicted differences represent percent increase over baseline condition. Although we use OLS in our primary analysis to facilitate coefficient interpretation, we note that all reported results are similar when satisfaction is estimated with an ordinal logistic model. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively (two-tailed tests).

In Table 3.14, column 1, we compare the mean overall satisfaction under each experimental condition. All conditions with delay announcement led to statistically large gains in overall satisfaction ($\alpha_1 = 0.693$, $p < 0.01$, $\alpha_2 = 0.830$, $p < 0.01$ and $\alpha_3 = 0.607$, $p < 0.01$). This result remain robust in the fully specified model, which controls for priority level, age, gender and previous experience with the ED. Wait-time satisfaction was 12% higher than the baseline when 70% overestimation of wait-time, 9.85% above the baseline when announcing 90% overestimation of wait-time and 9.78% above the baseline when announcing 50% overestimation of wait-time. This results suggests that delay announcement helped improving overall satisfaction of ED and the maximum increase achieved under C2. Table 3b suggests that overall satisfaction is positively correlated with the wait-time satisfaction and perceived fairness. We confirmed this observation

by estimating these relationships in a regression model with wait-time satisfaction and perceived fairness as a predictor of overall satisfaction. This model has adjusted R^2 of 53%, showing that wait-time satisfaction and perceived fairness explain a little bit more than 50% of the variability of overall satisfaction. This suggests that any improvement in wait-time satisfaction and perceived fairness can potentially have a significant impact on the overall satisfaction as well. The coefficient of wait-satisfaction and perceived wait-time are both positive and statistically significant in the corresponding regression model ($\alpha_1 = 0.3473$, $p < 0.01$, $\alpha_2 = 0.369$, $p < 0.01$, respectively).

Finally, we consider the potential impact of delay announcement on patients abandonments, which is measured as the left without being seen (LWBS) rate. Patients who learn that they have to wait for several hours to see the doctor may decide to leave the ED, which is a health risk for them and revenue loss for the hospital. Thus, this possible negative impact of announcing the delay on LWBS rates needs to be studied. Thus, we investigate whether delay announcement have an impact on LWBS rate using the following hypotheses.

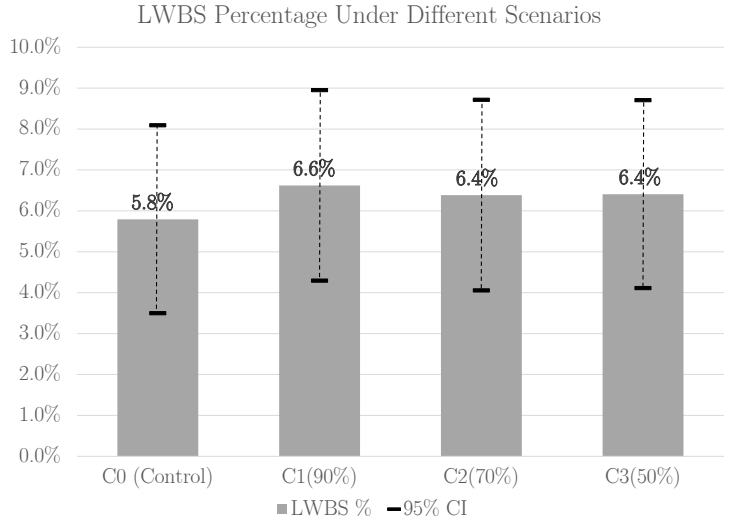
Hypothesis 8. *Delay announcement increases LWBS rate.*

If delay announcement increases LWBS rate (i.e., Hypothesis 8 is confirmed), hospitals may design policies or protocols to control this negative consequence. Providing more information about the medical status of the patient and communicate the reason for the delays may be good practices to control LWBS rates.

To test Hypothesis 8, we obtain data on whether a patient left without being seen by a doctor from the hospital's Electronic Medical Record (EMR) system. We compute

LWBS rate under each conditions with the corresponding 95% confidence interval to study this impact. We use series of two-proportion z-tests to compare the LWBS rate under experiment conditions. Figure 3.6 demonstrates the LWBS rate under each condition with the 95% Confidence Interval.

Figure 3.6. Left Without Being Seen Rate Under Different Conditions



As shown in Figure 3.6, even though the LWBS rate increased in presence of delay announcement compare to the base condition, where there was no delay announcement, the difference is not significant. The highest LWBS rate was observed in the week we announced 90% overestimation of wait-time, since on average longer wait-times were announced. We also run a series of two proportion z-test to confirm the LWBS rates are statistically not difference under experiment conditions. None of the tests were significant. Therefore, we cannot find enough evidence to support Hypothesis 8 and thus delay announcement did not increase LWBS rate significantly.

3.6. Engineering the Delay Announcement

In this section, we formulate a patient wait-time satisfaction as a function of her wait-time gap and actual wait-time, motivated by the results found in the previous sections. We then, fit the proposed function on the experiment data. We use this function to find what wait-time to announce to maximize the total average patient wait-time satisfaction, i.e., engineer the delay announcement. We define total average patient wait-time satisfaction to be the average wait-time satisfaction of all patients visiting ED in a specific period.

3.6.1. Wait-time Satisfaction Function

The utility from a service experience consists of two parts: acquisition and transition utility (Kahneman et al. 2003, Thaler 1985). The former reflects the value of receiving the service, and latter is the psychological value of the waiting, determined by the wait-time gap $\Delta = a - w$ between the $a$nnounced wait-time, a, and the actual $w$ait-time, w. Prospect theory (Kahneman and Tversky 1979) identifies key properties of the transaction utility. Specifically, utility increase in the magnitude of the reference gap, Δ, and it is more sensitive to longer-than-expected waits than shorter-than-expected waits of the same magnitude (i.e., loss aversion). The loss aversion behavior was confirmed in our study in Hypothesis 2B. Motivated by Hypotheses 2A-B and 3, we formally define satisfaction derived from a patient's wait-time experience, S, as a function of actual wait-time (w) and wait-time gap (Δ), presented in the following equation.

$$S(a,w) = S_0(w) + S_1(a-w) = S_0(w) + \begin{cases} G(a-w) & \text{if } w \le a \\ L(a-w) & \text{if } w > a \end{cases} \tag{3.11}$$

In Equation (3.11), $S_0(w)$ represents the direct impact of actual wait-time on patient wait-time satisfaction (i.e., the rational effect of wait-times). This term is motivated by the finding in Hypothesis 3. This can be a linear or nonlinear function of the actual wait-time (w). The behavioral effect of wait-times is formulated by $S_1(a-w)$. When the actual wait-time is shorter than the reference wait-time (i.e., $w \leq a$), the patient experience a gain in time and this gain is formulated as the function $G(a-w)$, which is a function of the difference of the actual wait-time and the announced wait-time. On the other hand, when the actual wait-time is longer than the announced wait-time (i.e., $w > a$), the patient experience a loss in time and this loss is formulated as the function $L(a-w)$, which is also a function of the difference of the actual wait-time and the announced wait-time. This term is motivated by the findings in Hypotheses 2A-B. In the following section, we use the data collected in our experiment at ED to fit Equation (3.11) on patient wait-time satisfaction. This functional form also matches what is proposed by Spiegler (2011), who also discusses loss aversion and reference dependent individual's decision making, but in the retail setting.

3.6.2. Fitting on the data

To fit Equation (3.11) on patient wait-time satisfaction, we need to assume a functional form for $S_0(w)$, $L(a-w)$ and $G(a-w)$. We assume that $S_0(w) = \beta_0 + \beta_1 \log(w)$, $L(a-w) = \tau^-(a-w)$ and $G(a-w) = \tau_0^+(a-w) + \tau_1^+(a-w)^3$. This choice of our functional forms are motivated by our findings in previous sections and data fitting considerations. For example, we choose the power 3 in function G to capture the observed decrease in satisfaction after a threshold and because it fits best to our data. Here, we expect

$\beta_0 > 0$ and $\beta_1 < 0$, based on Hypothesis 3. Also, we expect $\tau^- < 0$ and $\tau_0^+ > 0$ to capture patients' loss aversion behavior, which was confirmed in Hypothesis 2B. Finally, we expect $\tau_1^+ < 0$ to capture the finding that the wait-time gap should not be too large (i.e., Hypothesis 2A). Using data from our experiment, we find the best values for parameters in the Equation (3.11) using linear regression and least-squared methods. The wait-time satisfaction function fitted is as follows:

$$S(a,w) = 11.99 - 1.55\log(w) + \begin{cases} 0.007(a-w) - (3.2\text{e}{-07})(a-w)^3 & \text{if } w \le a \\ -0.017(a-w) & \text{if } w > a \end{cases} \quad (3.12)$$

Note that this function satisfies all the proposed properties for a utility function in Prospect Theory, expect than for the large positive wait-time gaps, we assume that the wait-time satisfaction decreases as the wait-time gap increases.

Figure 3.7. Fitted Wait-time Satisfaction Function Visualization of $S_1(\Delta)$

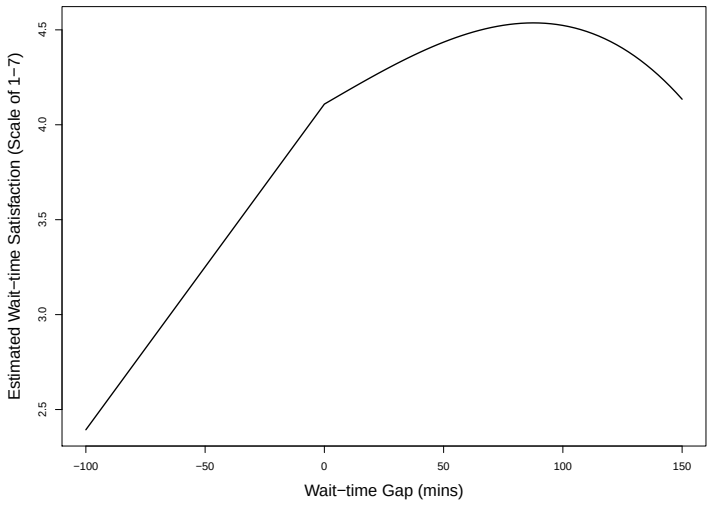


Figure 3.7 shows how wait-time satisfaction changes as a function of wait-time gap $\Delta = a - w$. We used our experiment data to generate the values and fixed the value of

w in $S_0(w)$ to the average wait-time of arriving patients. Even though, the loss aversion is picked-up by the model, there is a threshold on positive wait-time gap after which the estimated wait-time satisfaction starts to decrease. This fully matches with our empirical findings. With a utility model in mind, we now turn back the question of what wait-time to announce to maximize patient wait-time satisfaction.

3.7. What Wait-time to Announce?

To answer this question, we run two sets of regression models. First, we compare wait-time satisfaction between experiment conditions C0-C3. This comparison helps us evaluate the effect of announcing a wait-time median (or mean) with announcing an over-estimation of wait-times. Second, combining the data of weeks with delay announcement (i.e., week 2-4), we reclassify patients based on how much we overestimate their wait-times and compare the wait-time satisfaction of these classes with baseline. This comparison helps us find how much to overestimate to maximize the wait-time satisfaction. First, we model wait-time satisfaction, S, as a linear function of each treatment condition: (1) indicator variable C_{90}, which equals to 1 if 90% overestimation of wait-time is announced; (2) indicator variable C_{70}, which equals to 1 if 70% overestimation of wait-time is announced; (3) indicator variable C_{50}, which equals to 1 if the expected patients' wait-time is announced; (5) the actual wait-time, W; and (6) a vector of control variables, X.

$$E[S] = \alpha_0 + \alpha_1 C_{90} + \alpha_2 C_{70} + \alpha_3 C_{50} + \alpha_4 W + \beta' X \tag{3.13}$$

The coefficients α_1, α_2 and α_3 show the performance difference of Conditions C1 (i.e., announcing 90% overestimation of wait-time), C2 (i.e., announcing 70% overestimation of

wait-time), C3 (i.e., announcing 50% overestimation of wait-time) relative to the baseline condition, respectively. As shown in Table 3(a), the average wait-time satisfaction is higher in all scenarios with delay announcement than the no announcement scenario (i.e., average wait-time satisfaction under C0 = 3.52, under C1 = 4.19, under C2 = 4.74, and under C3 = 3.81). The highest average wait satisfaction achieved where 70% overestimation of wait-time is announced. By announcing 70% overestimation of wait-time we increased the average wait-time satisfaction from 3.52 to 4.74 (i.e., almost 35% increase).

Table 3.15. Models for Wait Satisfaction Survey Responses

	Dependent variable:			
	Wait-time Satisfaction			
	(1)	(2)	(3)	(4)
1. 90% overestimation of wait-time	0.673**	0.624***	0.631***	0.626***
	(0.272)	(0.029)	(0.023)	(0.040)
2. 70% overestimation of wait-time	1.219***	1.138***	1.141***	1.126***
	(0.124)	(0.370)	(0.357)	(0.368)
3. 50% overestimation of wait-time	0.294	0.350	0.365	0.367*
	(0.391)	(0.382)	(0.383)	(0.374)
4. Actual wait-time		−0.015***	−0.016***	−0.016***
		(0.001)	(0.001)	(0.001)
5. High acuity Indicator			−0.200	−0.180
			(0.127)	(0.150)
6. Age				−0.006
				(0.005)
7. Male Indicator				0.187
				(0.473)
8. First Visit Indicator				−0.161
				(0.229)
Constant	3.517***	6.030***	6.234***	6.546***
	(0.266)	(0.525)	(0.461)	(0.294)
Observations	373	373	373	373
Adjusted R^2	0.046	0.469	0.470	0.471
Pred. difference relative to baseline (90% overestimation of wait-time) (%)	19.41	10.35	10.12	9.56
Pred. difference relative to baseline (70% overestimation of wait-time) (%)	34.66	18.87	18.30	17.20
Pred. difference relative to baseline (50% overestimation of wait-time) (%)	8.36	5.80	5.85	5.61

Notes. Parentheses contain robust standard errors, clustered by shift of the day and day of the week. Predicted differences represent percent increase over baseline condition. Although we use OLS in our primary analysis to facilitate coefficient interpretation, we note that all reported results are similar when satisfaction is estimated with an ordinal logistic model. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively (two-tailed tests).

We estimate Equation (3.13) to access the impact of delay announcement on patient wait-time satisfaction. Table 3.15 presents different models for wait-time satisfaction. In

Table 3.15, column 1, we compare the mean wait-time satisfaction under each experimental condition. Both 90% and 70% overestimation of wait-time led to statistically large gains in wait-time satisfaction ($\alpha_1 = 0.626$, $p < 0.01$ and $\alpha_2 = 1.126$, $p < 0.01$). Even though announcing the average overestimation of wait-time did not significantly increase wait-time satisfaction ($\alpha_3 = 0.294$, $p = NS$) in the base model, after we control for the patients' actual wait-time, this effect become significant at the 10% level ($\alpha_3 = 0.367$, $p < 0.1$). This result remain robust in the fully specified model, which controls for priority level, age and gender.

No significant effect of priority level on wait-time satisfaction is identified ($\alpha_5 = -0.18$, $p = NS$). The largest VIF is 1.08, which falls well below the conventional threshold of 10, providing evidence that multicollinearity is not a concern (Wooldridge 2015).

While announcing the 90% overestimation of wait-time and the average wait-time resulted in 19% and 8% increase in average wait-time satisfaction, respectively, their positive impact was not as pronounced as the 70% overestimation of wait-time. We estimate Equation (3.13) to access the impact of delay announcement on patient wait-time satisfaction.

Both 90% and 70% overestimation of wait-time led to statistically large gains in wait-time satisfaction ($\alpha_1 = 0.626$, $p < 0.01$ and $\alpha_2 = 1.126$, $p < 0.01$). However, announcing the average overestimation of wait-time did not significantly increase wait-time satisfaction ($\alpha_3 = 0.367$, $p < 0.1$). This shows that announcing an overestimation of wait-time increases wait-time satisfaction more significantly than announcing the median (or mean) of wait-times. Hence, to maximize patients satisfaction, one need to overestimate the wait-times rather than sharing the mean estimated wait-times. In the fully specified model,

wait-time satisfaction was 17.2% higher than the baseline when 70% overestimation of wait-time and 9.56% above the baseline when announcing 90% overestimation of wait-time. Therefore, announcing 70% overestimation of wait-time has the highest predicted difference relative to the baseline.

How much to overestimate? The findings above suggest that under 70% overestimation of wait-time scenario, the highest predicted difference in wait-time satisfaction relative to baseline is achieved. In other words, when we overestimate the wait-time for around 70% of patients, we observe the highest average wait-time satisfaction in ED. How about the amount of overestimation? How much overestimation is optimal? If we know the actual wait-time, how many minutes should we add to that to maximize the satisfaction?

To address these questions, we classify the data based on how much the wait-time was overestimated or underestimated, using quartiles of wait-time gap, similar to Section 5.3. We use a linear regression model similar to the model in Equation 3.13 and find that except than Class 4, all other classes has significantly higher wait-time satisfaction relative to baseline.

As shown in Table 3.16, the mean wait-time satisfaction is significantly different than the baseline for Class 1 ($\alpha_1 = 0.999$ and $p < 0.01$), Class 2 ($\alpha_2 = 1.258$ and $p < 0.01$) and Class 1 ($\alpha_3 = 0.964$ and $p < 0.01$). However, the mean wait-time satisfaction is not significantly different than the baseline for Class 4. The predicted difference relative to baseline are 14.99%, 18.87% and 12.46% for Class 1-3, respectively. The highest wait-time satisfaction is achieved in Class 2, where patients' overestimation is between 40 to

Table 3.16. Models for Wait-time Satisfaction Survey Responses (Overestimation Classes)

	Dependent variable:
	Wait-time satisfaction
1. Class 1 (< 40 mins)	0.999^{***}
	(0.230)
2. Class 2 (40 mins << 80 mins)	1.258^{***}
	(0.002)
3. Class 3 (80 mins << 120 mins)	0.964^{***}
	(0.159)
4. Class 4 (>120 mins)	0.124
	(0.006)
5. Actual Wait-time	-0.017^{***}
	(0.425)
6. High acuity Indicator	−0.283
	(0.258)
7. Age	−0.004
	(0.008)
8. Male Indicator	0.130
	(0.338)
9 First Visit Indicator	−0.207
	(0.356)
Constant	6.666^{***}
	(0.322)
Observations	298
Adjusted R^2	0.429
Pred. difference relative to baseline (Class 1) (%)	14.99
Pred. difference relative to baseline (Class 2) (%)	18.87
Pred. difference relative to baseline (Class 3) (%)	12.46
Pred. difference relative to baseline (Class 4) (%)	1.86

Notes. Parentheses contain robust standard errors. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively (two-tailed tests).

80 minutes. This further supports the finding that overestimating too much may have a negative impact on patient satisfaction.

3.8. Robustness Tests

In this section, we explore the robustness of our main insights by discussing the alternatives for the reference point and the actual wait-time (see Appendix 5 for details).

3.8.1. Alternative for Reference Point

In previous sections, we assumed that patients take the announced wait-time as their expected wait-time since it is the only wait-related information provided to them. However, patients may have different expected wait-time when visiting the ED and they may not use the information provided by the system. One alternative for patients expected

wait-time, commonly used in delay announcement literature (Kőszegi and Rabin 2006, Yu et al. 2017), is the actual average wait time, which may be formed based on patients previous visits to ED. To test the robustness of our findings, for each patient, we computed the average wait-time of all patients who arrived on the same day of the week, same shift of the day and assigned to the same priority class as that patient over the last two years. We assume that patients take this average as their expected wait-time. We observed that our insights continue to hold with this alternative assumption of expected wait-time.

3.8.2. Alternative for Actual Wait Time

In previous sections, we assumed that patients compare their expected wait-time with how long they actually waited, when their wait-time satisfaction. However, Davis and Heineke (1998), for example, suggests that the perception of waiting time better predicts the satisfaction and the satisfaction initially defined as the difference between the expectation and perception not actual experience. However, perceived wait-time is a function of the actual wait-time and this is confirmed in our experiment by the high positive correlation ($\rho = 0.79$) between actual wait-time and perceived wait-time (see Table 3(b)). We used patients' reported perceived wait-time (i.e., question 1 on the survey) to recompute the wait-time gaps and tests the findings in Hypothesis 2. Our insights continue to hold when using perceived wait-times instead of the actual wait-times.

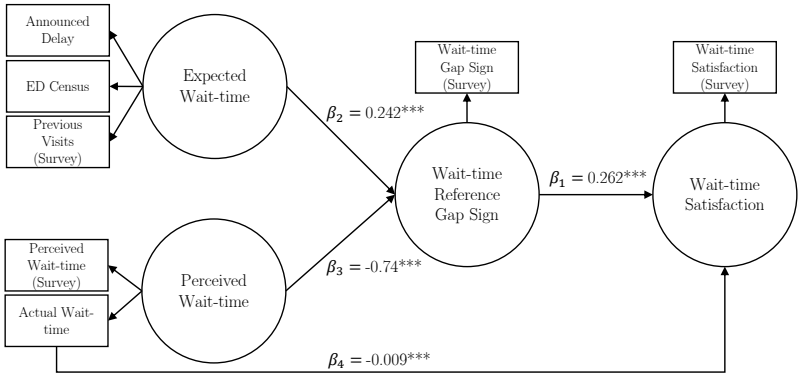
3.8.3. Latent Variable Analysis

Latent variables are used for theoretical framework or unobserved variables, such as personality characteristics, emotions, social status and etc and latent variables are widely

applied in many aspects (Bollen 2002). In this study, we model patients' expected wait-time, perceived wait-time and wait-time gap as latent variables. Figure 3.8 shows the integrative model of wait satisfaction and latent variables. To test the theory that delay announcement increases wait-time satisfaction, we used structural equation modeling to conduct a path analysis using the expected wait-time, perceived wait-time and wait-time gap measures.

Given our findings in previous sections and the conceptual model presented in Figure 3.2, we model the wait-time satisfaction again by taking the expected wait-time and perceived wait-time as latent variables. Figure 3.8 shows the wait-time satisfaction path analysis. This figure is drawn based on the relationships demonstrated in Figure 3.1. In this model, we assume that expected wait-times are formed based on announced wait-time, ED waiting room census and whether she has ever visited that ED. We model perceived wait-times based on actual wait-times and the answer to question 1 on the surveys. We also model wait-time reference gap sign and wait-time satisfaction based on the answers to question 2 and question 4 on surveys, respectively.

Figure 3.8. Wait-time Satisfaction Path Analysis



Coefficients in Figure 3.8 suggest that wait-time satisfaction is positively associated with wait-time reference gap sign ($\beta_1 = 0.262$, $p < 0.01$), which in turn is positively associated with expected wait-time ($\beta_2 = 0.242$, $p < 0.01$) and negatively associated with perceived wait-time ($\beta_3 = -0.74$, $p < 0.01$). Actual wait-time are also negatively associated with wait-time satisfaction as it was discussed in Hypothesis 3. Since wait-time satisfaction is positively associated wait-time gap sign ($\beta_1 = 0.262$, $p < 0.01$), this suggests that patients are more satisfied when the gap is positive, which further confirms the loss aversion behavior of the patients (i.e., Hypothesis 2B). The negative affect of actual wait-time, and therefore the perceived wait-time, on wait-time satisfaction is also picked up by this model as perceived wait-time is shown to be negatively associated with wait-time reference gap ($\beta_3 = -0.74$, $p < 0.01$). The model exhibits a good fit, with a high comparative fit index (CFI) 0.99 and a low root mean squared error of approximation (RMSEA) 0.039 ($p < 0.01$).

3.9. Discussion and Conclusion

Using two years of data from a hospital's ED, we developed a wait-time prediction model to estimate patients wait-time to see a doctor. We then conduct a field experiment to study the impact of informing patients of their wait-time on patient wait-time satisfaction. We find that patients are generally more satisfied with their wait-time when we inform them of their estimated wait-time. Higher patient satisfaction is a financial priority for hospitals, since the Centers for Medicare and Medicaid Services (CMS) is tying Medicare reimbursements to patients' assessment ratings. We also find that giving patients an overestimate of their wait-time make them more satisfied than giving them the

average estimated wait-time. In particular, since patients are loss-averse in their wait-time, overestimating their wait-time increases the probability of time gains than time losses. In other words, overestimating the wait-times allow the majority of patients to experience a "shorter than expected" wait-time and therefore become more satisfied with their waiting experience. We also find evidence that large wait-time gap may decrease the wait-time satisfaction, when patients experience time gains. The highest increase in wait-time satisfaction was obtained by adding 40 to 80 minutes to the expected wait-time (which corresponded with reporting the 70th percentile of the wait time distribution). This may be because of the negative impact of announcing long wait-times on satisfaction. These findings help system designers to design their delay announcement process to maximize patient satisfaction.

To quantify the impact of our findings, we compute the effect of announcing the wait-times under different delay announcement scenarios. We find that announcing 70% overestimation of wait-time is associated with almost 18% increase in wait-time satisfaction relative to the baseline (i.e., announcing nothing). We also observed a 6% and a 10% increase in wait-time satisfaction relative to the baseline when announcing the average and 90% overestimation of wait-time, respectively. This corresponds to more than one level increase in the seven-scale survey question. This positive impact was also observed on overall satisfaction and on the likelihood-to-recommend measured by Press Ganey survey questions, collected by the hospital. Even though separating out the effect of this intervention on Medicare payments is difficult, once we take into account the importance of patients rating on hospital's gains of Medicare payments, it becomes clear that the implications are substantial. If this findings are generalizable to other EDs, personalized

delay announcement would have significant practical financial and medical implications EDs across the country faced with unforeseen crowdedness and long wait-times for patients. It is also important to note that no statistically significant results found in terms of the possible effect of patients severity of conditions on their wait-time satisfaction.

This study contributes to the operations management literature on delay announcement and patient satisfaction in several ways. Our study is the among the first to conduct field experiment to examine the effect of announcing wait-times on satisfaction. Motivated by Batt and Terwiesch (2015) who called for more field experiment in EDs, we aimed to shed light on how providing information will influence service evaluation and alter behavior. It has not been clear what would happen if we start announcing delay and sharing news about the system's flow in EDs. If the news appears to be bad (e.g., a long wait-time), abandonment may increase and this is risky for the patients and economically bad for the hospital. Our field experiments helps determine how changes occur and what the net impact of the effects is. Lessons learned from such experiments serve to improve both ED management and the general understanding of human queuing behavior.

Implications for Practice. We suggest that patients in EDs should be provided with their wait-time while in the waiting room, estimated by an institution-specific wait-time predictor. Delay announcement is an inexpensive and easy-to-implement process to increase patients satisfaction in EDs and can be used along with other practices to reduce wait-times. To our knowledge, this is not currently in place at most EDs. Some EDs publish the average wait-times on their websites or on highway billboards but no personalized delays are provided for patients. We should also engineer the delay announcement such that for majority of patients the wait-time is overestimated. By minimizing the

number of patients with negative wait-time gap, we may be able to increase the total average wait-time satisfaction in the ED. Certainly, we need to be careful about how much we overestimate. Overestimating too much may cause patients to become dissatisfied or even leave the ED. The institution-specific wait-time predictor developed for the ED under study, is approved to be implemented and will be used by ED managers and nurses in the waiting room as part of their electronic medical record system to improve the communication and satisfaction in the ED.

There are also important positive and negative externalities for announcing the delays. Nurses and staff working in the waiting room area reported the waiting room to be calmer as a result of delay announcement. Calmer waiting rooms help nurses and staff to concentrate on their job rather than answering frequent wait-time related questions. The nurse who is trained to decide which patient gets the next available care space stated that "Before announcing the delays, many patients came to me to know about their wait-time and this disrupted my work. After announcing the delay this happens less frequently." Even though we did not observe a significant change in LWBS rate as a consequence of announcing wait-times, ED managers should still plan for providing patients clear explanation of the long waits and design protocols to help convince the patient not to leave before seeing a doctor.

Limitation and Future Research. There are several limitations in this study to take into account when considering the results. First, similar to many empirical studies, we note the threat of omitted variable bias. Even though it would have been helpful to control for more patient characteristics in our model, some of these data were protected information and some were not available. There are also other possible questions that we

could have asked in our surveys to help with understanding patients' evaluation process (e.g., about the emotions and affective responses). However, due to patients medical condition, we had some limits on the number of questions to ask. Since we wanted to capture patients' waiting assessment right after they experience it, we had to limit the number of questions to the level that best serves our purpose. Second, our study was done in a single hospital's ED. The fact that conducting such an experiment in several EDs, required developing an institution-specific wait-time predictor for each ED and building trust, which made it impossible for us to use another ED for implementation. Although the generalizability of our findings is limited because we studied only one ED, we believe our findings have strong theoretical backgrounds. Nevertheless, future research can examine a larger sample of EDs to study different delay announcement policies. Third, due to the limited number of days that this study was conducted over, situations beyond our control may have influenced some of the results (e.g., long waits due to arrival of a trauma case). It is important to note that this variability is part of a hospital's operation and we controlled for such effects by repeating the conditions that was influenced by extraordinary events.

Motivated by the findings of the current work, future work should include more analytical and empirical studies to find other delay announcement policies to maximize the patients satisfaction. In this study, we announced an estimated wait-time to all patients with no updates. ED managers may need to restrict the sharing of delay information with the patients to certain threshold on the estimated wait-time, to avoid possible negative consequences of announcing long announced wait-time on patient satisfaction. A dynamic policy may help managers find such a threshold. Also, we did not address the question of

whether we should provide an update on estimated wait-times if available and how frequently. Updating patients on what is happening in the system and revise the estimates provided, may have positive or negative impact on satisfaction and behavior, and needs to be carefully investigated in practice. Given that prior literature has found a variety of different mechanism such as operational transparency Buell and Norton (2011) to improve customers' perception of wait-time and assessment of their experience, the combination of these mechanisms with delay announcement to maximize satisfaction would also be a fruitful research area, especially in EDs, where satisfaction ties back to the service provider finance.

References

Abdellaoui, Mohammed, Emmanuel Kemel. 2013. Eliciting prospect theory when consequences are measured in time units:"time is not money". *Management Science* **60**(7) 1844–1859.

AHA. 2013. American hospital association. URL `http://www.aha.org/content/13/13-linkqualpaymnt.pdf`.

Ahn, Hyun-Soo, Izak Duenyas, Rachel Q Zhang. 1999. Optimal stochastic scheduling of a two-stage tandem queue with parallel servers. *Advances in Applied Probability* **31**(4) 1095–1117.

Ahn, Hyun-Soo, Rhonda Righter. 2006. Dynamic load balancing with flexible workers. *Advances in Applied Probability* 621–642.

Akşin, O Zeynep, Patrick T Harker. 2003. Capacity sizing in the presence of a common shared resource: Dimensioning an inbound call center. *European Journal of Operational Research* **147**(3) 464–483.

Aksin, Zeynep, Mor Armony, Vijay Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management* **16**(6) 665–688.

Akşin, Zeynep, Baris Ata, Seyed Morteza Emadi, Che-Lin Su. 2016. Impact of delay announcements in call centers: An empirical approach. *Operations Research* **65**(1) 242–265.

Allon, Gad, Achal Bassamboo, Itai Gurvich. 2011. "we will be right with you": Managing customer expectations with vague promises and cheap talk. *Operations research* **59**(6) 1382–1394.

Ammar, Achraf, Henri Pierreval, Sabeur Elkosentini. 2013. Workers assignment problems in manufacturing systems: A literature analysis. *Industrial Engineering and Systems Management (IESM), Proceedings of 2013 International Conference on*. IEEE, 1–7.

Andradóttir, Sigrún, Hayriye Ayhan. 2005. Throughput maximization for tandem lines with two stations and flexible servers. *Operations Research* **53**(3) 516–531.

Andradóttir, Sigrún, Hayriye Ayhan, Douglas G Down. 2007. Dynamic assignment of dedicated and flexible servers in tandem lines. *Probability in the Engineering and Informational Sciences* **21**(04) 497–538.

Andradóttir, Sigrún, Hayriye Ayhan, Eser Kırkızlar. 2012. Flexible servers in tandem lines with setup costs. *Queueing Systems* **70**(2) 165–186.

Andrews, Bruce, Henry Parsons. 1993. Establishing telephone-agent staffing levels through economic optimization. *Interfaces* **23**(2) 14–20.

Ang, Erjie, Sara Kwasnick, Mohsen Bayati, Erica L Plambeck, Michael Aratow. 2015. Accurate ed wait time prediction. *Manufaturing and Service Operations Management Forthcoming* .

Antonides, Gerrit, Peter C Verhoef, Marcel Van Aalst. 2002. Consumer perception and evaluation of waiting time: A field experiment. *Journal of Consumer Psychology* **12**(3) 193–202.

Armony, Mor, Nahum Shimkin, Ward Whitt. 2009. The impact of delay announcements in many-server queues with abandonment. *Operations Research* **57**(1) 66–81.

Ata, Barış, Mustafa H Tongarlak. 2013. On scheduling a multiclass queue with abandonments under general delay costs. *Queueing Systems* **74**(1) 65–104.

Atar, Rami, Chanit Giat, Nahum Shimkin. 2010. The cμ/θ rule for many-server queues with abandonment. *Operations Research* **58**(5) 1427–1439.

Atar, Rami, Chanit Giat, Nahum Shimkin. 2011. On the asymptotic optimality of the cμ/θ rule under ergodic cost. *Queueing Systems* **67**(2) 127–144.

Atar, Rami, Avi Mandelbaum, Martin I Reiman, et al. 2004. Scheduling a multi class queue with many exponential servers: Asymptotic optimality in heavy traffic. *The Annals of Applied Probability* **14**(3) 1084–1134.

Barberis, Nicholas C. 2013. Thirty years of prospect theory in economics: A review and assessment. *The Journal of Economic Perspectives* **27**(1) 173–195.

Batt, Robert J, Christian Terwiesch. 2015. Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science* **61**(1) 39–59.

Baumann, Hendrik, Werner Sandmann. 2017. Multi-server tandem queue with markovian arrival process, phase-type service times, and finite buffers. *European Journal of Operational Research* **256**(1) 187–195.

Bogle, Ariel. 2014. Is the infomercial dead? *Slate Magazine* .

Bolandifar, Ehsan, Nicole DeHoratius, Tava Olsen, Jennifer L Wiler. 2014. Modeling the behavior of patients who leave the emergency department without being seen by a physician. *Chicago Booth Research Paper* (12-14).

Bollen, Kenneth A. 2002. Latent variables in psychology and the social sciences. *Annual review of psychology* **53**(1) 605–634.

Boudreaux, Edwin D, Erin L O'Hea. 2004. Patient satisfaction in the emergency department: a review of the literature and implications for practice. *The Journal of emergency medicine* **26**(1) 13–26.

Breiman, Leo. 2001. Random forests. *Machine learning* **45**(1) 5–32.

Brown, Lawrence, Noah Gans, Avishai Mandelbaum, Anat Sakov, Haipeng Shen, Sergey Zeltyn, Linda Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association* **100**(469) 36–50.

Buell, Ryan W, Michael I Norton. 2011. The labor illusion: How operational transparency increases perceived value. *Management Science* **57**(9) 1564–1579.

Bühlmann, Peter, Torsten Hothorn. 2007. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science* 477–505.

Busby, John, Kristen Wisehart. 2016. How five changes can create happier customers and $100 million in revenue. *Marchex Institute* .

Campello, Fernanda, Armann Ingolfsson, Robert A Shumsky. 2016. Queueing models of case managers. *Management Science* **63**(3) 882–900.

Carmon, Ziv, Daniel Kahneman. 1996. The experienced utility of queuing: experience profiles and retrospective evaluations of simulated queues. *Durham, NC: Fuqua School, Duke University* .

Carmon, Ziv, J George Shanthikumar, Tali F Carmon. 1995. A psychological perspective on service segmentation models: The significance of accounting for consumers' perceptions of waiting and service. *Management Science* **41**(11) 1806–1815.

Chase, Richard B, Sriram Dasu. 2001. Want to perfect your company's service? use behavioral science. *Harvard business review* **79**(6) 78–84.

CMSb. 2017. Center of medicare and medicaid services. URL `https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/Hospital_VBPurchasing_Fact_Sheet_ICN907664.pdf`.

Cohen, Elizabeth L, Holley A Wilkin, Michael Tannebaum, Melissa S Plew, Leon L Haley Jr. 2013. When patients are impatient: the communication strategies utilized by emergency department employees to manage patients frustrated by wait times. *Health communication* **28**(3) 275–285.

Crawford, Vincent P, Juanjuan Meng. 2011. New york city cab drivers' labor supply revisited: Reference-dependent preferences with rationalexpectations targets for hours and income. *The American Economic Review* **101**(5) 1912–1932.

Davis, Mark M, Janelle Heineke. 1998. How disconfirmation, perception and actual waiting times impact customer satisfaction. *international Journal of Service industry Management* **9**(1) 64–73.

Dobson, Gregory, Tolga Tezcan, Vera Tilson. 2013. Optimal workflow decisions for investigators in systems with interruptions. *Management Science* **59**(5) 1125–1141.

Dong, Jing, Elad Yom-Tov, Galit B Yom-Tov. 2015. The impact of delay announcements on hospital network coordination and waiting times. Tech. rep., Working Paper.

Edwards, Corwin D. 1968. The meaning of quality. *Quality progress* **1**(10) 36–39.

Farrar, Timothy Martin. 1993. Optimal use of an extra server in a two station tandem queueing network. *IEEE Transactions on Automatic Control* **38**(8) 1296–1299.

Fitzsimmons, James A, Mona J Fitzsimmons, Sanjeev Bordoloi. 2006. *Service management: Operations, strategy, and information technology*. McGraw-Hill New York.

Forbes, Silke, Mara Lederman, Zhe Yuan. 2017. Do airlines pad their schedules? .

Fredrickson, Barbara L, Daniel Kahneman. 1993. Duration neglect in retrospective evaluations of affective episodes. *Journal of personality and social psychology* **65**(1) 45.

Friedman, Jerome H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 1189–1232.

Gans, Noah, Ger Koole, Avishai Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* **5**(2) 79–141.

GAO. 2011. United states government accountability office. URL `http://www.gao.gov/assets/100/97416.pdf`.

Garnett, Ofer, Avi Mandelbaum, M Reiman. 2002. Designing a call center with impatient customers. *Manufacturing & Service Operations Management* **4**(3) 208–227.

Garvin, David A. 1984. What does "hltoduct quality" really mean? *Sloan management review* 25.

Göransson, Katarina E, Anette von Rosen. 2010. Patient experience of the triage encounter in a swedish emergency department. *International emergency nursing* **18**(1) 36–40.

Guo, Pengfei, Paul Zipkin. 2007. Analysis and comparison of queues with different levels of delay information. *Management Science* **53**(6) 962–970.

Ha, Albert Y. 1997a. Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Science* **43**(8) 1093–1103.

Ha, Albert Y. 1997b. Optimal dynamic scheduling policy for a make-to-stock production system. *Operations Research* **45**(1) 42–53.

Hajek, Bruce. 1984. Optimal control of two interacting service stations. *IEEE transactions on automatic control* **29**(6) 491–499.

Hassin, Refael. 1986. Consumer information in markets with random product quality: The case of queues and balking. *Econometrica: Journal of the Econometric Society* 1185–1195.

Hassin, Refael. 2007. Information and uncertainty in a queuing system. *Probability in the Engineering and Informational Sciences* **21**(03) 361–380.

Hastie, T, R Tibshirani, J Friedman. 2009. *The elements of statistical learning 2nd edition*. New York: Springer.

Heskett, James L, Thomas O Jones, Gary W Loveman, W Earl Sasser, Leonard A Schlesinger, et al. 1994. Putting the service-profit chain to work. *Harvard business review* **72**(2) 164–174.

Hofner, Benjamin, Andreas Mayr, Nikolay Robinzonov, Matthias Schmid. 2014. Model-based boosting in r: a hands-on tutorial using the r package mboost. *Computational statistics* **29**(1-2) 3–35.

Hopp, Wallace J, Seyed MR Iravani, Biying Shou. 2005. Serial agile production systems with automation. *Operations research* **53**(5) 852–866.

Hopp, Wallace J, MARK P OYEN. 2004. Agile workforce evaluation: a framework for cross-training and coordination. *Iie Transactions* **36**(10) 919–940.

Hornik, Jacob. 1984. Subjective vs. objective time measures: A note on the perception of time in consumer behavior. *Journal of consumer research* **11**(1) 615–618.

Hu, Ming, Yang Li, Jianfu Wang. 2017. Efficient ignorance: Information heterogeneity in a queue. *Management Science* .

Hui, Michael K, David K Tse. 1996. What to tell consumers in waits of different lengths: An integrative model of service evaluation. *The Journal of Marketing* 81–90.

Iravani, Seyed MR, Morton J. M. Posner, John A. Buzacott. 1997. A two-stage tandem queue attended by a moving server with holding and switching costs. *Queueing Systems* **26**(3-4) 203–228.

Javidi, Tara, Nah-Oak Song, Demosthenis Teneketzis. 2001. Expected makespan minimization on identical machines in two interconnected queues. *Probability in the Engineering and Informational Sciences* **15**(04) 409–443.

Johnson, Mary Beth, Edward M Castillo, James Harley, David A Guss. 2012. Impact of patient and family communication in a pediatric emergency department on likelihood to recommend. *Pediatric emergency care* **28**(3) 243–246.

Johnson, Rosser. 2013. The emergence of the infomercial in new zealand 1993–1997. *The Political Economy of Communication* **1**(1).

Jouini, Oualid, Zeynep Akşin, Yves Dallery. 2011. Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management* **13**(4) 534–548.

Jouini, Oualid, Yves Dallery, Zeynep Akşin. 2009. Queueing models for full-flexible multi-class call centers with real-time anticipated delays. *International Journal of Production Economics* **120**(2) 389–399.

Kahneman, Daniel, Barbara L Fredrickson, Charles A Schreiber, Donald A Redelmeier. 1993. When more pain is preferred to less: Adding a better end. *Psychological science* **4**(6) 401–405.

Kahneman, Daniel, D Kahneman, A Tversky, et al. 2003. Experienced utility and objective happiness: A moment-based approach. *The psychology of economic decisions* **1** 187–208.

Kahneman, Daniel, Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society* 263–291.

Kalekar, Prajakta S. 2004. Time series forecasting using holt-winters exponential smoothing. *Kanwal Rekhi School of Information Technology* **4329008** 1–13.

Kim, Jeunghyun, Ramandeep S Randhawa, Amy R Ward. 2016. Dynamic scheduling in a many-server multi-class system: The role of customer impatience in large systems. *Available at SSRN* .

Kim, Jeunghyun, Amy R Ward. 2013. Dynamic scheduling of a gi/gi/1+ gi queue with multiple customer classes. *Queueing Systems* **75**(2-4) 339–384.

Kırkızlar, Eser, Sigrún Andradóttir, Hayriye Ayhan. 2010. Robustness of efficient server assignment policies to service time distributions in finite-buffered lines. *Naval Research Logistics (NRL)* **57**(6) 563–582.

Kırkızlar, Eser, Sigrún Andradóttir, Hayriye Ayhan. 2014. Profit maximization in flexible serial queueing networks. *Queueing Systems* 1–38.

Kőszegi, Botond, Matthew Rabin. 2006. A model of reference-dependent preferences. *The Quarterly Journal of Economics* **121**(4) 1133–1165.

Krishel, Scott, Larry J Baraff. 1993. Effect of emergency department information on patient satisfaction. *Annals of emergency medicine* **22**(3) 568–572.

Lippman, Steven A. 1975. Applying a new device in the optimization of exponential queuing systems. *Operations Research* **23**(4) 687–710.

Maister, David H. 1984. *The psychology of waiting lines*. Harvard Business School.

Mowen, John C, Jane W Licata, Jeannie McPhail. 1993. Waiting in the emergency room: how to improve patient satisfaction. *Marketing Health Services* **13**(2) 26.

Norman, Donald A. 2009. Designing waits that work. *MIT Sloan Management Review* **50**(4) 23.

Osuna, Edgar Elias. 1985. The psychological cost of waiting. *Journal of Mathematical Psychology* **29**(1) 82–105.

Pandelis, Dimitrios G. 2007. Optimal use of excess capacity in two interconnected queues. *Mathematical Methods of Operations Research* **65**(1) 179–192.

Pandelis, Dimitrios G. 2008. Optimal stochastic scheduling of two interconnected queues with varying service rates. *Operations Research Letters* **36**(4) 492–495.

Pandelis, Dimitrios G, Demosthenis Teneketzis. 1994. Optimal multiserver stochastic scheduling of two interconnected priority queues. *Advances in Applied Probability* 258–279.

Phung-Duc, Tuan, Ken'ichi Kawanishi. 2014. Performance analysis of call centers with abandonment, retrial and after-call work. *Performance Evaluation* **80** 43–62.

Plambeck, Erica, Mohsen Bayati, Erjie Ang, Sara Kwasnick, Mike Aratow, et al. 2014. Forecasting emergency department wait times. Tech. rep.

Pruyn, A Th H, Ale Smidts. 1993. Customers' evaluations of queues: Three exploratory studies. *ACR European Advances* .

Puterman, Martin L. 1990. Markov decision processes. *Handbooks in operations research and management science* **2** 331–434.

Righter, Rhonda. 2000. Expulsion and scheduling control for multiclass queues with heterogeneous servers. *Queueing Systems* **34**(1-4) 289–300.

Rosberg, Zvi, P Varaiya, J Walrand. 1982. Optimal control of service in tandem queues. *IEEE Transactions on Automatic Control* **27**(3) 600–610.

Salch, Alexandre, J-P Gayon, Pierre Lemaire. 2013. Optimal static priority rules for stochastic scheduling with impatience. *Operations Research Letters* **41**(1) 81–85.

Salway, RJ, R Valenzuela, JM Shoenberger, WK Mallon, A Viccellio. 2017. Emergency department (ed) overcrowding: evidence-based answers to frequently asked questions. *Revista Médica Clínica Las Condes* **28**(2) 213–219.

Schiefermayr, Klaus, Josef Weichbold. 2005. A complete solution for the optimal stochastic scheduling of a two-stage tandem queue with two flexible servers. *Journal of applied probability* 778–796.

Schmitt, Bernd H, France Leclerc. 2002. The value of time in the context of waiting and delays. *Consumer Value*. Routledge, 45–58.

Schreiber, Charles A, Daniel Kahneman. 2000. Determinants of the remembered utility of aversive sounds. *Journal of Experimental Psychology: General* **129**(1) 27.

Sennott, Linn I. 1996. The convergence of value iteration in average cost markov decision chains. *Operations research letters* **19**(1) 11–16.

Sennott, Linn I, Mark P Van Oyen, Seyed MR Iravani. 2006. Optimal dynamic assignment of a flexible worker on an open production line with specialists. *European Journal of Operational Research* **170**(2) 541–566.

Shah, Shital, Anay Patel, Dino P Rumoro, Samuel Hohmann, Francis Fullam. 2015. Managing patient expectations at emergency department triage. *Patient Experience Journal* **2**(2) 31–44.

Soman, Dilip, Mengze Shi. 2003. Virtual progress: The effect of path characteristics on perceptions of progress and choice. *Management Science* **49**(9) 1229–1250.

Spiegler, Ran. 2011. *Bounded rationality and industrial organization*. Oxford University Press.

Sun, Yan, Kiok Liang Teow, Bee Hoon Heng, Chee Kheong Ooi, Seow Yian Tay. 2012. Real-time prediction of waiting time in the emergency department, using quantile regression. *Annals of emergency medicine* **60**(3) 299–308.

Tang, Jiashan, Yiqiang Q Zhao. 2008. Stationary tail asymptotics of a tandem queue with feedback. *Annals of Operations Research* **160**(1) 173–189.

Taylor, Shirley, Gordon Fullerton. 1999. Perceptions management of the wait experience. *Handbook of services marketing and management* 171.

Thaler, Richard. 1985. Mental accounting and consumer choice. *Marketing science* **4**(3) 199–214.

Thompson, David A, Paul R Yarnold, Stephen L Adams, Alan B Spacone. 1996a. How accurate are waiting time perceptions of patients in the emergency department? *Annals of emergency medicine* **28**(6) 652–656.

Thompson, David A, Paul R Yarnold, Diana R Williams, Stephen L Adams. 1996b. Effects of actual waiting time, perceived waiting time, information delivery, and expressive quality on patient satisfaction in the emergency department. *Annals of emergency medicine* **28**(6) 657–665.

Tom, Gail, Scott Lucey. 1995. Waiting time delays and customer satisfaction in supermarkets. *Journal of Services Marketing* **9**(5) 20–29.

Tom, Gail, Scott Lucey. 1997. A field study investigating the effect of waiting time on customer satisfaction. *The Journal of psychology* **131**(6) 655–660.

Varey, Carol, Daniel Kahneman. 1992. Experiences extended across time: Evaluation of moments and episodes. *Journal of Behavioral Decision Making* **5**(3) 169–185.

Veatch, Michael H, Lawrence M Wein. 1994. Optimal control of a two-station tandem production/inventory system. *Operations Research* **42**(2) 337–350.

Ward, Amy R, Peter W Glynn. 2003. A diffusion approximation for a markovian queue with reneging. *Queueing Systems* **43**(1) 103–128.

Weichbold, Josef, Klaus Schiefermayr. 2006. The optimal control of a general tandem queue. *Probability in the Engineering and Informational Sciences* **20**(02) 307–327.

Welch, Shari Jule. 2009. Twenty years of patient satisfaction research applied to the emergency department: a qualitative review. *American Journal of Medical Quality* .

White, P, DP Bahner, A Fishman, J Glinski, S Khandelwal, D Post, J Gatto, S Damewood. 2005. The effect of information delivery on patient satisfaction in the emergency department. *Annals of Emergency Medicine* **46**(3) 120–121.

Whitt, Ward. 1999. Improving service by informing customers about anticipated delays. *Management science* **45**(2) 192–207.

Wooldridge, Jeffrey M. 2015. *Introductory econometrics: A modern approach*. Nelson Education.

Wu, Cheng-Hung, Mark E Lewis, Michael Veatch. 2006. Dynamic allocation of reconfigurable resources ina two-stage tandem queueing system with reliability considerations. *Automatic Control, IEEE Transactions on* **51**(2) 309–314.

Xie, Bin, Sabrina Youash. 2011. The effects of publishing emergency department wait time on patient utilization patterns in a community with two emergency department sites: a retrospective, quasi-experiment design. *International journal of emergency medicine* **4**(1) 29.

Yang, Liu, Francis De Vericourt, Peng Sun. 2013. Time-based competition with benchmark effects. *Manufacturing & Service Operations Management* **16**(1) 119–132.

Yu, Qiuping, Gad Allon, Achal Bassamboo. 2016. How do delay announcements shape customer behavior? an empirical study. *Management Science* **63**(1) 1–20.

Yu, Qiuping, Gad Allon, Achal Bassamboo. 2017. The reference effect of delay announcements: A field experiment .

Yu, Qiuping, Gad Allon, Achal Bassamboo, Seyed Iravani. 2015. Managing customer expectations and priorities in service systems. *Available at SSRN* .

Zayas-Cabán, Gabriel, Jingui Xie, Linda V Green, Mark E Lewis. 2016. Dynamic control of a tandem system with abandonments. *Queueing Systems* **84**(3-4) 279–293.

Zeltyn, Sergey, Avishai Mandelbaum. 2005. Call centers with impatient customers: many-server asymptotics of the m/m/n+ g queue. *Queueing Systems* **51**(3-4) 361–402.

APPENDIX A

Appendix of Chapter 1: Proof of analytical results

Proof of Proposition 1.1. Given the job arrival rate at Stage 1 (i.e., λ), and given that jobs that leave Stage 1 arrive at Stage 2 with probability 1, the average arrival rate at both Stages 1 and 2 are λ. Since jobs completed at Stage 2 arrive at Stage 3 with probability p, the average arrival rate at Stage 3 is $p\lambda$. We use λ_i to denote the arrival rate at Stage i. Let ρ_i represent the utilization at Stage i. In order to have a stable system, we need to ensure that there exists a server control policy such that $\rho_i = \lambda_i/\mu_i < 1,\ \forall i$.

The system cannot be stable if $\exists i$ such that $\rho_i \geq 1$ (i.e., $\mu_i \leq \lambda,\ i = 1, 2$ and $\mu_3 \leq p\lambda$). When $\mu_i > \lambda,\ i = 1, 2$ and $\mu_3 > p\lambda$ (i.e., $\rho_i < 1$, $\forall i$), we show that the system is stable if the last condition in Proposition 1.1 holds.

Let $k \in (0,\ 1)$ denote the percentage of time that Server S1 works at Stage 1. So the average processing rate at Stage 1 is $k\mu_1$ and the processing rate is at most $(1-k)\mu_3$ at Stage 3. We show that if the condition in Proposition 1.1 holds, then there exists a solution for k that makes the system stable.

To make Stages 1 and 3 stable we require:

$$k\mu_1 > \lambda$$

$$(1-k)\mu_3 > p\lambda$$

Since $\mu_i > 0$, $\forall i$, we have

$$k > \frac{\lambda}{\mu_1}$$

$$k < 1 - \frac{p\lambda}{\mu_3}$$

To ensure there exists one solution for k, we need

$$\frac{\lambda}{\mu_1} < 1 - \frac{p\lambda}{\mu_3}$$

It is equivalent to

$$\frac{\mu_1\mu_3}{p\mu_1+\mu_3} > \lambda$$

Thus, if all conditions in Proposition 1.1 holds, there exists a solution for k and the system is stable. □

Proof of Theorem 1.1. We first prove that there exists an average-cost optimal stationary policy for the MDP that has a constant average cost. To prove this, we need to show that (i) the Markov chain corresponding to the class of non-idling policy that we consider in the MDP model is irreducible; (ii) there exists a stationary policy under which the Markov chain is positive recurrent; and (iii) the system has finite mean queue lengths.

Consider state $\mathbf{0} = (0, 0, 0)$. Since the arrivals are Poisson, state $\mathbf{0}$ can reach any state $\mathbf{n}$ where $n_1 > 0$. State $\mathbf{n}$ can reach state $(\mathbf{n} - \mathbf{e^1} + \mathbf{e^2})$ where $n_1 > 0, n_2 \geq 0$. Since Server S2 never idles when $n_2 > 0$, state $(\mathbf{n})$ with $n_2 > 0$ can reach $(\mathbf{n} - \mathbf{e^2} + \mathbf{e^3})$ with a positive probability. On the other hand, since any policy we consider does not idle Server S1 at state $(\mathbf{n})$ when she is at Stage 3 with $n_3 > 0$, and does not idle state $\mathbf{n}$ when she is at Stage 1 with $n_1 > 0$, then $\mathbf{0}$ is reachable by any state $\mathbf{n}$.

Therefore, the Markov chain corresponding to the class of non-idling policy we consider in the MDP model is irreducible. We can use results from Puterman Puterman (1990) to prove the existence of an average-cost optimal stationary policy. First, we need to show that when conditions in Proposition 1.1 hold, there exists a stationary policy under which the Markov chain is positive recurrent and the system has finite mean queue lengths.

Consider policy γ that assigns Server S1 in the following way: Server S1 works at Stage 1 and switches to Stage 3 when there is at least 1 job at Stage 3. After switching to Stage 3, she processes the job then switches back to Stage 1. Since we have $\frac{\mu_1\mu_3}{p\mu_1+\mu_3} > \lambda$, using the same argument as the proof of Proposition 1.1, the utilization of Server S1 is less than 1 and the system is stable under policy γ and has finite mean queue lengths. Moreover, under policy γ every state can reach state $(\mathbf{0})$ within a finite time, the Markov chain is positive recurrent. Let g represent the average cost induced by policy γ. We need to show that set $A = \{s \in S : C(s,a) < g \text{ for some } a \in A_S\}$ is not empty and finite, where $C(s,a) = (n_1 + n_2 + n_3)/\Lambda$. First, at state $s_1 = (\mathbf{0})$ we have $C(s_1, a) = 0$. Since $g > 0$ and $s_1 \in S$, set A is not empty. Second, $C(s,a)$ is increasing in n_1, n_2, and n_3, and since g is finite, there are only finite number of states that satisfy $C(s,a) < g$. Therefore, set A is finite, and Theorem 8.10.9 in Puterman (1990) holds. Thus, we showed that there exists a stationary policy under which the Markov chain is positive recurrent. Hence, the MDP has an average-cost optimal stationary policy.

We now prove that the value iteration algorithm converges. Proposition 4.3 of Sennott (1996) states that if there exists a stationary policy inducing an irreducible and positive recurrent Markov chain with finite average cost g, and if there exists $\epsilon > 0$ such that $D = \{s \in S : \text{there exists } a \text{ such that } C(s,a) < g+\epsilon\}$ is finite, then the value iteration algorithm converges. We have a stationary policy γ such that the induced Markov chain is irreducible and positive recurrent, and has finite average cost g. Since $C(s,a)$ is increasing in n_1, n_2, and n_3 and since $g + \epsilon$ is finite, like set A, set D is also finite. Therefore, conditions of Proposition 4.3 of Sennott (1996) is satisfied and the value iteration algorithm converges.

□

Proof of Proposition 1.2(a). To prove that **C1**, **C2**, and **M1-M3** are preserved by operator $\mathbf{T}$, we first show **C1** and **C2** holds for $\mathbf{T}v(\mathbf{n})$, then we drive a general inequalities to prove **M1** through **M3**. For any $v \in \Upsilon$, we have

$$\begin{aligned}\mathbf{T}v(\mathbf{n}) =& \frac{1}{\Lambda}\Big\{n_1 + n_2 + n_3 + \lambda v(\mathbf{n}+\mathbf{e}^1) + \mu_2 B(\mathbf{n}) \\ &+ \min\Big\{A_I(\mathbf{n}), A_{P1}(\mathbf{n}), A_{P3}(\mathbf{n})\Big\}\Big\},\end{aligned} \tag{A.1}$$

$$B(\mathbf{n}) = \begin{cases} v(\mathbf{n}) & : \text{ if } n_2 = 0 \\ pv(\mathbf{n}-\mathbf{e}^2+\mathbf{e}^3) + (1-p)v(\mathbf{n}-\mathbf{e}^2) & : \text{ if } n_2 > 0 \end{cases}$$

$$A_I(\mathbf{n}) = (\mu_1 + \mu_3)v(\mathbf{n})$$

$$A_{P1}(\mathbf{n}) = \begin{cases} \mu_1 v(\mathbf{n}) + \mu_3 v(\mathbf{n}) & : \text{ if } n_1 = 0 \\ \mu_1 v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) + \mu_3 v(\mathbf{n}) & : \text{ if } n_1 > 0 \end{cases}$$

$$A_{P3}(\mathbf{n}) = \begin{cases} \mu_3 v(\mathbf{n}) + \mu_1 v(\mathbf{n}) & : \text{ if } n_3 = 0 \\ \mu_3 v(\mathbf{n}-\mathbf{e}^3) + \mu_1 v(\mathbf{n}) & : \text{ if } n_3 > 0 \end{cases}$$

To prove **C1**, first, we consider **C1** as follows:

C1: $v(\mathbf{n})$ is nondecreasing in $n_1 \geq 0$, $n_2 \geq 0$ and $n_3 \geq 0$.

We want to show if $v \in \Upsilon$, then $\mathbf{T}v \in \Upsilon$. We show property **C1** is preserved by operator $\mathbf{T}$, i.e., $\mathbf{T}v(\mathbf{n})$ is nondecreasing in $n_1 \geq 0$, as an example and the rest can be shown similarly. There are 8 possible cases for **C1** in terms of $n_i > 0$ or $n_i = 0$,

$i \in \{1,2,3\}$. First, consider the case ($n_1 > 0$, $n_2 > 0$, $n_3 > 0$). Define

$$AC_a(\mathbf{n}) = \begin{cases} (\mu_1 + \mu_3)v(\mathbf{n}), & a = I \\ \mu_1 v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2) + \mu_3 v(\mathbf{n}), & a = P1 \\ \mu_1 v(\mathbf{n}) + \mu_3 v(\mathbf{n} - \mathbf{e}^3), & a = P3 \end{cases}$$

Next, we define $AC_{a^*_{(\mathbf{n})}}(\mathbf{n}) = \min_a AC_a(\mathbf{n})$ as the value of $AC_a(\mathbf{n})$ under action $a^*_{(\mathbf{n})} \in \{I, P1, P3\}$, where $a^*_{(\mathbf{n})}$ is the optimal action at state $(\mathbf{n})$. More generally, we define $AC_{a^*_{(\mathbf{n}-\mathbf{e}^i)}}(\mathbf{n} - \mathbf{e}^j)$ $(i, j = 1, 2, 3)$ as the value of $AC_a(\mathbf{n} - \mathbf{e}^j)$ under action $a^*_{(\mathbf{n}-\mathbf{e}^i)}$, where $a^*_{(\mathbf{n}-\mathbf{e}^i)}$ is the optimal action at state $(\mathbf{n} - \mathbf{e}^i)$. Therefore, we can rewrite Equation (A.1) as follows:

$$\mathbf{T}v(\mathbf{n}) = \frac{1}{\Lambda}\Big\{ n_1 + n_2 + n_3 + \lambda v(\mathbf{n} + \mathbf{e}^1) + \mu_2 \left[pv(\mathbf{n} - \mathbf{e}^2 + \mathbf{e}^3) + (1-p)v(\mathbf{n} - \mathbf{e}^2) \right] + AC_{a^*_{(\mathbf{n})}}(\mathbf{n}) \Big\}.$$

Since $v \in \Upsilon$, $v(\mathbf{n})$ is nondecreasing in n_1, for $n_1 \geq 0, n_2 \geq 0, n_3 \geq 0$. Thus, **C1** immediately holds for the first three $v(.)$ terms of right hand sides of $\mathbf{T}v(\mathbf{n})$. Therefore, we only need to show that **C1** holds for $AC_{a^*_{(\mathbf{n})}}(\mathbf{n})$.

If the optimal action at state $(\mathbf{n} + \mathbf{e}^1)$ is idling (i.e., $a^*_{(\mathbf{n}+\mathbf{e}^1)} = I$), then

$$AC_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n} + \mathbf{e}^1) = (\mu_1 + \mu_3)v(\mathbf{n} + \mathbf{e}^1)$$

$$AC_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n}) = (\mu_1 + \mu_3)v(\mathbf{n})$$

If the optimal action at state $(\mathbf{n}+\mathbf{e}^1)$ is working at Station 1 (i.e., $a^*_{(\mathbf{n}+\mathbf{e}^1)} = P1$), then

$$AC_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n}+\mathbf{e}^1) = \mu_1 v(\mathbf{n}+\mathbf{e}^{\mathbf{2}}) + \mu_3 v(\mathbf{n}+\mathbf{e}^1)$$

$$AC_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n}) = \mu_1 v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) + \mu_3 v(\mathbf{n})$$

If the optimal action at state $(\mathbf{n}+\mathbf{e}^1)$ is working at Station 3 (i.e., $a^*_{(\mathbf{n}+\mathbf{e}^1)} = P3$), then

$$AC_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n}+\mathbf{e}^1) = \mu_1 v(\mathbf{n}+\mathbf{e}^1) + \mu_3 v(\mathbf{n}+\mathbf{e}^1-\mathbf{e}^3)$$

$$AC_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n}) = \mu_1 v(\mathbf{n}) + \mu_3 v(\mathbf{n}-\mathbf{e}^3)$$

According to **C1**, we have $AC_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n}+\mathbf{e}^1) \geq AC_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n})$. On the other hand, $AC_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n})$ may not select the optimal action at state $(\mathbf{n})$ (i.e., $a^*_{(\mathbf{n}+\mathbf{e}^1)}$ may not equal $a^*_{(\mathbf{n})}$) but $AC_{a^*_{(\mathbf{n})}}(\mathbf{n})$ always does. Thus, we have $AC_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n}) \geq AC_{a^*_{(\mathbf{n})}}(\mathbf{n})$. Therefore, we get $AC_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n}+\mathbf{e}^1) \geq AC_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n}) \geq AC_{a^*_{(\mathbf{n})}}(\mathbf{n})$, i.e., $AC_{a^*_{(\mathbf{n})}}(\mathbf{n})$ is nondecreasing in n_1. Thus, we prove property **C1** holds. Similarly, we can easily show that **C1** is preserved by operator **T** in all 8 cases.

To prove **C2**, first, we write **C2** as follows:

$v(\mathbf{n}) \geq v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)$, for $n_1 > 0, n_2 \geq 0, n_3 \geq 0$.

There are 8 possible cases for **C2** in terms of $n_1 > 1$ or $n_1 = 1$ and $n_i > 0$ or $n_i = 0$, $i \in \{2, 3\}$. First, consider the case ($n_1 > 1$, $n_2 > 0$, $n_3 > 0$). Define

$$AC_a(\mathbf{n}) = \begin{cases} (\mu_1+\mu_3)v(\mathbf{n}), & a = I \\ \mu_1 v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) + \mu_3 v(\mathbf{n}), & a = P1 \\ \mu_1 v(\mathbf{n}+\mathbf{e}^3) + \mu_3 v(\mathbf{n}), & a = P3 \end{cases}$$

We define $AC_{a^*_{(\mathbf{n})}}(\mathbf{n})$ and $AC_{a^*_{(\mathbf{n}-\mathbf{e}^i)}}(\mathbf{n}-\mathbf{e}^j)$ $(i,j=1,2,3)$ as it was previously defined. we can rewrite Equation A.1 as follows:

$$\mathbf{T}v(\mathbf{n}) = \frac{1}{\Lambda}\Big\{n_1+n_2+n_3+\lambda v(\mathbf{n}+\mathbf{e}^1) + \mu_2\left[pv(\mathbf{n}-\mathbf{e}^2+\mathbf{e}^3)+(1-p)v(\mathbf{n}-\mathbf{e}^2)\right] + AC_{a^*_{(\mathbf{n})}}(\mathbf{n})\Big\}.$$

Since $v\in\Upsilon$, $v(\mathbf{n})\geq v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)$, for $n_1\geq 1, n_2\geq 0, n_3\geq 0$. Thus, **C2** immediately holds for the first three $v(.)$ terms of right hand sides of $\mathbf{T}v(\mathbf{n})$. Therefore, we only need to show that **C2** holds for $AC_{a^*_{(\mathbf{n})}}(\mathbf{n})$.

If the optimal action at state $(\mathbf{n})$ is idling (i.e., $a^*_{(\mathbf{n})}=I$), then

$$AC_{a^*_{(\mathbf{n})}}(\mathbf{n}) = (\mu_1+\mu_3)v(\mathbf{n})$$

$$AC_{a^*_{(\mathbf{n})}}(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) = (\mu_1+\mu_3)v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)$$

If the optimal action at state $(\mathbf{n})$ is working at Station 1 (i.e., $a^*_{(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)}=P1$), then

$$AC_{a^*_{(\mathbf{n})}}(\mathbf{n}) = \mu_1 v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)+\mu_3 v(\mathbf{n})$$

$$AC_{a^*_{(\mathbf{n})}}(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) = \mu_1 v(\mathbf{n}-2\mathbf{e}^1+2\mathbf{e}^2)+\mu_3 v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)$$

If the optimal action at state $(\mathbf{n})$ is working at Station 3 (i.e., $a^*_{(\mathbf{n})}=P3$), then

$$AC_{a^*_{(\mathbf{n})}}(\mathbf{n}) = \mu_1 v(\mathbf{n})+\mu_3 v(\mathbf{n}-\mathbf{e}^3)$$

$$AC_{a^*_{(\mathbf{n})}}(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) = \mu_1 v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)+\mu_3 v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2-\mathbf{e}^3)$$

According to **C2**, we have $AC_{a^*_{(\mathbf{n})}}(\mathbf{n}) \geq AC_{a^*_{(\mathbf{n})}}(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)$. On the other hand, $AC_{a^*_{(\mathbf{n})}}(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)$ may not select the optimal action at state $(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)$ (i.e., $a^*_{(\mathbf{n})}$ may not equal $a^*_{(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)}$) but $AC_{a^*_{(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)}}(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)$ always does. Thus, we have $AC_{a^*_{(\mathbf{n})}}(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) \geq AC_{a^*_{(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)}}(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)$. Therefore, we get $AC_{a^*_{(\mathbf{n})}}(\mathbf{n}) \geq AC_{a^*_{(\mathbf{n})}}(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) \geq AC_{a^*_{(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)}}(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)$. Thus, we prove property **C2**, i.e., $v(\mathbf{n}) - v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) \geq 0$, which implies $v(\mathbf{n}) \geq v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)$. Similarly, it is tedious but easy to verify that **C2** is preserved by operator **T** in all 8 cases.

To prove **M1** to **M3** are preserved by operator **T**, we first define

$$A = v(\mathbf{n}'-\mathbf{e}^3) - v(\mathbf{n}')$$

$$B = v(\mathbf{n}') - v(\mathbf{n}'-\mathbf{e}^1+\mathbf{e}^2)$$

$$C = v(\mathbf{n}-\mathbf{e}^3) - v(\mathbf{n})$$

$$D = v(\mathbf{n}) - v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)$$

Therefore, we can rewrite **M1** to **M3** as follows:

$$\mu_3 A + \mu_1 B - \mu_3 C - \mu_1 D \leq 0.$$

Properties **M1** to **M3** can then be directly obtained by replacing $\mathbf{n}'$ with $(\mathbf{n}-\mathbf{e}^1)$, $(\mathbf{n}+\mathbf{e}^2)$, and $(\mathbf{n}+\mathbf{e}^3)$, respectively. To prove **M1** to **M3** are preserved by operator **T**, we need to show that:

$$\mu_3 \mathbf{T}A + \mu_1 \mathbf{T}B - \mu_3 \mathbf{T}C - \mu_1 \mathbf{T}D \leq 0.$$

Using Equation A.5, for $n_1 > 0$, $n_2 > 0$, $n_3 > 0$, we have:

$$\begin{aligned}
\mathbf{T}A = & \frac{-1}{\Lambda} + \frac{\lambda}{\Lambda}\left[v(\mathbf{n}' + \mathbf{e}^1 - \mathbf{e}^3) - v(\mathbf{n}' + \mathbf{e}^1)\right] \\
& + \frac{\mu_2}{\Lambda}\left[p\{v(\mathbf{n}' - \mathbf{e}^2) - v(\mathbf{n}' - \mathbf{e}^2 + \mathbf{e}^3)\} + (1-p)\{v(\mathbf{n}' - \mathbf{e}^2 - \mathbf{e}^3) - v(\mathbf{n}' - \mathbf{e}^2)\}\right] \\
& + \frac{1}{\Lambda} min \begin{cases} \mu_1 v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) + \mu_3 v(\mathbf{n}' - \mathbf{e}^3) \\ \mu_3 v(\mathbf{n}' - 2\mathbf{e}^3) + \mu_1 v(\mathbf{n}' - \mathbf{e}^3) \end{cases} \\
& - \frac{1}{\Lambda} min \begin{cases} \mu_1 v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2) + \mu_3 v(\mathbf{n}') \\ \mu_3 v(\mathbf{n}' - \mathbf{e}^3) + \mu_1 v(\mathbf{n}') \end{cases}
\end{aligned}$$

$$\begin{aligned}
\mathbf{T}B = & \frac{-1}{\Lambda} + \frac{\lambda}{\Lambda}\left[v(\mathbf{n}' + \mathbf{e}^1) - v(\mathbf{n}' + \mathbf{e}^2)\right] \\
& + \frac{\mu_2}{\Lambda}\left[p\{v(\mathbf{n}' - \mathbf{e}^2 + \mathbf{e}^3) - v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^3)\} + (1-p)\{v(\mathbf{n}' - \mathbf{e}^2) - v(\mathbf{n}' - \mathbf{e}^1)\}\right] \\
& + \frac{1}{\Lambda} min \begin{cases} \mu_1 v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2) + \mu_3 v(\mathbf{n}') \\ \mu_3 v(\mathbf{n}' - \mathbf{e}^3) + \mu_1 v(\mathbf{n}') \end{cases} \\
& - \frac{1}{\Lambda} min \begin{cases} \mu_1 v(\mathbf{n}' - 2\mathbf{e}^1 + 2\mathbf{e}^2) + \mu_3 v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2) \\ \mu_3 v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) + \mu_1 v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2) \end{cases}
\end{aligned}$$

$$\begin{aligned}
\mathbf{T}C = & \frac{-1}{\Lambda} + \frac{\lambda}{\Lambda}\left[v(\mathbf{n} - \mathbf{e}^1 - \mathbf{e}^3) - v(\mathbf{n} + \mathbf{e}^1)\right] \\
& + \frac{\mu_2}{\Lambda}\left[p\{v(\mathbf{n} - \mathbf{e}^2) - v(\mathbf{n} - \mathbf{e}^2 + \mathbf{e}^3)\} + (1-p)\{v(\mathbf{n} - \mathbf{e}^2 - \mathbf{e}^3) - v(\mathbf{n} - \mathbf{e}^2)\}\right] \\
& + \frac{1}{\Lambda} min \begin{cases} \mu_1 v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) + \mu_3 v(\mathbf{n} - \mathbf{e}^3) \\ \mu_3 v(\mathbf{n} - 2\mathbf{e}^3) + \mu_1 v(\mathbf{n} - \mathbf{e}^3) \end{cases} \\
& - \frac{1}{\Lambda} min \begin{cases} \mu_1 v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2) + \mu_3 v(\mathbf{n}) \\ \mu_3 v(\mathbf{n} - \mathbf{e}^3) + \mu_1 v(\mathbf{n}) \end{cases}
\end{aligned}$$

$$\begin{aligned}
\mathbf{T}D = & \frac{\lambda}{\Lambda}\left[v(\mathbf{n}+\mathbf{e}^1) - v(\mathbf{n}+\mathbf{e}^2)\right] \\
& + \frac{\mu_2}{\Lambda}\left[p\{v(\mathbf{n}-\mathbf{e}^2+\mathbf{e}^3) - v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^3)\} + (1-p)\{v(\mathbf{n}-\mathbf{e}^2) - v(\mathbf{n}-\mathbf{e}^1)\}\right] \\
& + \frac{1}{\Lambda} min \begin{cases} \mu_1 v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) + \mu_3 v(\mathbf{n}) \\ \mu_3 v(\mathbf{n}-\mathbf{e}^3) + \mu_1 v(\mathbf{n}) \end{cases} \\
& - \frac{1}{\Lambda} min \begin{cases} \mu_1 v(\mathbf{n}-2\mathbf{e}^1+2\mathbf{e}^2) + \mu_3 v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) \\ \mu_3 v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2-\mathbf{e}^3) + \mu_1 v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) \end{cases}
\end{aligned}$$

Furthermore, we let

$$\begin{aligned}
A' = & \tfrac{1}{\Lambda} min \begin{cases} \mu_1 v(\mathbf{n}'-\mathbf{e}^1+\mathbf{e}^2-\mathbf{e}^3) + \mu_3 v(\mathbf{n}'-\mathbf{e}^3) \\ \mu_3 v(\mathbf{n}'-2\mathbf{e}^3) + \mu_1 v(\mathbf{n}'-\mathbf{e}^3) \end{cases} \\
& - \tfrac{1}{\Lambda} min \begin{cases} \mu_1 v(\mathbf{n}'-\mathbf{e}^1+\mathbf{e}^2) + \mu_3 v(\mathbf{n}') \\ \mu_3 v(\mathbf{n}'-\mathbf{e}^3) + \mu_1 v(\mathbf{n}') \end{cases} \\
B' = & \tfrac{1}{\Lambda} min \begin{cases} \mu_1 v(\mathbf{n}'-\mathbf{e}^1+\mathbf{e}^2) + \mu_3 v(\mathbf{n}') \\ \mu_3 v(\mathbf{n}'-\mathbf{e}^3) + \mu_1 v(\mathbf{n}') \end{cases} \\
& - \tfrac{1}{\Lambda} min \begin{cases} \mu_1 v(\mathbf{n}'-2\mathbf{e}^1+2\mathbf{e}^2) + \mu_3 v(\mathbf{n}'-\mathbf{e}^1+\mathbf{e}^2) \\ \mu_3 v(\mathbf{n}'-\mathbf{e}^1+\mathbf{e}^2-\mathbf{e}^3) + \mu_1 v(\mathbf{n}'-\mathbf{e}^1+\mathbf{e}^2) \end{cases} \\
C' = & \tfrac{1}{\Lambda} min \begin{cases} \mu_1 v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2-\mathbf{e}^3) + \mu_3 v(\mathbf{n}-\mathbf{e}^3) \\ \mu_3 v(\mathbf{n}-2\mathbf{e}^3) + \mu_1 v(\mathbf{n}-\mathbf{e}^3) \end{cases} \\
& - \tfrac{1}{\Lambda} min \begin{cases} \mu_1 v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) + \mu_3 v(\mathbf{n}) \\ \mu_3 v(\mathbf{n}-\mathbf{e}^3) + \mu_1 v(\mathbf{n}) \end{cases} \\
D' = & \tfrac{1}{\Lambda} min \begin{cases} \mu_1 v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) + \mu_3 v(\mathbf{n}) \\ \mu_3 v(\mathbf{n}-\mathbf{e}^3) + \mu_1 v(\mathbf{n}) \end{cases}
\end{aligned}$$

$$-\tfrac{1}{\Lambda} min \begin{cases} \mu_1 v(\mathbf{n}-2\mathbf{e}^1+2\mathbf{e}^2)+\mu_3 v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) \\ \mu_3 v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2-\mathbf{e}^3)+\mu_1 v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) \end{cases}$$

Thus, we have:

$$\begin{aligned}
&\mu_3\mathbf{T}A+\mu_1\mathbf{T}B-\mu_3\mathbf{T}C-\mu_1\mathbf{T}D \\
&\quad=\tfrac{\lambda}{\Lambda}\Big\{\mu_3\left[v(\mathbf{n}'+\mathbf{e}^1-\mathbf{e}^3)-v(\mathbf{n}'+\mathbf{e}^1)\right]+\mu_1\left[v(\mathbf{n})-v(\mathbf{n}'+\mathbf{e}^1)-v(\mathbf{n}'+\mathbf{e}^2)\right] \\
&\qquad\quad-\mu_3\left[v(\mathbf{n}+\mathbf{e}^1-\mathbf{e}^3)-v(\mathbf{n}+\mathbf{e}^1)\right]-\mu_1\left[v(\mathbf{n}+\mathbf{e}^1)-v(\mathbf{n}+\mathbf{e}^2)\right]\Big\} \\
&\quad+\tfrac{\mu_2}{\Lambda}p\Big\{\mu_3\left[v(\mathbf{n}'-\mathbf{e}^2)-v(\mathbf{n}'-\mathbf{e}^2+\mathbf{e}^3)\right]+\mu_1\left[v(\mathbf{n}'-\mathbf{e}^2+\mathbf{e}^3)-v(\mathbf{n}'-\mathbf{e}^1+\mathbf{e}^3)\right] \\
&\qquad\quad-\mu_3\left[v(\mathbf{n}-\mathbf{e}^2)-v(\mathbf{n}-\mathbf{e}^2+\mathbf{e}^3)\right]-\mu_1\left[v(\mathbf{n}-\mathbf{e}^2+\mathbf{e}^3)-v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^3)\right]\Big\} \\
&\quad+\tfrac{\mu_2}{\Lambda}(1-p)\Big\{\mu_3\left[v(\mathbf{n}'-\mathbf{e}^2-\mathbf{e}^3)-v(\mathbf{n}'-\mathbf{e}^2)\right]+\mu_1\left[v(\mathbf{n}'-\mathbf{e}^2)-v(\mathbf{n}'-\mathbf{e}^1)\right] \\
&\qquad\quad-\mu_3\left[v(\mathbf{n}-\mathbf{e}^2-\mathbf{e}^3)-v(\mathbf{n}-\mathbf{e}^2)\right]-\mu_1\left[v(\mathbf{n}-\mathbf{e}^2)-v(\mathbf{n}-\mathbf{e}^1)\right]\Big\} \\
&\quad+\mu_3A'+\mu_1B'-\mu_3C'-\mu_1D'.
\end{aligned}$$

Since $v\in\Upsilon$, $\mu_3A+\mu_1B-\mu_3C-\mu_1D\le 0$, for $n_1>0, n_2\ge 0, n_3>0$. Thus, **M1-M3** are immediately holds for the first three terms of right hand side. Therefore, we only need to show that **M1-M3** hold for $\mu_3A'+\mu_1B'-\mu_3C'-\mu_1D'\le 0$. Note that we consider without loss of generality $n_2>0$. The case $n_2=0$ can be proved similarly. Considering the fact that **M1-M3** hold for any $v\in\Upsilon$, there are 5 possible cases we need to discuss in terms of which stage is optimal for Server S1 to work at states $(\mathbf{n}'-\mathbf{e}^3)$, $(\mathbf{n}')$, $(\mathbf{n}'-\mathbf{e}^1+\mathbf{e}^2)$, $(\mathbf{n}-\mathbf{e}^3)$, $(\mathbf{n})$, and $(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)$. Note that even though there are total of 64 cases in terms of which stage is optimal for Server S1 to work at aforementioned states, most of those cases are not valid, considering the fact that **M1-M3** hold for any $v\in\Upsilon$. For example, we can show that if at state $(\mathbf{n})$ working at Stage 1 is optimal, then at state $(\mathbf{n}-\mathbf{e}^3)$ working at Stage 1 is also optimal. More specifically, if at state $(\mathbf{n})$, it is optimal for

Server 1 to work at Stage 1, then $A^{P1}(\mathbf{n}) \leq A^{P3}(\mathbf{n})$. Therefore, for $n_1 > 0, n_2 \geq 0, n_3 > 0$, we have:

(A.2)

$$\mu_1 v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2) + \mu_3 v(\mathbf{n}) \leq \mu_3 v(\mathbf{n} - \mathbf{e}^3) + \mu_1 v(\mathbf{n})$$

$$\Rightarrow \mu_1 \left[v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2) - v(\mathbf{n}) \right] + \mu_3 \left[v(\mathbf{n}) - v(\mathbf{n} - \mathbf{e}^3) \right] \leq 0$$

To show that working at Stage 1 is also optimal at state $(\mathbf{n} - \mathbf{e}^3)$, we need to show $A^{P1}(\mathbf{n} - \mathbf{e}^3) \leq A^{P3}(\mathbf{n} - \mathbf{e}^3)$. According to **M3**, for $n_1 > 0, n_2 \geq 0, n_3 > 0$, we have:

$$\mu_1 \left[v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) - v(\mathbf{n} - \mathbf{e}^3) \right] + \mu_3 \left[v(\mathbf{n} - \mathbf{e}^3) - v(\mathbf{n} - 2\mathbf{e}^3) \right]$$

$$\leq \mu_1 \left[v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2) - v(\mathbf{n}) \right] + \mu_3 \left[v(\mathbf{n}) - v(\mathbf{n} - \mathbf{e}^3) \right]$$

When it is optimal for Server 1 to work at Stage 1 at $(\mathbf{n})$, using inequality (A.2), for $n_1 > 0, n_2 \geq 0, n_3 > 0$, we have:

$$\mu_1 \left[v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) - v(\mathbf{n} - \mathbf{e}^3) \right] + \mu_3 \left[v(\mathbf{n} - \mathbf{e}^3) - v(\mathbf{n} - 2\mathbf{e}^3) \right]$$

$$\leq \mu_1 \left[v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2) - v(\mathbf{n}) \right] + \mu_3 \left[v(\mathbf{n}) - v(\mathbf{n} - \mathbf{e}^3) \right] \leq 0$$

Hence, we can conclude:

$$\mu_1 \left[v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) - v(\mathbf{n} - \mathbf{e}^3) \right] + \mu_3 \left[v(\mathbf{n} - \mathbf{e}^3) - v(\mathbf{n} - 2\mathbf{e}^3) \right] \leq 0$$

$$\Rightarrow \mu_1 v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) + \mu_3 v(\mathbf{n} - \mathbf{e}^3) \leq \mu_3 v(\mathbf{n} - 2\mathbf{e}^3) + \mu_1 v(\mathbf{n} - \mathbf{e}^3)$$

This shows that $A^{P1}(\mathbf{n} - \mathbf{e}^3) \leq A^{P3}(\mathbf{n} - \mathbf{e}^3)$ and thus, it is optimal for Server 1 to work at Stage 1 at $(\mathbf{n} - \mathbf{e}^3)$, for $n_1 > 0, n_2 \geq 0, n_3 > 0$. It is tedious but easy to verify, in all 5 cases, $\mu_3 A' + \mu_1 B' - \mu_3 C' - \mu_1 D' \leq 0$, which implies $\mu_3 \mathbf{T} A + \mu_1 \mathbf{T} B - \mu_3 \mathbf{T} C - \mu_1 \mathbf{T} D \leq 0$. For example, for one of the most general cases where working at Stage 1 is optimal at $(\mathbf{n}' - \mathbf{e}^3)$, $(\mathbf{n}')$, $(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2)$, $(\mathbf{n} - \mathbf{e}^3)$, $(\mathbf{n})$, and $(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2)$, we will have

$$A' = \tfrac{1}{\Lambda}\Big\{ [\mu_1 v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) + \mu_3 v(\mathbf{n}' - \mathbf{e}^3)] - [\mu_1 v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2) + \mu_3 v(\mathbf{n}')] \Big\}$$
$$B' = \tfrac{1}{\Lambda}\Big\{ [\mu_1 v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2) + \mu_3 v(\mathbf{n}')] - [\mu_1 v(\mathbf{n}' - 2\mathbf{e}^1 + 2\mathbf{e}^2) + \mu_3 v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2)] \Big\}$$
$$C' = \tfrac{1}{\Lambda}\Big\{ [\mu_1 v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) + \mu_3 v(\mathbf{n} - \mathbf{e}^3)] - [\mu_1 v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2) + \mu_3 v(\mathbf{n})] \Big\}$$
$$D' = \tfrac{1}{\Lambda}\Big\{ [\mu_1 v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2) + \mu_3 v(\mathbf{n})] - [\mu_1 v(\mathbf{n} - 2\mathbf{e}^1 + 2\mathbf{e}^2) + \mu_3 v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2)] \Big\}$$

Therefore,

$$\begin{aligned}
&\Lambda\,[A' - C'] \\
&= [\mu_1 v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) + \mu_3 v(\mathbf{n}' - \mathbf{e}^3)] - [\mu_1 v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2) + \mu_3 v(\mathbf{n}')] \\
&\qquad - \Big\{ [\mu_1 v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) + \mu_3 v(\mathbf{n} - \mathbf{e}^3)] - [\mu_1 v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2) + \mu_3 v(\mathbf{n})] \Big\} \\
&= \mu_1 [v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) - v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2)] + \mu_3 [v(\mathbf{n}' - \mathbf{e}^3) - v(\mathbf{n}')] \\
&\qquad - \mu_1 [v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) - v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2)] - \mu_3 [v(\mathbf{n} - \mathbf{e}^3) - v(\mathbf{n})] \\
&= \mu_3 [v(\mathbf{n}' - \mathbf{e}^3) - v(\mathbf{n}')] + \mu_1 [v(\mathbf{n} - \mathbf{e}^1) - v(\mathbf{n} - 2\mathbf{e}^1 + \mathbf{e}^2)] \\
&\qquad - \mu_3 [v(\mathbf{n} - \mathbf{e}^3) - v(\mathbf{n})] - \mu_1 [v(\mathbf{n}) - v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2)] \\
&\qquad + \mu_1 \Big\{ [v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) - v(\mathbf{n} - \mathbf{e}^1)] - [v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) - v(\mathbf{n})] \Big\}
\end{aligned}$$

Considering the fact that **M1-M3** hold, we have

$$\Lambda\,[A' - C'] \leq \mu_1 \{ [v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) - v(\mathbf{n}')] - [v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) - v(\mathbf{n})] \}$$

Similarly,

$$\begin{aligned}
&\Lambda\,[B' - D'] \\
&= [\mu_1 v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2) + \mu_3 v(\mathbf{n}')] - [\mu_1 v(\mathbf{n}' - 2\mathbf{e}^1 + 2\mathbf{e}^2) + \mu_3 v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2)] \\
&\qquad - \{ [\mu_1 v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2) + \mu_3 v(\mathbf{n})] - [\mu_1 v(\mathbf{n} - 2\mathbf{e}^1 + 2\mathbf{e}^2) + \mu_3 v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2)] \} \\
&= \mu_1 [v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2) - v(\mathbf{n}' - 2\mathbf{e}^1 + 2\mathbf{e}^2)] + \mu_3 [v(\mathbf{n}') - v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2)]
\end{aligned}$$

$$
\begin{aligned}
&- \mu_1 \left[v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2) - v(\mathbf{n} - 2\mathbf{e}^1 + 2\mathbf{e}^2)\right] - \mu_3 \left[v(\mathbf{n}) - v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2)\right] \\
&= \mu_3 \left[v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) - v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2)\right] + \mu_1 \left[v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2) - v(\mathbf{n}' - 2\mathbf{e}^{\mathbf{1}} + 2\mathbf{e}^2)\right] \\
&\quad - \mu_3 \left[v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) - v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2)\right] - \mu_1 \left[v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2) - v(\mathbf{n} - 2\mathbf{e}^1 + 2\mathbf{e}^2)\right] \\
&\quad + \mu_3 \{[v(\mathbf{n}') - v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3)] - [v(\mathbf{n}) - v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3)]\}
\end{aligned}
$$

Considering the fact that **M1-M3** hold at state $(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2)$, we have

$$\Lambda \left[B' - D'\right] \leq \mu_3 \{[v(\mathbf{n}') - v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3)] - [v(\mathbf{n}) - v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3)]\}$$

Therefore, we have:

$$
\begin{aligned}
&\Lambda \left[\mu_3 A' + \mu_1 B' - \mu_3 C' - \mu_1 D'\right] = \Lambda \mu_3 \left[A' - C'\right] + \Lambda \mu_1 \left[B' - D'\right] \\
&\leq \mu_3 \mu_1 \{[v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) - v(\mathbf{n}')] - [v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3) - v(\mathbf{n})]\} \\
&\qquad + \mu_1 \mu_3 \{[v(\mathbf{n}') - v(\mathbf{n}' - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3)] - [v(\mathbf{n}) - v(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2 - \mathbf{e}^3)]\} = 0
\end{aligned}
$$

which implies that $\mu_3 A' + \mu_1 B' - \mu_3 C' - \mu_1 D' \leq 0$. This completes the proof that **M1-M3** are preserved under operator $\mathbf{T}$ and therefore the proof of Proposition 1.2 is complete.

□

Proof of Proposition 1.2(b). The proof is very similar to the proofs in [12] and [13]. Using the L^∞ metric, the limit of any convergent sequence of functions in Υ will be in Υ as well. Therefore, Υ is complete. Now, define a structure decision rules with the

state-dependent thresholds:

$$n^*_{1,(n_2,n_3)} = \min\{n_1 | V(\mathbf{n}) \geq 0, n_1 > 0\}$$

$$n^*_{2,(n_1,n_3)} = \min\{n_2 | V(\mathbf{n}) \geq 0, n_2 > 0\}$$

$$n^*_{3,(n_1,n_2)} = \min\{n_3 | V(\mathbf{n}) \geq 0, n_3 > 0\}$$

That is, given $n_2 \geq 0$ and $n_3 \geq 0$, the decision for Server S1 is to work at Stage 1, if the number of task at Stage 1 (i.e., n_1) is more than $n^*_{1,(n_2,n_3)}$, or work at Stage 3 if the number of task at Stage 1 (i.e., n_1) is less than $n^*_{1,(n_2,n_3)}$ and $n_3 > 0$, or stay idle otherwise. Similarly, given $n_1 \geq 0$ and $n_3 \geq 0$, the decision for Server S1 is to work at Stage 1, if the number of task at Stage 2 (i.e., n_2) is less than $n^*_{2,(n_1,n_3)}$ and $n_1 > 0$, or work at Stage 3 if the number of task at Stage 2 (i.e., n_2) is more than $n^*_{2,(n_1,n_3)}$ and $n_3 > 0$, or stay idle otherwise. Finally, given $n_1 \geq 0$ and $n_2 \geq 0$, the decision for Server S1 is to work at Stage 3, if the number of task at Stage 3 (i.e., n_3) is more than $n^*_{3,(n_1,n_2)}$, or work at Stage 1 if the number of task at Stage 3 (i.e., n_3) is less than $n^*_{3,(n_1,n_2)}$ and $n_1 > 0$, or stay idle otherwise. It can be shown that the structured decision rules satisfy the optimality equation (1). By Theorem 5.1 of [23], the optimal value function V is structured and has properties **C1**, **C2**, and **M1-M3**.

□

Proof of Theorem 1.2. This Theorem is a direct result of the fact that value function V satisfies properties **C1** and **C2**, as shown in Proposition 1.2. At state ($\mathbf{n}$), we

have:

$$A_I(\mathbf{n}) = (\mu_1 + \mu_3)V(\mathbf{n})$$

$$A_{P1}(\mathbf{n}) = \begin{cases} \mu_1 V(\mathbf{n}) + \mu_3 V(\mathbf{n}) & : \text{ if } n_1 = 0 \\ \mu_1 V(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2) + \mu_3 V(\mathbf{n}) & : \text{ if } n_1 > 0 \end{cases}$$

$$A_{P3}(\mathbf{n}) = \begin{cases} \mu_3 V(\mathbf{n}) + \mu_1 V(\mathbf{n}) & : \text{ if } n_3 = 0 \\ \mu_3 V(\mathbf{n} - \mathbf{e}^3) + \mu_1 V(\mathbf{n}) & : \text{ if } n_3 > 0 \end{cases}$$

According to Proposition 1.2(b), since $V \in \Upsilon$, V satisfies property **C1** and this implies that $V(\mathbf{n} + \mathbf{e}^3) \geq V(\mathbf{n})$. Therefore, we have $A_{P3}(\mathbf{n}) \leq A_I(\mathbf{n})$ if $n_3 > 0$. Therefore, idling cannot be optimal at state $(\mathbf{n})$ if $n_3 > 0$. Similarly, since V satisfies property **C2**, $V(\mathbf{n}) \geq V(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2)$, we have $A_{P1}(\mathbf{n}) \leq A_I(\mathbf{n})$ if $n_1 > 0$. Therefore, idling cannot be optimal at state $(\mathbf{n})$ if $n_1 > 0$. This completes the proof of Theorem 1.2. □

Proof of Theorem 1.3. Theorem 1.3 has two parts: Work at Stage 3 and Work at Stage 1. We first prove the first part of Theorem 1.3.

Work at Stage 3: At state $(\mathbf{n})$, if it is optimal for Server S1 to work at Stage 3, then $A_{P3}(\mathbf{n}) \leq A_{P1}(\mathbf{n})$. Therefore, for $n_1 > 0, n_2 \geq 0, n_3 > 0$, we have:

$$\begin{aligned} &\mu_3 V(\mathbf{n} - \mathbf{e}^3) + \mu_1 V(\mathbf{n}) \leq \mu_1 V(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2) + \mu_3 V(\mathbf{n}) \\ &\qquad \Rightarrow \mu_3[V(\mathbf{n} - \mathbf{e}^3) - V(\mathbf{n})] + \mu_1[V(\mathbf{n}) - V(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2)] \leq 0 \end{aligned} \tag{A.3}$$

There are 3 cases to discuss:

noitemsep,nolistsep Case 1: at state $(\mathbf{n} - \mathbf{e}^1)$,

noitemsep,nolistsep Case 2: at state $(\mathbf{n} + \mathbf{e}^2)$,

noitemsep,nolistsep Case 3: at state $(\mathbf{n}+\mathbf{e}^3)$.

Case 1 $(\mathbf{n}-\mathbf{e}^1)$: To show that working at Stage 3 is also optimal at state $(\mathbf{n}-\mathbf{e}^1)$, we need to show $A_{P3}(\mathbf{n}-\mathbf{e}^1) \leq A_{P1}(\mathbf{n}-\mathbf{e}^1)$. According to Proposition 1.2, since $V \in \Upsilon$, V satisfies property **M1**, for $n_1 > 1, n_2 \geq 0, n_3 > 0$, and we have:

$$\mu_3[V(\mathbf{n}-\mathbf{e}^1-\mathbf{e}^3)-V(\mathbf{n}-\mathbf{e}^1)]+\mu_1[V(\mathbf{n}-\mathbf{e}^1)-V(\mathbf{n}-2\mathbf{e}^1+\mathbf{e}^2)]$$
$$\leq \mu_3[V(\mathbf{n}-\mathbf{e}^3)-V(\mathbf{n})]+\mu_1[V(\mathbf{n})-V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)]$$

When it is optimal for Server S1 to work at Stage 3 at $(\mathbf{n})$, using inequality (A.3), for $n_1 > 1, n_2 \geq 0, n_3 > 0$, we have:

$$\mu_3[V(\mathbf{n}-\mathbf{e}^1-\mathbf{e}^3)-V(\mathbf{n}-\mathbf{e}^1)]+\mu_1[V(\mathbf{n}-\mathbf{e}^1)-V(\mathbf{n}-2\mathbf{e}^1+\mathbf{e}^2)]$$
$$\leq \mu_3[V(\mathbf{n}-\mathbf{e}^3)-V(\mathbf{n})]+\mu_1[V(\mathbf{n})-V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)] \leq 0$$

Hence, we can conclude:

$$\mu_3[V(\mathbf{n}-\mathbf{e}^1-\mathbf{e}^3)-V(\mathbf{n}-\mathbf{e}^1)]+\mu_1[V(\mathbf{n}-\mathbf{e}^1)-V(\mathbf{n}-2\mathbf{e}^1+\mathbf{e}^2)] \leq 0$$
$$\Rightarrow \mu_3 V(\mathbf{n}-\mathbf{e}^1-\mathbf{e}^3)+\mu_1 V(\mathbf{n}-\mathbf{e}^1) \leq \mu_1 V(\mathbf{n}-2\mathbf{e}^1+\mathbf{e}^2)+\mu_3 V(\mathbf{n}-\mathbf{e}^1)$$

This shows that $A_{P3}(\mathbf{n}-\mathbf{e}^1) \leq A_{P1}(\mathbf{n}-\mathbf{e}^1)$ and thus, it is optimal for Server S1 to work at Stage 3 at $(\mathbf{n}-\mathbf{e}^1)$, for $n_1 > 0, n_2 \geq 0, n_3 > 0$.

Case 2 $(\mathbf{n}+\mathbf{e}^2)$: To show that working at Stage 3 is also optimal at state $(\mathbf{n}+\mathbf{e}^2)$, we need to show $A_{P3}(\mathbf{n}+\mathbf{e}^2) \leq A_{P1}(\mathbf{n}+\mathbf{e}^2)$. According to Proposition 1.2, since $V \in \Upsilon$, V satisfies property **M2**, for $n_1 > 0, n_2 \geq 0, n_3 > 0$, and we have:

$$\mu_3[V(\mathbf{n}+\mathbf{e}^2-\mathbf{e}^3)-V(\mathbf{n}+\mathbf{e}^2)]+\mu_1[V(\mathbf{n}+\mathbf{e}^2)-V(\mathbf{n}-\mathbf{e}^1+2\mathbf{e}^2)]$$
$$\leq \mu_3[V(\mathbf{n}-\mathbf{e}^3)-V(\mathbf{n})]+\mu_1[V(\mathbf{n})-V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)]$$

When it is optimal for Server S1 to work at Stage 3 at $(\mathbf{n})$, using inequality (A.3), for $n_1 > 0, n_2 \geq 0, n_3 > 0$, we have:

$$\mu_3[V(\mathbf{n}+\mathbf{e}^2-\mathbf{e}^3)-V(\mathbf{n}+\mathbf{e}^2)]+\mu_1[V(\mathbf{n}+\mathbf{e}^2)-V(\mathbf{n}-\mathbf{e}^1+2\mathbf{e}^2)]$$
$$\leq \mu_3[V(\mathbf{n}-\mathbf{e}^3)-V(\mathbf{n})]+\mu_1[V(\mathbf{n})-V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)] \leq 0$$

Hence, we can conclude:

$$\mu_3[V(\mathbf{n}+\mathbf{e}^2-\mathbf{e}^3)-V(\mathbf{n}+\mathbf{e}^2)]+\mu_1[V(\mathbf{n}+\mathbf{e}^2)-V(\mathbf{n}-\mathbf{e}^1+2\mathbf{e}^2)] \leq 0$$
$$\Rightarrow \mu_3 V(\mathbf{n}+\mathbf{e}^2-\mathbf{e}^3)+\mu_1 V(\mathbf{n}+\mathbf{e}^2) \leq \mu_1 V(\mathbf{n}-\mathbf{e}^1+2\mathbf{e}^2)+\mu_3 V(\mathbf{n}+\mathbf{e}^2)$$

This shows that $A_{P3}(\mathbf{n}+\mathbf{e}^2) \leq A_{P1}(\mathbf{n}+\mathbf{e}^2)$ and thus, it is optimal for Server S1 to work at Stage 3 at $(\mathbf{n}+\mathbf{e}^2)$, for $n_1 > 0, n_2 \geq 0, n_3 > 0$.

Case 3 $(\mathbf{n}+\mathbf{e}^3)$: To show that working at Stage 3 is also optimal at state $(\mathbf{n}+\mathbf{e}^3)$, we need to show $A_{P3}(\mathbf{n}+\mathbf{e}^3) \leq A_{P1}(\mathbf{n}+\mathbf{e}^3)$. According to Proposition 1.2, since $V \in \Upsilon$, V satisfies property **M3**, for $n_1 > 0, n_2 \geq 0, n_3 > 0$, and we have:

$$\mu_3[V(\mathbf{n})-V(\mathbf{n}+\mathbf{e}^3)]+\mu_1[V(\mathbf{n}+\mathbf{e}^3)-V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2+\mathbf{e}^3)]$$
$$\leq \mu_3[V(\mathbf{n}-\mathbf{e}^3)-V(\mathbf{n})]+\mu_1[V(\mathbf{n})-V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)]$$

When it is optimal for Server S1 to work at Stage 3 at $(\mathbf{n})$, using inequality (A.3), for $n_1 > 0, n_2 \geq 0, n_3 > 0$, we have:

$$\mu_3[V(\mathbf{n})-V(\mathbf{n}+\mathbf{e}^3)]+\mu_1[V(\mathbf{n}+\mathbf{e}^3)-V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2+\mathbf{e}^3)]$$
$$\leq \mu_3[V(\mathbf{n}-\mathbf{e}^3)-V(\mathbf{n})]+\mu_1[V(\mathbf{n})-V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)] \leq 0$$

Hence, we can conclude:

$$\mu_3[V(\mathbf{n})-V(\mathbf{n}+\mathbf{e}^3)]+\mu_1[V(\mathbf{n}+\mathbf{e}^3)-V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2+\mathbf{e}^3)] \leq 0$$
$$\Rightarrow \mu_3 V(\mathbf{n})+\mu_1 V(\mathbf{n}+\mathbf{e}^3) \leq \mu_1 V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2+\mathbf{e}^3)+\mu_3 V(\mathbf{n}+\mathbf{e}^3)$$

This shows that $A_{P3}(\mathbf{n}+\mathbf{e}^3) \leq A_{P1}(\mathbf{n}+\mathbf{e}^3)$ and thus, it is optimal for Server S1 to work at Stage 3 at $(\mathbf{n}+\mathbf{e}^3)$, for $n_1 > 0, n_2 \geq 0, n_3 > 0$. This concludes the proof of the first part of the Theorem. Now, we prove the second part of Theorem 1.3.

Work at Stage 1: At state $(\mathbf{n})$, if it is optimal for Server S1 to work at Stage 1, then $A_{P1}(\mathbf{n}) \leq A_{P3}(\mathbf{n})$. Therefore, for $n_1 > 0, n_2 \geq 0, n_3 > 0$, we have:

$$\begin{aligned} &\mu_1 V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) + \mu_3 V(\mathbf{n}) \leq \mu_3 V(\mathbf{n}-\mathbf{e}^3) + \mu_1 V(\mathbf{n}) \\ &\qquad \Rightarrow \mu_1[V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) - V(\mathbf{n})] + \mu_3[V(\mathbf{n}) - V(\mathbf{n}-\mathbf{e}^3)] \leq 0 \end{aligned} \tag{A.4}$$

There are 3 cases to discuss:

noitemsep,nolistsep Case 1: at state $(\mathbf{n}+\mathbf{e}^1)$,

noitemsep,nolistsep Case 2: at state $(\mathbf{n}-\mathbf{e}^2)$,

noitemsep,nolistsep Case 3: at state $(\mathbf{n}-\mathbf{e}^3)$.

Case 1 $(\mathbf{n}+\mathbf{e}^1)$: To show that working at Stage 1 is also optimal at state $(\mathbf{n}+\mathbf{e}^1)$, we need to show $A_{P1}(\mathbf{n}+\mathbf{e}^1) \leq A_{P3}(\mathbf{n}+\mathbf{e}^1)$. According to Proposition 1.2, since $V \in \Upsilon$, V satisfies property **M1**, for $n_1 > 0, n_2 \geq 0, n_3 > 0$, and we have:

$$\begin{aligned} &\mu_1[V(\mathbf{n}+\mathbf{e}^2) - V(\mathbf{n}+\mathbf{e}^1)] + \mu_3[V(\mathbf{n}+\mathbf{e}^1) - V(\mathbf{n}+\mathbf{e}^1-\mathbf{e}^3)] \\ &\qquad \leq \mu_1[V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) - V(\mathbf{n})] + \mu_3[V(\mathbf{n}) - V(\mathbf{n}-\mathbf{e}^3)] \end{aligned}$$

When it is optimal for Server S1 to work at Stage 1 at $(\mathbf{n})$, using inequality (A.4), for $n_1 > 0, n_2 \geq 0, n_3 > 0$, we have:

$$\begin{aligned} &\mu_1[V(\mathbf{n}+\mathbf{e}^2) - V(\mathbf{n}+\mathbf{e}^1)] + \mu_3[V(\mathbf{n}+\mathbf{e}^1) - V(\mathbf{n}+\mathbf{e}^1-\mathbf{e}^3)] \\ &\qquad \leq \mu_1[V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) - V(\mathbf{n})] + \mu_3[V(\mathbf{n}) - V(\mathbf{n}-\mathbf{e}^3)] \leq 0 \end{aligned}$$

Hence, we can conclude:

$$\mu_1[V(\mathbf{n}+\mathbf{e}^2) - V(\mathbf{n}+\mathbf{e}^1)] + \mu_3[V(\mathbf{n}+\mathbf{e}^1) - V(\mathbf{n}+\mathbf{e}^1-\mathbf{e}^3)] \leq 0$$

$$\Rightarrow \mu_1 V(\mathbf{n}+\mathbf{e}^2) + \mu_3 V(\mathbf{n}+\mathbf{e}^1) \leq \mu_3 V(\mathbf{n}+\mathbf{e}^1-\mathbf{e}^3) + \mu_1 V(\mathbf{n}+\mathbf{e}^1)$$

This shows that $A_{P1}(\mathbf{n}+\mathbf{e}^1) \leq A_{P3}(\mathbf{n}+\mathbf{e}^1)$ and thus, it is optimal for Server S1 to work at Stage 1 at $(\mathbf{n}+\mathbf{e}^1)$, for $n_1 > 0, n_2 \geq 0, n_3 > 0$.

Case 2 $(\mathbf{n}-\mathbf{e}^2)$: To show that working at Stage 1 is also optimal at state $(\mathbf{n}-\mathbf{e}^2)$, we need to show $A_{P1}(\mathbf{n}-\mathbf{e}^2) \leq A_{P3}(\mathbf{n}-\mathbf{e}^2)$. According to Proposition 1.2, since $V \in \Upsilon$, V satisfies property **M2**, for $n_1 > 0, n_2 \geq 0, n_3 > 0$, and we have:

$$\mu_1[V(\mathbf{n}-\mathbf{e}^1) - V(\mathbf{n}-\mathbf{e}^2)] + \mu_3[V(\mathbf{n}-\mathbf{e}^2) - V(\mathbf{n}-\mathbf{e}^2-\mathbf{e}^3)]$$
$$\leq \mu_1[V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) - V(\mathbf{n})] + \mu_3[V(\mathbf{n}) - V(\mathbf{n}-\mathbf{e}^3)]$$

When it is optimal for Server S1 to work at Stage 1 at $(\mathbf{n})$, using inequality (A.4), for $n_1 > 0, n_2 \geq 0, n_3 > 0$, we have:

$$\mu_1[V(\mathbf{n}-\mathbf{e}^1) - V(\mathbf{n}-\mathbf{e}^2)] + \mu_3[V(\mathbf{n}-\mathbf{e}^2) - V(\mathbf{n}-\mathbf{e}^2-\mathbf{e}^3)]$$
$$\leq \mu_1[V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) - V(\mathbf{n})] + \mu_3[V(\mathbf{n}) - V(\mathbf{n}-\mathbf{e}^3)] \leq 0$$

Hence, we can conclude:

$$\mu_1[V(\mathbf{n}-\mathbf{e}^1) - V(\mathbf{n}-\mathbf{e}^2)] + \mu_3[V(\mathbf{n}-\mathbf{e}^2) - V(\mathbf{n}-\mathbf{e}^2-\mathbf{e}^3)] \leq 0$$
$$\Rightarrow \mu_1 V(\mathbf{n}-\mathbf{e}^1) + \mu_3 V(\mathbf{n}-\mathbf{e}^2) \leq \mu_3 V(\mathbf{n}-\mathbf{e}^2-\mathbf{e}^3) + \mu_1 V(\mathbf{n}-\mathbf{e}^2)$$

This shows that $A_{P1}(\mathbf{n}-\mathbf{e}^2) \leq A_{P3}(\mathbf{n}-\mathbf{e}^2)$ and thus, it is optimal for Server S1 to work at Stage 1 at $(\mathbf{n}-\mathbf{e}^2)$, for $n_1 > 0, n_2 \geq 0, n_3 > 0$.

Case 3 $(\mathbf{n}-\mathbf{e}^3)$: To show that working at Stage 1 is also optimal at state $(\mathbf{n}-\mathbf{e}^3)$, we need to show $A_{P1}(\mathbf{n}-\mathbf{e}^3) \leq A_{P3}(\mathbf{n}-\mathbf{e}^3)$. According to Proposition 1.2, since $V \in \Upsilon$, V satisfies property **M3**, for $n_1 > 0, n_2 \geq 0, n_3 > 0$, and we have:

$$\mu_1[V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2-\mathbf{e}^3) - V(\mathbf{n}-\mathbf{e}^3)] + \mu_3[V(\mathbf{n}-\mathbf{e}^3) - V(\mathbf{n}-2\mathbf{e}^3)]$$
$$\leq \mu_1[V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2) - V(\mathbf{n})] + \mu_3[V(\mathbf{n}) - V(\mathbf{n}-\mathbf{e}^3)]$$

When it is optimal for Server S1 to work at Stage 1 at state $(\mathbf{n})$, using inequality (A.4), for $n_1 > 0, n_2 \geq 0, n_3 > 0$, we have:

$$\mu_1[V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2-\mathbf{e}^3)-V(\mathbf{n}-\mathbf{e}^3)]+\mu_3[V(\mathbf{n}-\mathbf{e}^3)-V(\mathbf{n}-2\mathbf{e}^3)]$$
$$\leq \mu_1[V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)-V(\mathbf{n})]+\mu_3[V(\mathbf{n})-V(\mathbf{n}-\mathbf{e}^3)] \leq 0$$

Hence, we can conclude:

$$\mu_1[V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2-\mathbf{e}^3)-V(\mathbf{n}-\mathbf{e}^3)]+\mu_3[V(\mathbf{n}-\mathbf{e}^3)-V(\mathbf{n}-2\mathbf{e}^3)] \leq 0$$
$$\Rightarrow \mu_1 V(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2-\mathbf{e}^3)+\mu_3 V(\mathbf{n}-\mathbf{e}^3) \leq \mu_3 V(\mathbf{n}-2\mathbf{e}^3)+\mu_1 V(\mathbf{n}-\mathbf{e}^3)$$

This shows that $A_{P1}(\mathbf{n}-\mathbf{e}^3) \leq A_{P3}(\mathbf{n}-\mathbf{e}^3)$ and thus, it is optimal for Server S1 to work at Stage 1 at $(\mathbf{n}-\mathbf{e}^3)$, for $n_1 > 0, n_2 \geq 0, n_3 > 0$. This concludes the proof of the second part of the Theorem. This completes the proof of Theorem 1.3.

□

Proof of Proposition 1.3. In this proposition, we want to show when there is a new independent arrival to Stage 2, the structural properties presented in Proposition 1.2, Theorem 1.2 and Theorem 1.1 still hold.

In this case, system stability conditions will be changed to:

$$\mu_1 > \lambda,\ \mu_2 > \lambda + \lambda^e$$
$$\mu_3 > p(\lambda + \lambda^e)$$
$$\frac{\mu_1\mu_3 - p\mu_1\lambda^e}{p\mu_1 + \mu_3} > \lambda$$

The proof is similar to the proof of Proposition 1.1. If these conditions hold, Theorem 1.1 still holds.

Here, we discuss that the value function $V^e(\mathbf{n})$ satisfies the structural properties and corresponding results presented for value function $V(\mathbf{n})$ in Proposition 1.2, Theorem 1.2 and Theorem 1.3.

The proof of Proposition 1.2 for the value function V^e is very similar to that for the value function V. Similar to Proposition 1.2, let Υ^e be the set of functions defined on $\mathscr{U}$ such that if function $v^e \in \Upsilon^e$, then v^e satisfies

C1: $v^e(\mathbf{n})$ is nondecreasing in $n_1 \geq 0$, $n_2 \geq 0$ and $n_3 \geq 0$.

C2: $v^e(\mathbf{n}) \geq v^e(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2)$, for $n_1 > 1, n_2 \geq 0, n_3 \geq 0$.

M1: For $n_1 > 1, n_2 \geq 0, n_3 > 0$,

$$\mu_3[v^e(\mathbf{n} - \mathbf{e}^1 - \mathbf{e}^3) - v^e(\mathbf{n} - \mathbf{e}^1)] + \mu_1[v^e(\mathbf{n} - \mathbf{e}^1) - v^e(\mathbf{n} - 2\mathbf{e}^1 + \mathbf{e}^2)]$$
$$\leq \mu_3[v^e(\mathbf{n} - \mathbf{e}^3) - v^e(\mathbf{n})] + \mu_1[v^e(\mathbf{n}) - v^e(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2)].$$

M2: For $n_1 > 0, n_2 \geq 0, n_3 > 0$,

$$\mu_3[v^e(\mathbf{n} + \mathbf{e}^2 - \mathbf{e}^3) - v^e(\mathbf{n} + \mathbf{e}^2)] + \mu_1[v^e(\mathbf{n} + \mathbf{e}^2) - v^e(\mathbf{n} - \mathbf{e}^1 + 2\mathbf{e}^2)]$$
$$\leq \mu_3[v^e(\mathbf{n} - \mathbf{e}^3) - v^e(\mathbf{n})] + \mu_1[v^e(\mathbf{n}) - v^e(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2)].$$

M3: For $n_1 > 0, n_2 \geq 0, n_3 > 0$,

$$\mu_3[v^e(\mathbf{n}) - v^e(\mathbf{n} + \mathbf{e}^3)] + \mu_1[v^e(\mathbf{n} + \mathbf{e}^3) - v^e(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2 + \mathbf{e}^3)]$$
$$\leq \mu_3[v^e(\mathbf{n} - \mathbf{e}^3) - v^e(\mathbf{n})] + \mu_1[v^e(\mathbf{n}) - v^e(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2)].$$

For any $v^e \in \Upsilon^e$, we have

(A.5)

$$\mathbf{T}v^e(\mathbf{n}) = \frac{1}{\Lambda}\Big\{n_1 + n_2 + n_3 + \lambda v^e(\mathbf{n} + \mathbf{e}^1) + \lambda^e v^e(\mathbf{n} + \mathbf{e}^2) + \mu_2 B^e(\mathbf{n}) + \min\Big\{A_I^e(\mathbf{n}), A_{P1}^e(\mathbf{n}), A_{P3}^e(\mathbf{n})\Big\}\Big\},$$

$$B^e(\mathbf{n}) = \begin{cases} v^e(\mathbf{n}) & : \text{ if } n_2 = 0 \\ pv^e(\mathbf{n} - \mathbf{e}^2 + \mathbf{e}^3) + (1-p)v^e(\mathbf{n} - \mathbf{e}^2) & : \text{ if } n_2 > 0 \end{cases}$$

$$A^e_I(\mathbf{n}) = (\mu_1 + \mu_3)v^e(\mathbf{n})$$

$$A^e_{P1}(\mathbf{n}) = \begin{cases} \mu_1 v^e(\mathbf{n}) + \mu_3 v^e(\mathbf{n}) & : \text{ if } n_1 = 0 \\ \mu_1 v^e(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2) + \mu_3 v^e(\mathbf{n}) & : \text{ if } n_1 > 0 \end{cases}$$

$$A^e_{P3}(\mathbf{n}) = \begin{cases} \mu_3 v^e(\mathbf{n}) + \mu_1 v^e(\mathbf{n}) & : \text{ if } n_3 = 0 \\ \mu_3 v^e(\mathbf{n} - \mathbf{e}^3) + \mu_1 v^e(\mathbf{n}) & : \text{ if } n_3 > 0 \end{cases}$$

We want to show if $v^e \in \Upsilon^e$, then (a) $\mathbf{T}v^e \in \Upsilon^e$, and (b) $V^e \in \Upsilon^e$. To prove part (a), we show property **C1** is preserved by operator $\mathbf{T}$, i.e., $\mathbf{T}v^e(\mathbf{n})$ is nondecreasing in $n_1 \geq 0$, as an example and the rest can be shown similarly.

Consider property **C1**: $v^e(\mathbf{n})$ is nondecreasing in n_1, for $n_1 \geq 0$, $n_2 \geq 0$ and $n_3 \geq 0$. To prove that **C1** is preserved by operator $\mathbf{T}$, we need to consider 8 possible cases for **C1** in terms of $n_i > 0$ or $n_i = 0$, $i \in \{1, 2, 3\}$. First, consider the case ($n_1 > 0$, $n_2 > 0$, $n_3 > 0$). Define

$$AC^e_a(\mathbf{n}) = \begin{cases} (\mu_1 + \mu_3)v^e(\mathbf{n}), & a = I \\ \mu_1 v^e(\mathbf{n} - \mathbf{e}^1 + \mathbf{e}^2) + \mu_3 v^e(\mathbf{n}), & a = P1 \\ \mu_1 v^e(\mathbf{n}) + \mu_3 v^e(\mathbf{n} - \mathbf{e}^3), & a = P3 \end{cases}$$

Next, we define $AC^e_{a^*_{(\mathbf{n})}}(\mathbf{n}) = \min_a AC^e_a(\mathbf{n})$ as the value of $AC^e_a(\mathbf{n})$ under action $a^*_{(\mathbf{n})} \in \{I, P1, P3\}$, where $a^*_{(\mathbf{n})}$ is the optimal action at state $(\mathbf{n})$. More generally, we define

$AC^e_{a^*_{(\mathbf{n}+\mathbf{e}^i)}}(\mathbf{n}+\mathbf{e}^j)$ $(i,j=1,2,3)$ as the value of $AC^e_a(\mathbf{n}-\mathbf{e}^j)$ under action $a^*_{(\mathbf{n}+\mathbf{e}^i)}$, where $a^*_{(\mathbf{n}-\mathbf{e}^i)}$ is the optimal action at state $(\mathbf{n}-\mathbf{e}^i)$. Therefore, we can rewrite Equation (A.5) as follows:

$$\mathbf{T}v^e(\mathbf{n}) = \frac{1}{\Lambda}\Big\{ n_1+n_2+n_3+\lambda v^e(\mathbf{n}+\mathbf{e}^1)+\lambda^e v^e(\mathbf{n}+\mathbf{e}^2) + \mu_2\left[pv^e(\mathbf{n}-\mathbf{e}^2+\mathbf{e}^3)+(1-p)v^e(\mathbf{n}-\mathbf{e}^2)\right]+AC^e_{a^*_{(\mathbf{n})}}(\mathbf{n})\Big\}.$$

Since $v^e \in \Upsilon^e$, $v^e(\mathbf{n})$ is nondecreasing in n_1, for $n_1 \geq 0, n_2 \geq 0, n_3 \geq 0$. Thus, **C1** immediately holds for the first four $v^e(.)$ terms of right hand sides of $\mathbf{T}v^e(\mathbf{n})$. Therefore, we only need to show that **C1** holds for $AC^e_{a^*_{(\mathbf{n})}}(\mathbf{n})$. If the optimal action at state $(\mathbf{n}+\mathbf{e}^1)$ is idling (i.e., $a^*_{(\mathbf{n}+\mathbf{e}^1)} = I$), then

$$AC^e_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n}+\mathbf{e}^1) = (\mu_1+\mu_3)v(\mathbf{n}+\mathbf{e}^1)$$

$$AC^e_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n}) = (\mu_1+\mu_3)v(\mathbf{n})$$

If the optimal action at state $(\mathbf{n}+\mathbf{e}^1)$ is working at Station 1 (i.e., $a^*_{(\mathbf{n}+\mathbf{e}^1)} = P1$), then

$$AC^e_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n}+\mathbf{e}^1) = \mu_1 v(\mathbf{n}+\mathbf{e}^2)+\mu_3 v(\mathbf{n}+\mathbf{e}^1)$$

$$AC^e_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n}) = \mu_1 v(\mathbf{n}-\mathbf{e}^1+\mathbf{e}^2)+\mu_3 v(\mathbf{n})$$

If the optimal action at state $(\mathbf{n}+\mathbf{e}^1)$ is working at Station 3 (i.e., $a^*_{(\mathbf{n}+\mathbf{e}^1)} = P3$), then

$$AC^e_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n}+\mathbf{e}^1) = \mu_1 v(\mathbf{n}+\mathbf{e}^1)+\mu_3 v(\mathbf{n}+\mathbf{e}^1-\mathbf{e}^3)$$

$$AC^e_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n}) = \mu_1 v(\mathbf{n})+\mu_3 v(\mathbf{n}-\mathbf{e}^3)$$

According to **C1**, we have $AC^e_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n}+\mathbf{e}^1) \geq AC^e_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n})$. On the other hand, $AC^e_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n})$ may not select the optimal action at state $(\mathbf{n})$ (i.e., $a^*_{(\mathbf{n}+\mathbf{e}^1)}$ may not equal $a^*_{(\mathbf{n})}$) but $AC^e_{a^*_{(\mathbf{n})}}(\mathbf{n})$ always does. Thus, we have $AC^e_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n}) \geq AC^e_{a^*_{(\mathbf{n})}}(\mathbf{n})$. Therefore, we get $AC^e_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n}+\mathbf{e}^1) \geq AC^e_{a^*_{(\mathbf{n}+\mathbf{e}^1)}}(\mathbf{n}) \geq AC^e_{a^*_{(\mathbf{n})}}(\mathbf{n})$, i.e., $AC^e_{a^*_{(\mathbf{n})}}(\mathbf{n})$ is nondecreasing in n_1. Thus, we prove property **C1** holds. Similarly, we can easily show that **C1** is preserved by operator **T** in all 8 cases.

The proof of properties **C2** and **M1-M3** are exactly the same as that of Proposition 1.2(a) (available as an On-line Appendix). Proof of Proposition 1.2(b) is very similar to the proofs in Ha (1997b) and Ha (1997a). Using the L^∞ metric, the limit of any convergent sequence of functions in Υ^e will be in Υ^e as well. Therefore, Υ^e is complete. Now, define a structure decision rules with the state-dependent thresholds:

$$n^*_{1,(n_2,n_3)} = \min\{n_1 | V^e(\mathbf{n}) \geq 0, n_1 > 0\}$$

$$n^*_{2,(n_1,n_3)} = \min\{n_2 | V^e(\mathbf{n}) \geq 0, n_2 > 0\}$$

$$n^*_{3,(n_1,n_2)} = \min\{n_3 | V^e(\mathbf{n}) \geq 0, n_3 > 0\}$$

That is, given $n_2 \geq 0$ and $n_3 \geq 0$, the decision for Server S1 is to work at Stage 1, if the number of jobs at Stage 1 (i.e., n_1) is more than $n^*_{1,(n_2,n_3)}$, or work at Stage 3 if the number of jobs at Stage 1 (i.e., n_1) is less than $n^*_{1,(n_2,n_3)}$ and $n_3 > 0$, or stay idle otherwise. Similarly, given $n_1 \geq 0$ and $n_3 \geq 0$, the decision for Server S1 is to work at Stage 1, if the number of jobs at Stage 2 (i.e., n_2) is less than $n^*_{2,(n_1,n_3)}$ and $n_1 > 0$, or work at Stage 3 if the number of jobs at Stage 2 (i.e., n_2) is more than $n^*_{2,(n_1,n_3)}$ and $n_3 > 0$, or stay idle otherwise. Finally, given $n_1 \geq 0$ and $n_2 \geq 0$, the decision for Server S1 is to work

at Stage 3, if the number of jobs at Stage 3 (i.e., n_3) is more than $n^*_{3,(n_1,n_2)}$, or work at Stage 1 if the number of jobs at Stage 3 (i.e., n_3) is less than $n^*_{3,(n_1,n_2)}$ and $n_1 > 0$, or stay idle otherwise. It can be shown that the structured decision rules satisfy the optimality equation (1.5). By Theorem 5.1 of [23], the optimal value function V^e is structured and has properties **C1**, **C2**, and **M1-M3**. Therefore, the proof of Proposition 1.2 is complete for the value function V^e.

Theorem 1.2 and Theorem 1.3 are direct results of properties presented in Proposition 1.2. Since Proposition 1.2 has been shown to hold for the value function V^e, the results of Theorem 1.2 and Theorem 1.3 will hold as well. Thus, the proof of Proposition 1.3 is complete.

□

Proof of Theorem 1.4. Taking the processing time at each Stage i, G_i, to be generally distributed, we want to show that, when $P(G_2 < G_3 < G_1) = 1$, the policy that gives priority to Stage 3 incurs less N than any other policy at any time as long as Stage 2 is not empty (i.e., $n_2 > 0$).

Consider a policy Φ that at time t_0 assigns Server S1 to Stage 1 to process m jobs, when there is at least one job at Stage 3 (i.e., $n_3 > 0$). After serving m jobs at Stage 1, Server S1 is then assigned to Stage 3 to process a job. We construct a policy Ω that mimics policy Φ until time t_0. At time t_0, Ω assigns Server S1 to Stage 1 for $m-1$ jobs then assigns Server S1 to Stage 3 to process a job, after finishing one job at Stage 3, policy Ω will then assigns Server S1 back to Stage 1 to process a job. We also define the time point after processing $m-1$ jobs at Stage 1 as time t_1. At time t_1, policy Ω assigns Server S1 to Stage 3 and policy Φ continues assigning Server S1 to Stage 1. Note that

policy Φ and policy Ω are identical and are always at the same state before time t_1. To show that giving priority to Stage 3 is optimal is equivalent to show that policy Φ is not optimal for any $m > 0$. In other words, we need to show that policy Ω results in less N than policy Φ for any $m > 0$.

Let us assume that the state of the system is $(\mathbf{n})$ at time t_1, with $n_1 > 0$, $n_2 > 0$ and $n_3 > 0$. We define $n_s = n_1 + n_2 + n_3$. Considering that we have $P(G_3 < G_1) = 1$, there are three different cases for n_2 at time t_1:

- Case 1: $n_2 = 1$,
- Case 2: $n_2 > 1$ and Server S2 is not fully utilized (i.e. is idle) under policy Ω from time t_1 to $t_1 + G_1 + G_3$,
- Case 3: $n_2 > 1$ and Server S2 never idles under policy Ω from time t_1 to $t_1 + G_1 + G_3$.

Case 1: $n_2 = 1$.

If $n_2 = 1$, at time t_1, Ω assigns Server S1 to Stage 3 and policy Φ assigns Server S1 to Stage 1. Therefore, Server S2 works under both policy Φ and policy Ω at time t_1. See table below for actions and states for both policies during each time period. The first row is the incremental time period starting at t_1 (e.g., $[G_2, G_3)$ corresponds to time $[t_1 + G_2$ to $t_1 + G_3)$, the second and third row show the actions for Server S1 and Server S2 under policy Φ and policy Ω (i.e., (1,work) means Server S1 works at Stage 1 and Server S2 works at Stage 2 and is not idle). The fourth and fifth row show the states of the system at the beginning of each time span under policy Φ and policy Ω, respectively. Finally, the last row shows the difference in N under two policies.

Time Period	$(0, G_2)$	$[G_2, G_3)$	$[G_3, G_1)$	$[G_1, G_1 + G_3)$	$[G_1 + G_3, G_1 + G_2 + G_3]$
Actions(Φ)	(1,work)	(1,idle)	(1,idle)	(3,work)	(1,idle)
Actions(Ω)	(3,work)	(3,idle)	(1,idle)	(1,idle)	(1,work)
State(Φ)	n_s	n_s	n_s	n_s	$n_s - 1$
State(Ω)	n_s	n_s	$n_s - 1$	$n_s - 1$	$n_s - 1$
$N(\Phi)$-$N(\Omega)$	0	0	1	1	0

- **Time period** $(0, G_2)$**:**

  Since $n_2 = 1$ and $P(G_2 < G_3 < G_1) = 1$, the first event for both policies is Server S2 processing the only job at Stage 2. Thus, both policy Φ and policy Ω have the same N from t_1 to $t_1 + G_2$.

- **Time period** $[G_2, G_3)$**:**

  At time $t_1 + G_2$, Server S2 finishes a job at Stage 2 (under both policy) then becomes idle. Both policy Φ and policy Ω still have the same N from $t_1 + G_2$ to $t_1 + G_3$.

- **Time period** $[G_3, G_1)$**:**

  At time $t_1 + G_3$, under policy Ω, Server S1 finishes a job at Stage 3 and starts working on a job at Stage 1 and under policy Φ, Server S1 is working on a job at Stage 1. Note that Server S1 finishes a job at Stage 3 earlier under policy Ω than finishes a job at Stage 1 under policy Φ, since $P(G_3 < G_1) = 1$. Therefore, from $t_1 + G_3$ to $t_1 + G_1$, policy Ω has 1 less N than policy Φ.

- **Time period** $[G_1, G_1 + G_3)$**:**

  At time $t_1 + G_1$, under policy Φ, Server S1 finishes the job at Stage 1. Then, she starts working at Stage 3. Under this policy, Server S2 also starts working since now there is a job to be processed at Stage 2. Under policy Ω, Server S1 is

working on a job at Stage 1 and Server S2 remains idle. Therefore, from $t_1 + G_1$ to $t_1 + G_1 + G_3$, policy Ω has still 1 less N than policy Φ.

- **Time period** $[G_1 + G_3, G_1 + G_2 + G_3]$**:**

  At time $t_1 + G_1 + G_3$, under policy Φ, Server S1 finishes a job at Stage 3 and under policy Ω, Server S1 finishes the job at Stage 1 at the same time (Notice that same sample path realization is considered for both policies). The two policies have the same N at this time. From $t_1 + G_1 + G_3$ to $t_1 + G_1 + G_2 + G_3$, both policies have the same N. At time $t_1 + G_1 + G_2 + G_3$, Server S2 finishes a job under policy Ω and the two policies reach the same state.

Policy Ω and policy Φ will have exactly the same N after time $t_1 + G_1 + G_2 + G_3$. From time t_1 to time $t_1 + G_1 + G_2 + G_3$, policy Φ has more N than policy Ω, so policy Φ cannot be optimal.

Case 2: $n_2 = k, k > 1$ and Server S2 is not fully utilized (i.e., is idle) under policy Ω from time t_1 to $t_1 + G_1 + G_3$.

Same as before, at time t_1, policy Ω assigns Server S1 to Stage 3 and policy Φ assigns Server S1 to Stage 1. Server S2 will start working under both policy Φ and policy Ω. Since $P(G_2 < G_3 < G_1) = 1$, it is possible for Server S2 to become idle under policy Ω after t_1 (i.e., while Server S1 is working at Stage 3, Server S2 may finishes up all the jobs at Stage 2). Now suppose Server S2 idles under policy Ω at time $t_1 + G_2'$, then the idle time for Server S2 under this policy is $G_1 + G_3 - G_2'$. If $P(G_2' < G_1) = 1$, this case is the same as Case 1. If $P(G_2' < G_1) = 1$, we need to study each time periods from t_1 to $t_1 + G_1 + G_2 + G_3$. Table below shows the actions and states for both policies during each time period.

Time Period	$(0, G_3)$	$[G_3, G_1)$	$[G_1, G_2')$	$[G_2', G_1 + G_3)$	$[G_1 + G_3, G_1 + G_2 + G_3]$
Actions(Φ)	(1,work)	(1,work)	(3,work)	(3,work)	(1,idle)
Actions(Ω)	(3,work)	(1,work)	(1,work)	(1,idle)	(1,work)
State(Φ)	n_s	n_s	n_s	n_s	$n_s - 1$
State(Ω)	n_s	$n_s - 1$	$n_s - 1$	$n_s - 1$	$n_s - 1$
$N(\Phi)$-$N(\Omega)$	0	1	1	1	0

- **Time period** $(0, G_3)$**:**

  At time t_1, policy Ω assigns Server S1 to Stage 3 and policy Φ assigns Server S1 to Stage 1. Server S2 starts working under both policy Φ and policy Ω. Thus, both policy Φ and policy Ω have the same N from t_1 to $t_1 + G_3$.

- **Time period** $[G_3, G_1)$**:**

  At time $t_1 + G_3$, under policy Ω, Server S1 finishes the job at Stage 3 and starts working on a job at Stage 1. Under policy Φ, Server S1 is working on a job at Stage 1. Since $P(G_3 < G_1) = 1$, from $t_1 + G_3$ to $t_1 + G_1$, policy Ω has 1 less N than policy Φ.

- **Time period** $[G_1, G_2')$**:**

  At time $t_1 + G_1$, under policy Φ, Server S1 finishes the job at Stage 1 and under policy Ω, Server S1 is still working on a job at Stage 1. Therefore, from $t_1 + G_1$ to $t_1 + G_2'$, policy Ω has still 1 less N than policy Φ.

- **Time period** $[G_2', G_1 + G_3)$**:**

  At time $t_1 + G_2'$, under policy Ω, Server S2 becomes idle. Under policy Φ, Server S2 does not become idle since there are one more job at Stage 2 under policy Φ. From $t_1 + G_1$ to $t_1 + G_2'$, policy Ω has still 1 less N than policy Φ.

- **Time period** $[G_1 + G_3, G_1 + G_2 + G_3]$**:**

  At time $t_1 + G_3 + G_1$, under policy Ω, Server S1 finishes the job at Stage 1 and

Server S2 starts working again. At time $t_1 + G_1 + G_2 + G_3$, Server S2 under policy Ω finishes the job. Under policy Φ, Server S2 works until $t_1 + G_2' + G_2$ then remains idle until $t_1 + G_1 + G_2 + G_3$. The idle time for Server S2 under policy Φ is $t_1 + G_1 + G_2 + G_3 - (t_1 + G_2' + G_2) = G_1 + G_3 - G_2'$, which is the same as idle time of Server S2 under policy Ω.

Similar to case 1, policy Ω and Φ will have exactly the same N after time $t_1 + G_1 + G_2 + G_3$. From time t_1 to time $t_1 + G_1 + G_2 + G_3$, policy Φ has more N than policy Ω, so policy Φ cannot be optimal.

Case 3: $n_2 = k, k > 1$ and Server S2 never idles under policy Ω from time t_1 to $t_1 + G_1 + G_3$.

In this case, Server S2 never idles under policy Ω from time t_1 to $t_1 + G_1 + G_3$. If this is true, then we want to show she also never idles under policy Φ since Stage 2 under policy Φ always has the same or more N than Stage 2 under policy Ω from time t_1 to $t_1 + G_1 + G_3$. Table below shows the actions and states for both policies during each time period in this time span.

Time Period	$(0, G_3)$	$[G_3, G_1)$	$[G_1, G_1 + G_3]$
Actions(Φ)	(1,work)	(1,work)	(3,work)
Actions(Ω)	(3,work)	(1,work)	(1,work)
State(Φ)	n_s	n_s	n_s
State(Ω)	n_s	$n_s - 1$	$n_s - 1$
$N(\Phi)$-$N(\Omega)$	0	1	1

- **Time period $(0, G_3)$:**

  At time t_1, policy Ω assigns Server S1 to Stage 3 and policy Φ assigns Server S1

to Stage 1. Server S2 starts working under both policy Φ and policy Ω. Thus, both policy Φ and policy Ω have the same N from t_1 to $t_1 + G_3$.

- **Time period** $[G_3, G_1)$**:**

  At time $t_1 + G_3$, under policy Ω, Server S1 finishes the job at Stage 3 and starts working on a job at Stage 1. From $t_1 + G_3$ to $t_1 + G_1$, policy Ω has 1 less N than policy Φ.

- **Time period** $[G_1, G_1 + G_3]$**:**

  At time $t_1 + G_1$, under policy Φ, Server S1 finishes the job at Stage 1. Then, she starts working at Stage 3. At time $t_1 + G_3 + G_1$, policy Φ finishes the job at Stage 3 and Ω finishes the job at Stage 1 (at the same time). The two policies reach the same state at this time.

From time $t_1 + G_3 + G_1$, policy Ω will mimic policy Φ. Policy Ω and policy Φ will have exactly the same N after time $t_1 + G_1 + G_3$. From time t_1 to time $t_1 + G_1 + G_3$, policy Φ has more N than policy Ω, so policy Φ cannot be optimal.

Therefore any policy Φ with $m > 0$ has more N than policy Ω, so the optimal dynamic policy must have $m = 0$, which means the optimal dynamic policy should not assign Server S1 to Stage 1 as long as there is at least one job at Stage 3. Therefore, the policy which gives priority to Stage 3 is optimal under conditions in this theorem. □

Extended Numerical Analysis for Robustness Check

Robustness of Optimal NIT Policy

In this section, we check the robustness of optimal NIT policy with respect to the errors in setting the optimal thresholds (R_1^*, R_3^*) by running additional numerical studies. We

recompute the long-run average number of jobs in the system when a threshold is set 10% and 20% below or above the optimal thresholds, while keeping the other threshold fixed. This allows us to isolate the effect of change in one threshold on the performance of the policy without being influenced by the change in the other threshold. We then compare the performance of the policy with sub-optimal thresholds with that of the policy with the optimal thresholds as well as with that of the optimal dynamic policy. Table A.1 summarizes the robustness check results on thresholds used for optimal NIT policy. The top part of Table A.1 shows the the performance of the policy with sub-optimal thresholds with that of the same policy with optimal thresholds. The bottom part of Table A.1 shows the the performance of the policy with sub-optimal thresholds with that of the optimal dynamic policy (obtained using MDP).

Table A.1. Summary of Robustness Analysis on Threshold (R_1, R_3) for NIT Policy

	R_1				R_3			
Item	20% below	10% below	10% above	20% above	20% below	10% below	10% above	20% above
	Compared to NIT policy with Optimal Thresholds (R_1^*, R_3^*)							
Average PL	0.0%	0.0%	0.0%	0.0%	0.5%	0.2%	0.2%	0.4%
Max PL	0.6%	0.6%	0.4%	0.4%	6.5%	3.0%	4.9%	7.1%
% of cases with $PL < 10\%$	100%	100%	100%	100%	100%	100%	100%	100%
% of cases with $PL < 5\%$	100%	100%	100%	100%	96%	100%	100%	99%
	Compared to Optimal Dynamic Policy							
Average PL	3.2%	3.2%	3.1%	3.2%	3.7%	3.3%	3.3%	3.5%
Max PL	12.7%	12.7%	12.7%	12.7%	16.6%	14.1%	14.1%	19.2%
% of cases with $PL < 10\%$	95%	95%	95%	95%	93%	95%	93%	93%
% of cases with $PL < 5\%$	77%	77%	77%	77%	73%	77%	77%	77%

Robustness of PIT Policy

In this section, we perform a similar robustness analysis the robustness of optimal PIT policy with respect to the errors in setting the optimal thresholds $(N_2^*, Z_1^*, Z_3^*, S_1^*, S_3^*)$ by running additional numerical studies. Similarly, we recompute the long-run average number of jobs in the system when a threshold is set 10% and 20% below or above the

optimal thresholds, while keeping the other thresholds fixed. Again, this allows us to isolate the effect of change in one threshold on the performance of the policy without being influenced by the change in the other thresholds.

Table A.2. Summary of Robustness Analysis on Threshold N_2 for PIT Policy

	R_1			
Item	20% below	10% below	10% above	20% above
	Compared to NIT policy with Optimal Threshold N_2^*			
Average PL	0.9%	0.8%	0.4%	0.5%
Max PL	11.3%	11.3%	2.2%	2.2%
% of cases with $PL < 10\%$	99%	99%	100%	100%
% of cases with $PL < 5\%$	96%	98%	100%	100%
	Compared to Optimal Dynamic Policy			
Average PL	1.5%	1.3%	1.0%	1.0%
Max PL	15.3%	15.3%	3.9%	3.9%
% of cases with $PL < 10\%$	99%	99%	100%	100%
% of cases with $PL < 5\%$	94%	95%	100%	100%

Table A.3. Summary of Robustness Analysis on Thresholds (Z_1, Z_3) for PIT Policy

	Z_1				Z_3			
Item	20% below	10% below	10% above	20% above	20% below	10% below	10% above	20% above
	Compared to NIT policy with Optimal Thresholds (Z_1^*, Z_3^*)							
Average PL	0.0%	0.0%	1.3%	1.9%	0.6%	0.2%	0.1%	0.1%
Max PL	0.0%	0.0%	6.9%	7.0%	7.0%	3.5%	1.2%	1.2%
% of cases with $PL < 10\%$	100%	100%	100%	100%	100%	100%	100%	100%
% of cases with $PL < 5\%$	100%	100%	90%	89%	96%	100%	100%	100%
	Compared to Optimal Dynamic Policy							
Average PL	0.6%	0.6%	1.9%	1.9%	1.1%	0.8%	0.6%	0.7%
Max PL	3.6%	3.6%	7.0%	7.0%	7.3%	4.4%	3.8%	4.1%
% of cases with $PL < 10\%$	100%	100%	100%	100%	100%	100%	100%	100%
% of cases with $PL < 5\%$	100%	100%	89%	89%	93%	100%	100%	100%

Table A.4. Summary of Robustness Analysis on Thresholds (S_1, S_3) for PIT Policy

	S_1				S_3			
Item	20% below	10% below	10% above	20% above	20% below	10% below	10% above	20% above
	Compared to NIT policy with Optimal Thresholds (S_1^*, S_3^*)							
Average PL	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.0%	2.0%
Max PL	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	9.6%	9.6%
% of cases with $PL < 10\%$	100%	100%	100%	100%	100%	100%	100%	100%
% of cases with $PL < 5\%$	100%	100%	100%	100%	100%	100%	90%	90%
	Compared to Optimal Dynamic Policy							
Average PL	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	2.6%	2.6%
Max PL	3.6%	3.6%	3.6%	3.6%	3.6%	3.6%	9.8%	9.8%
% of cases with $PL < 10\%$	100%	100%	100%	100%	100%	100%	100%	100%
% of cases with $PL < 5\%$	100%	100%	100%	100%	100%	100%	89%	89%

Non-exponential Service Times

In this section, we present the details of numerical studies designed to test the performance of our proposed policies when service time at Stage 2 has a non-exponential distribution.

Consider the MDP model presented in the chapter, except than the service time at Stage 2 is not exponentially distributed. To analyze this system, we discretize the time horizon into equal, nonoverlapping infinitesimal intervals δt, where $\delta t \to 0$. At the beginning of each interval, Server S1 decides whether to process (or continue processing) a job at Stage 1 or at Stage 3 or remain idle, using the system's state information, which includes the number of jobs at each Stage i, n_i, and the time-interval index t_2 (in units of δt) since Server S2's last type-2 job processing started. Once the Server S1 decides to process a job, the service process at Stage 1 is a Poisson process with service rate μ_1, independent of the arrival process and service process at other stages. Similarly, the service process at Stage 3 is a Poisson process with service rate μ_3. Thus, in each period (time interval δt), the probability that one job is processed at Stage i is close to $\mu_i \delta t$. Similarly, a job may arrive at Stage 1 during this period, and the probability of having one arrival at Stage 1 in each period is close to $\lambda \delta t$ as $\delta t \to 0$. The service time at Stage 2 follows an independent random variable with gamma(α, β) distribution. We let δt be small enough so that the followings are true:

- The probability that Server S2 processes a type-2 job in period $[t_2, t_2 + 1]$ is close to $\phi_2(t_2 \delta t)\delta t$, where t_2 is the total number of time intervals (in length of δt) elapsed from when Server S2's last started to process a job, and $\phi_2(t)$ is the *hazard function* of gamma(α, β).

- The probability that Server S1 processes more than one job (either type-1 or type-3) during the time interval of length δt is almost zero.
- The probability of arrival of more than one job at Stage 1 during an in interval of length δt is almost zero.

Based on these assumptions, we can develop an MDP model with state space $\mathscr{U} = \{(n_1, n_2, n_3, t_2) | n_i \geq 0 \ \forall i, t_2 \geq 0\}$, action space $\mathscr{A} = \{I, P1, P3\}$, as we defined in Section 4, and decision epochs being the beginning of each period.

Let $\eta_{(i,j)}(t_2)$ be the joint probability that during period $[t_2, t_2+1]$ (one unit of δt), i jobs arrive at Stage 1, and j jobs are processed at Stage 2. As an example, $\eta_{(1,0)}(t_2)$ represents the probability that in period $[t_2, t_2 + 1]$, a job arrives at Stage 1 and no departure from Stage 2. Therefore, when $\delta t \to 0$,

$$\eta_{(0,0)}(t_2) = [1 - \lambda\delta t][1 - \phi_2(t_2\delta t)\delta t]$$

$$\eta_{(1,0)}(t_2) = [\lambda\delta t][1 - \phi_2(t_2\delta t)\delta t]$$

$$\eta_{(0,1)}(t_2) = [1 - \lambda\delta t][\phi_2(t_2\delta t)\delta t]$$

$$\eta_{(1,1)}(t_2) = [\lambda\delta t][\phi_2(t_2\delta t)\delta t]$$

and $\eta_{(i,j)}(t_2)$ for all $i, j \geq 2$. Thus, we get $\sum_{i,j} \eta_{(i,j)}(t_2) = 1$. The optimality equation under the long-run average number of jobs in the system criterion is

$$\begin{aligned} g\delta t + V(n_1, n_2, n_3, t_2) = \Big\{ & (n_1 + n_2 + n_3)\delta t + \min \Big\{ m(n_1, n_2, n_3, t_2), \\ & (\mu_1\delta t) m([n_1 - 1]^+, n_2, n_3, t_2) + (1 - \mu_1\delta t) m(n_1, n_2, n_3, t_2), \\ & (\mu_3\delta t) m(n_1, n_2, [n_3 - 1]^+, t_2) + (1 - \mu_3\delta t) m(n_1, n_2, n_3, t_2) \Big\} \Big\}, \end{aligned}$$

where $\forall n_1, n_2, n_3, t_2 \geq 0$ we have

$$m(n_1, n_2, n_3, t_2) = \eta_{(0,0)}(t_2)[pV(n_1, n_2, n_3, t_2 + 1) + (1-p)V(n_1, n_2, n_3, t_2 + 1)]$$

$$+ \eta_{(1,0)}(t_2)[pV(n_1 + 1, n_2, n_3, t_2 + 1) + (1-p)V(n_1 + 1, n_2, n_3, t_2 + 1)]$$

$$+ \eta_{(0,1)}(t_2)[pV(n_1, [n_2 - 1]^+, n_3 + 1, 0) + (1-p)V(n_1, [n_2 - 1]^+, n_3, 0)]$$

$$+ \eta_{(1,1)}(t_2)[pV(n_1 + 1, [n_2 - 1]^+, n_3 + 1, 0) + (1-p)V(n_1 + 1, [n_2 - 1]^+, n_3, 0)],$$

and g is the optimal average cost per unit time and $[x]^+ = \max(x, 0)$. Table 2.1 shows the performance of the optimal static, optimal NIT and optimal PIT when CV is 0.5 or 2.

Table A.5. The performance of the optimal static, optimal NIT and optimal PIT when CV = 0.5 or 2

	optimal static policy		optimal NIT policy		optimal PIT policy	
	CV = 0.5	CV = 2	CV = 0.5	CV = 2	CV = 0.5	CV = 2
Average PL	7.3%	16.6%	4.4%	10.4%	0.9%	2.0%
Max PL	81%	91.5%	56%	73.9%	5.6%	9.2%
% of cases with $PL < 10\%$	80%	58%	84%	68%	100%	100%
% of cases with $PL < 5\%$	69%	41%	74%	52%	98%	84%

Our numerical study includes a total of 81 scenarios generated according to the range of parameters presented in Table 2.1. We calculate μ_2 as described in Section 5.3 and then set the parameters of the Gamma distributed service time at Stage 2 such that the coefficient of variation is equal to 0.5 and 2. We considered a cycle of length $T = 10$, which is discretized into $T \geq 1000$ periods of length $\delta t \leq 0.01$. Due to the large number of states in our MDP model (since T is a very large number, i.e., $T = 1000$), we had to truncate the number of jobs in each stage in our numerical study. Table A.5 summarizes the performance of optimal static, optimal NIT and optimal PIT policies compared to the optimal dynamic policy under full information.

APPENDIX B

Appendix of Chapter 2: Proof of analytical results

PROOF OF THEOREM 2.1: To show that there exists a stationary average-cost optimal policy for MDP problem presented with the optimality equation (2.1), we must show that: (i) state space $\mathcal{S}$ is finite; (ii) action set $\mathcal{A}$ is finite; (iii) per-unit transition costs incurred at any state for any allowable action are bounded; and (iv) the MDP model is unichain (Theorem 8.4.5, Puterman [21]).

Clearly, our problem satisfies condition (i) and (ii). To show that condition (iii) holds for our problem, let $c(a, \mathbf{n})$ denote the cost incurred per unit transition at state $\mathbf{n} \in \mathcal{S}$ for allowable action $a \in A_{\mathbf{n}}$. Thus, $c(a, \mathbf{n}) \leq \Gamma N^T / \Lambda < \infty$, i.e., per-unit transition costs are finite, where Γ is the row vector of $b_i\theta_i$'s and N is the row vector of N_i's, where N_i is the maximum limit for the number of type-i customers allowed in the system.

To show that out MDP is *unichain* for every deterministic policy, we need to show that it consists of a single recurrent class plus a possibly empty set of transient states. Since the state space $\mathcal{S}$ is finite in our MDP, there exists at least one positive recurrent class. We can show by contradiction that there is only *one* positive recurrent class. Suppose that there exists more than one positive recurrent class, $R_1, R_2, \cdots, R_K$. Considering state N as defined above, note that every state $\mathbf{n} \in \mathcal{S}$ leads to state N. If N is transient, it leads to a closed class R_k for some k. But since every state in R_k also leads to N, then R_k cannot be closed and this is a contradiction. Thus, N belongs to a positive recurrent class. Without loss of generality, let $N \in R_1$. With the same line of argument, we can state the R_1 is the only positive recurrent class in the model. This concludes the proof of Theorem 2.1. □

PROOF OF PROPOSITION 2.1: We use induction and value iteration algorithm to prove property **P1**. The optimality equation for the value iteration algorithm is as

follows:

$$V_{t+1}(\mathbf{n}) = \frac{1}{\Lambda}\Big\{ \sum_{i=1}^{M} b_i\theta_i n_i + \sum_{i=1}^{M} \lambda_i V_t(\mathbf{n} + \mathbf{I}^i_{\{n_i \le N_i\}}) \\ + \sum_{i=1}^{M} n_i\theta_i V_t(\mathbf{n} - \mathbf{I}^i_{\{n_i>0\}}) + \sum_{i=1}^{M}(N_i - n_i)\theta_i V_t(\mathbf{n}) + f_t(\mathbf{n})\Big\},$$

where

$$f_t(\mathbf{n}) = \min \begin{cases} \sum_{i=1}^{M} \mu_i V_t(\mathbf{n}) & \text{Idling} \\ \min_{j\in\mathcal{J}_\mathbf{n}}\{\mu_j V_t(\mathbf{n}-\mathbf{e}^j) + \sum_{i=1, i\neq j}^{M} \mu_i V_t(\mathbf{n})\} & \text{Serve} \end{cases} \tag{B.3}$$

We first show that property **P1** holds at iteration $t = 1$. Assuming **P1** holds at iteration t, we then show that it holds at iteration $t + 1$.

(P1) Iteration 1: Since $V_0(\mathbf{n}) = 0$, $\forall \mathbf{n} \in \mathcal{S}$, then at iteration 1, we will have $V_1(\mathbf{n}) = \frac{1}{\Lambda}(\sum_i b_i\theta_i n_i)$ for all $\mathbf{n} \in \mathcal{S}$, by (B.3). For type-$j$ customer when $j \in \mathcal{J}_\mathbf{n}$, we have:

$$\begin{aligned} \mathbf{D}_j V_1(\mathbf{n}) &= V_1(\mathbf{n}) - V_1(\mathbf{n} - \mathbf{e}^j) \\ &= \frac{1}{\Lambda}\left[b_j\theta_j n_j + \sum_{i=1, i\neq j}^{M} b_i\theta_i n_i\right] - \frac{1}{\Lambda}\left[b_j\theta_j(n_j - 1) + \sum_{i=1, i\neq j}^{M} b_i\theta_i n_i\right] \\ &= \frac{b_j\theta_j}{\Lambda}. \end{aligned}$$

Thus, property **P1** holds since $0 \le \frac{b_j\theta_j}{\Lambda} = \mathbf{D}_j V_1(\mathbf{n}) \le b_j$, $\forall \mathbf{n} \in \mathcal{S}$ and $j \in \mathcal{J}_\mathbf{n}$.

(P1) Iteration t: We assume that property **P1** holds at iteration t. That is:

$$0 < \frac{b_j\theta_j}{\Lambda} \le \mathbf{D}_j V_t(\mathbf{n}) \le b_j, \quad \forall \mathbf{n} \in \mathcal{S} \quad \text{and} \quad j \in \mathcal{J}_\mathbf{n}. \tag{B.4}$$

(P1) Iteration $t+1$: We complete the proof by showing that property **P1** holds at iteration $t+1$. By the optimality equation (B.3), we have:

(B.5)

$$\begin{aligned}
\mathbf{D}_j V_{t+1}(\mathbf{n}) &= V_{t+1}(\mathbf{n}) - V_{t+1}(\mathbf{n}-\mathbf{e}^j) \\
&= \frac{1}{\Lambda}\Big\{ b_j\theta_j + \sum_{i=1}^{M} \lambda_i \left[V_t(\mathbf{n}+\mathbf{I}^i_{\{n_i<N_i\}}) - V_t(\mathbf{n}-\mathbf{e}^j+\mathbf{I}^i_{\{n_i<N_i\}}) \right] \\
&\quad + \sum_{i=1}^{M} n_i\theta_i \left[V_t(\mathbf{n}-\mathbf{I}^i_{\{n_i>0\}}) - V_t(\mathbf{n}-\mathbf{e}^j-\mathbf{I}^i_{\{n_i>0\}}) \right] \\
&\quad + \sum_{i=1}^{M} (N_i-n_i)\theta_i \left[V_t(\mathbf{n}) - V_t(\mathbf{n}-\mathbf{e}^j) \right] \\
&\quad - \theta_j \left[V_t(\mathbf{n}-\mathbf{e}^j) - V_t(\mathbf{n}-\mathbf{2e}^j) \right] + \left[f_t(\mathbf{n}) - f_t(\mathbf{n}-\mathbf{e}^j) \right] \Big\} \\
&= \frac{1}{\Lambda}\Big\{ b_j\theta_j + \sum_{i=1}^{M} \lambda_i \mathbf{D}_j V_t(\mathbf{n}+\mathbf{I}^i_{\{n_i<N_i\}}) \\
&\quad + \sum_{i=1}^{M} n_i\theta_i \mathbf{D}_j V_t(\mathbf{n}-\mathbf{I}^i_{\{n_i>0\}}) + \sum_{i=1}^{M} (N_i-n_i)\theta_i \mathbf{D}_j V_t(\mathbf{n}) - \theta_j \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j) \\
&\quad + \mathbf{D}_j f_t(\mathbf{n}) \Big\}
\end{aligned}$$

where $\mathbf{D}_j f_t(\mathbf{n}) = f_t(\mathbf{n}) - f_t(\mathbf{n}-\mathbf{e}^j)$, $\forall \mathbf{n} \in \mathcal{S}$ and $j \in \mathcal{J}_{\mathbf{n}}$.

The term $\mathbf{D}_j f_t(\mathbf{n})$ depends on the optimal policy at states $\mathbf{n}$ and $\mathbf{n}-\mathbf{e}^j$. By induction assumption (B.4), we know that idling is optimal if there is no customer in system. Therefore, there are 3 cases to consider.

CASE 1: $n_j = 1$ and $n_k = 0$, $\forall k \neq j$ at state $\mathbf{n}$, then serving type j at state $\mathbf{n}$ and idling at $\mathbf{n}-\mathbf{e}^j = (0,0,\ldots,0)$ are the only options.

CASE 2: Serving the same customer of type $z \in \{1,2,\ldots,M\}$ is optimal at both states $\mathbf{n}$ and $\mathbf{n}-\mathbf{e}^j$.

CASE 3: Serving different customer of types $z, l \in \{1, 2, \ldots, M\}$ is optimal at states $\mathbf{n}$ and $\mathbf{n} - \mathbf{e}^j$, respectively.

<u>*CASE 1:*</u> In this case, $n_j = 1$ and $n_k = 0$, $\forall k \neq j$. Considering serving type-j and idling as the only options at states $\mathbf{n}$ and $\mathbf{n} - \mathbf{e}^j$, respectively, we have:

$$\begin{aligned}
\mathbf{D}_j f_t(\mathbf{n}) &= \sum_{i=1, i\neq j}^{M} \mu_i \mathbf{D}_j V_t(\mathbf{n}) \\
\mathbf{D}_j V_{t+1}(\mathbf{n}) &= \frac{1}{\Lambda}\Big\{ b_j\theta_j + \sum_{i=1}^{M} \lambda_i \mathbf{D}_j V_t(\mathbf{n} + \mathbf{I}^i_{\{n_i < N_i\}}) \\
&\quad + \sum_{i=1}^{M} n_i\theta_i \mathbf{D}_j V_t(\mathbf{n} - \mathbf{I}^i_{\{n_i > 0\}}) + \sum_{i=1}^{M} (N_i - n_i)\theta_i \mathbf{D}_j V_t(\mathbf{n}) - \theta_j \mathbf{D}_j V_t(\mathbf{n} - \mathbf{e}^j) \\
&\quad + \sum_{i=1, i\neq j}^{M} \mu_i \mathbf{D}_j V_t(\mathbf{n}) \Big\}.
\end{aligned}$$

By induction assumption (B.4), we have:

$$\begin{aligned}
\mathbf{D}_j V_{t+1}(\mathbf{n}) &\leq \frac{1}{\Lambda}\Big\{ b_j\theta_j + \sum_{i=1}^{M} \lambda_i b_j + \sum_{i=1}^{M} n_i\theta_i b_j + \sum_{i=1}^{M} (N_i - n_i)\theta_i b_j - \theta_j b_j + \sum_{i=1, i\neq j}^{M} \mu_i b_j \Big\} \\
&= \frac{1}{\Lambda}\left[\sum_{i=1}^{M} \lambda_i + \sum_{i=1}^{M} N_i\theta_i + \sum_{i=1, i\neq j}^{M} \mu_i \right] b_j = \left[\frac{\Lambda - \mu_j}{\Lambda} \right] b_j \leq b_j,
\end{aligned}$$

since $\Lambda = \sum_{i=1}^{M} (\lambda_i + \mu_i + N_i\theta_i)$. Similarly,

$$\begin{aligned}
\mathbf{D}_j V_{t+1}(\mathbf{n}) &\geq \frac{1}{\Lambda}\Big\{ b_j\theta_j + \left[\sum_{i=1}^{M} \lambda_i + \sum_{i=1}^{M} n_i\theta_i + \sum_{i=1}^{M} (N_i - n_i)\theta_i - \theta_j + \sum_{i=1, i\neq j}^{M} \mu_i \right] \frac{b_j\theta_j}{\Lambda} \Big\} \\
&= \frac{1}{\Lambda}\Big\{ b_j\theta_j + \left[\sum_{i=1}^{M} \lambda_i + \sum_{i=1}^{M} N_i\theta_i - \theta_j + \sum_{i=1, i\neq j}^{M} \mu_i \right] \frac{b_j\theta_j}{\Lambda} \Big\} \\
&= \frac{1}{\Lambda}\Big\{ b_j\theta_j + \left[\Lambda - (\theta_j + \mu_i) \right] \frac{b_j\theta_j}{\Lambda} \Big\} \geq \frac{b_j\theta_j}{\Lambda}.
\end{aligned}$$

CASE 2: In this case, serving the same customer of type $z \in \{1, 2, \ldots, M\}$ is optimal at both states $\mathbf{n}$ and $\mathbf{n} - \mathbf{e}^j$. Thus, by (B.3) and (B.5), we have:

(B.6)

$$\begin{aligned}
\mathbf{D}_j f_t(\mathbf{n}) &= \left[\mu_z V(\mathbf{n} - \mathbf{e}^z) + \sum_{i=1, i\neq z}^{M} \mu_i V(\mathbf{n})\right] - \left[\mu_z V(\mathbf{n} - \mathbf{e}^z - \mathbf{e}^j) + \sum_{i=1, i\neq z}^{M} \mu_i V(\mathbf{n} - \mathbf{e}^j)\right] \\
&= \mu_z \mathbf{D}_j V(\mathbf{n} - \mathbf{e}^z) + \sum_{i=1, i\neq z}^{M} \mu_i \mathbf{D}_j V_t(\mathbf{n}) \\
\mathbf{D}_j V_{t+1}(\mathbf{n}) &= \frac{1}{\Lambda}\Big\{ b_j\theta_j + \sum_{i=1}^{M} \lambda_i \mathbf{D}_j V_t(\mathbf{n} + \mathbf{I}^i_{\{n_i < N_i\}}) \\
&\quad + \sum_{i=1}^{M} n_i\theta_i \mathbf{D}_j V_t(\mathbf{n} - \mathbf{I}^i_{\{n_i > 0\}}) + \sum_{i=1}^{M} (N_i - n_i)\theta_i \mathbf{D}_j V_t(\mathbf{n}) \\
&\quad - \theta_j \mathbf{D}_j V_t(\mathbf{n} - \mathbf{e}^j) + \mu_z \mathbf{D}_j V(\mathbf{n} - \mathbf{e}^z) + \sum_{i=1, i\neq z}^{M} \mu_i \mathbf{D}_j V_t(\mathbf{n})\Big\}
\end{aligned}$$

By induction assumption (B.4), we have:

$$\begin{aligned}
\mathbf{D}_j V_{t+1}(\mathbf{n}) &\leq \frac{1}{\Lambda}\Big\{ b_j\theta_j + \sum_{i=1}^{M} \lambda_i b_j + \sum_{i=1}^{M} n_i\theta_i b_j + \sum_{i=1}^{M} (N_i - n_i)\theta_i b_j - \theta_j b_j + \mu_z b_j + \sum_{i=1, i\neq z}^{M} \mu_i b_j\Big\} \\
&= \frac{1}{\Lambda}\left[\sum_{i=1}^{M} \lambda_i + \sum_{i=1}^{M} N_i\theta_i + \sum_{i=1}^{M} \mu_i\right] b_j = \frac{\Lambda}{\Lambda} b_j \leq b_j
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{D}_j V_{t+1}(\mathbf{n}) &\geq \frac{1}{\Lambda}\Big\{ b_j\theta_j + \left[\sum_{i=1}^{M} \lambda_i + \sum_{i=1}^{M} n_i\theta_i + \sum_{i=1}^{M} (N_i - n_i)\theta_i - \theta_j + \sum_{i=1}^{M} \mu_i\right] \frac{b_j\theta_j}{\Lambda}\Big\} \\
&= \frac{1}{\Lambda}\Big\{ b_j\theta_j + \left[\sum_{i=1}^{M} \lambda_i + \sum_{i=1}^{M} N_i\theta_i - \theta_j + \sum_{i=1}^{M} \mu_i\right] \frac{b_j\theta_j}{\Lambda}\Big\} \\
&= \frac{1}{\Lambda}\Big\{ b_j\theta_j + [\Lambda - \theta_j] \frac{b_j\theta_j}{\Lambda}\Big\} \geq \frac{b_j\theta_j}{\Lambda}.
\end{aligned}$$

CASE 3: In this case, serving different customer of types $z, l \in \{1, 2, \ldots, M\}$ is optimal at states $\mathbf{n}$ and $\mathbf{n} - \mathbf{e}^j$, respectively.

Define $[f_t(\mathbf{n})]_x$ to be the value of function $f_t(\mathbf{n})$ when action "Serving type-x customer" is chosen at state $\mathbf{n}$. Thus, since $[f_t(\mathbf{n})]_z \leq [f_t(\mathbf{n})]_l$, we have:

$$\mathbf{D}_j f_t(\mathbf{n}) = [f_t(\mathbf{n})]_z - \left[f_t(\mathbf{n} - \mathbf{e}^j)\right]_l \leq [f_t(\mathbf{n})]_l - \left[f_t(\mathbf{n} - \mathbf{e}^j)\right]_l . \tag{B.7}$$

Note that it is always feasible to serve l at state $\mathbf{n}$ if it is feasible to serve l at $\mathbf{n} - \mathbf{e}^j$. Thus, consider $\mathbf{D}_j V_{t+1}(\mathbf{n})$ for this case,

$$\begin{aligned}\mathbf{D}_j V_{t+1}(\mathbf{n}) = \frac{1}{\Lambda}\Big\{ & b_j\theta_j + \sum_{i=1}^{M} \lambda_i \mathbf{D}_j V_t(\mathbf{n} + \mathbf{I}^i_{\{n_i<N_i\}}) \\ & + \sum_{i=1}^{M} n_i\theta_i \mathbf{D}_j V_t(\mathbf{n} - \mathbf{I}^i_{\{n_i>0\}}) + \sum_{i=1}^{M}(N_i - n_i)\theta_i \mathbf{D}_j V_t(\mathbf{n}) \\ & - \theta_j \mathbf{D}_j V_t(\mathbf{n} - \mathbf{e}^j) + [f_t(\mathbf{n})]_z - \left[f_t(\mathbf{n} - \mathbf{e}^j)\right]_l \Big\}\end{aligned}$$

Using (B.7), we get

$$\begin{aligned}\mathbf{D}_j V_{t+1}(\mathbf{n}) = \frac{1}{\Lambda}\Big\{ & b_j\theta_j + \sum_{i=1}^{M} \lambda_i \mathbf{D}_j V_t(\mathbf{n} + \mathbf{I}^i_{\{n_i<N_i\}}) \\ & + \sum_{i=1}^{M} n_i\theta_i \mathbf{D}_j V_t(\mathbf{n} - \mathbf{I}^i_{\{n_i>0\}}) + \sum_{i=1}^{M}(N_i - n_i)\theta_i \mathbf{D}_j V_t(\mathbf{n}) \\ & - \theta_j \mathbf{D}_j V_t(\mathbf{n} - \mathbf{e}^j) + [f_t(\mathbf{n})]_l - \left[f_t(\mathbf{n} - \mathbf{e}^j)\right]_l \Big\}\end{aligned}$$

where $[f_t(\mathbf{n})]_l - [f_t(\mathbf{n}-\mathbf{e}^j)]_l = \mu_l \mathbf{D}_j V(\mathbf{n}-\mathbf{e}^l) + \sum_{\forall i \neq l} \mu_i \mathbf{D}_j V_t(\mathbf{n})$, which was obtained in (B.6) in Case 2, except for z being replaced by l. Thus, we have:

$$\begin{aligned}\mathbf{D}_j V_{t+1}(\mathbf{n}) = \frac{1}{\Lambda}\Big\{ & b_j\theta_j + \sum_{i=1}^{M} \lambda_i \mathbf{D}_j V_t(\mathbf{n} + \mathbf{I}^i_{\{n_i<N_i\}}) \\ & + \sum_{i=1}^{M} n_i\theta_i \mathbf{D}_j V_t(\mathbf{n} - \mathbf{I}^i_{\{n_i>0\}}) + \sum_{i=1}^{M}(N_i - n_i)\theta_i \mathbf{D}_j V_t(\mathbf{n}) \\ & - \theta_j \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j) + \mu_l \mathbf{D}_j V(\mathbf{n}-\mathbf{e}^l) + \sum_{i=1, i\neq l}^{M} \mu_i \mathbf{D}_j V_t(\mathbf{n})\Big\}.\end{aligned}$$

The right hand side is the same as the right hand side of the second equation in (B.6) in Case 2, except for z being replaced by l. Therefore, similar to Case 2 the upper bound of **P1** holds at iteration $t+1$ for Case 3.

For the lower bound we need to consider two cases, namely $n_z = 1$ and $n_z > 1$. When $n_z = 1$, we have:

$$\mathbf{D}_j f_t(\mathbf{n}) = [f_t(\mathbf{n})]_z - \left[f_t(\mathbf{n}-\mathbf{e}^j)\right]_l \geq [f_t(\mathbf{n})]_l - \left[f_t(\mathbf{n}-\mathbf{e}^j)\right]_{idling},$$

since idling has higher cost than serving l at state $\mathbf{n}-\mathbf{e}^j$. Note that serving serving l is optimal at state $\mathbf{n}-\mathbf{e}^j$. Since the actions at states $\mathbf{n}$ and $\mathbf{n}-\mathbf{e}^j$ are serving a customer and idling, respectively, the proof for the lower bound if analogous to the proof for lower bound in Case 1. Note that the proof for Case 1 is general and does require that $n_j = 0, \forall j \in \mathcal{J}_{\mathbf{n}}$.

When $n_z > 1$, we will have:

$$\mathbf{D}_j f_t(\mathbf{n}) = [f_t(\mathbf{n})]_z - \left[f_t(\mathbf{n}-\mathbf{e}^j)\right]_l \geq [f_t(\mathbf{n})]_z - \left[f_t(\mathbf{n}-\mathbf{e}^j)\right]_z,$$

since at state $\mathbf{n}-\mathbf{e}^j$ serving type-l customer has a lower cost than serving type-z customer. Thus, the proof for the lower bound is analogous to the proof for lower bound in Case 2. This concludes the proof for lower bound of **P1** in Case 3. This completes the proof of Proposition 2.1. □

PROOF OF PROPOSITION 2.2:

This proposition states that the optimality equation (2.1) has the following property: $V^k(\mathbf{n})-V^j(\mathbf{n})\geq 0$ is non-decreasing in n_z, *for all* $z\neq j,k$ *and* $j,k,z\in\mathcal{J}_{\mathbf{n}}$.

First, we present and prove following Lemmas.

Lemma B.1. *If* ***P2*** *holds for* $n_j,n_k\neq 0$, $\forall j,k\in\mathcal{J}_{\mathbf{n}}$, *then the following equation is non-decreasing in* n_z, $\forall z\neq j,k$ *and* $z\in\mathcal{J}_{\mathbf{n}}$*:*

$$\mu_k\left[V(\mathbf{n}-\mathbf{e}^k)-V(\mathbf{n})\right]+\mu_j\left[V(\mathbf{n})-V(\mathbf{n}-\mathbf{e}^j)\right],\ \forall j,k\in\mathcal{J}_{\mathbf{n}}\ \textit{and}\ j\neq k.$$

PROOF OF LEMMA B.1: Using equation (2.1), we have:

$$\begin{aligned}
V^k(\mathbf{n})-V^j(\mathbf{n}) =&\frac{1}{\Lambda}\Bigg\{\left[\sum_{i=1}^{M}b_i\theta_i n_i-\sum_{i=1}^{M}b_i\theta_i n_i\right]\\
&+\left[\sum_{i=1}^{M}\lambda_i V(\mathbf{n}+\mathbf{I}^i_{\{n_i\leq N_i\}})-\sum_{i=1}^{M}\lambda_i V(\mathbf{n}+\mathbf{I}^i_{\{n_i\leq N_i\}})\right]\\
&+\left[\sum_{i=1}^{M}n_i\theta_i V(\mathbf{n}-\mathbf{I}^i_{\{n_i>0\}})-\sum_{i=1}^{M}n_i\theta_i V(\mathbf{n}-\mathbf{I}^i_{\{n_i>0\}})\right]\\
&+\left[\sum_{i=1}^{M}(N_i-n_i)\theta_i V(\mathbf{n})-\sum_{i=1}^{M}(N_i-n_i)\theta_i V(\mathbf{n})\right]\\
&+\left[\mu_k V(\mathbf{n}-\mathbf{e}^k)+\sum_{i=1,i\neq k}^{M}\mu_i V(\mathbf{n})-\mu_j V(\mathbf{n}-\mathbf{e}^j)+\sum_{i=1,i\neq j}^{M}\mu_i V(\mathbf{n})\right]\Bigg\}
\end{aligned}$$

Therefore,

$$\begin{aligned}\Lambda\left[V^k(\mathbf{n}) - V^j(\mathbf{n})\right] &= \sum_{i=1}^{M} \mu_i \left[V(\mathbf{n}) - V(\mathbf{n})\right] \\ &\quad + \mu_k \left[V(\mathbf{n} - \mathbf{e}^k) - V(\mathbf{n})\right] \\ &\quad + \mu_j \left[V(\mathbf{n}) - V(\mathbf{n} - \mathbf{e}^j)\right] \\ &= \mu_k \left[V(\mathbf{n} - \mathbf{e}^k) - V(\mathbf{n})\right] + \mu_j \left[V(\mathbf{n}) - V(\mathbf{n} - \mathbf{e}^j)\right]\end{aligned}$$

Therefore, since **P2** holds, $V^k(\mathbf{n}) - V^j(\mathbf{n})$ (the left-hand-side) is non-decreasing in n_z, the right-hand-side is also non-decreasing in n_i. This completes the proof. □

Lemma B.2. *If **P2** holds for $n_j = 0$ and $n_k \neq 0$, $\forall j, k \in \mathcal{J}_{\mathbf{n}}$ and $j \neq k$, then the following equation is non-decreasing in n_z, $\forall z \neq j, k$ and $z \in \mathcal{J}_{\mathbf{n}}$:*

$$\mu_j \left[V(\mathbf{n}) - V(\mathbf{n} - \mathbf{e}^j)\right], \ \forall j \in \mathcal{J}_{\mathbf{n}} \text{ and } j \neq k.$$

PROOF OF LEMMA B.2: Using equation (2.1) and similar to Lemma B.1, we have:

$$\begin{aligned}\Lambda\left[V^k(\mathbf{n}) - V^j(\mathbf{n})\right] &= \sum_{i=1}^{M} \mu_i \left[V(\mathbf{n}) - V(\mathbf{n})\right] \\ &\quad + \mu_k \left[V(\mathbf{n}) - V(\mathbf{n})\right] \\ &\quad + \mu_j \left[V(\mathbf{n}) - V(\mathbf{n} - \mathbf{e}^j)\right] \\ &= \mu_j \left[V(\mathbf{n}) - V(\mathbf{n} - \mathbf{e}^j)\right]\end{aligned}$$

Therefore, since **P2** holds, $V^k(\mathbf{n}) - V^j(\mathbf{n})$ (the left-hand-side) is non-decreasing in n_z , the right-hand-side is also non-decreasing in n_z. This completes the proof. □

Lemma B.3. *If **P2** holds for $n_j \neq 0$ and $n_k = 0$, $\forall j, k \in \mathcal{J}_{\mathbf{n}}$ and $j \neq k$, then the following equation is non-decreasing in n_z, $\forall z \neq j, k$ and $z \in \mathcal{J}_{\mathbf{n}}$:*

$$\mu_k \left[V(\mathbf{n} - \mathbf{e}^k) - V(\mathbf{n}) \right], \ \forall k \in \mathcal{J}_{\mathbf{n}} \text{ and } k \neq j.$$

PROOF OF LEMMA B.3: Using equation (2.1), we have:

$$\begin{aligned} \Lambda \left[V^k(\mathbf{n}) - V^j(\mathbf{n}) \right] = & \sum_{i=1}^{M} \mu_i \left[V(\mathbf{n}) - V(\mathbf{n}) \right] \\ & + \mu_k \left[V(\mathbf{n} - \mathbf{e}^k) - V(\mathbf{n}) \right] \\ & + \mu_j \left[V(\mathbf{n}) - V(\mathbf{n}) \right] \\ & = \mu_k \left[V(\mathbf{n} - \mathbf{e}^k) - V(\mathbf{n}) \right] \end{aligned}$$

Therefore, since **P2** holds, $V^k(\mathbf{n}) - V^j(\mathbf{n})$ (the left-hand-side) is non-decreasing in n_z, the right-hand-side is also non-decreasing in n_z. This completes the proof. □

Proof of P2:

We use induction and value iteration algorithm to prove property **P2**.

(P2) Iteration 1: At iteration 0, $V_0(\mathbf{n}) = 0$, $\forall \mathbf{n} \in \mathcal{S}$. Therefore, **P2** holds for all $j, k, \mathbf{n}$.

(P2) Iteration t: We assume that property **P2** holds at iteration t. That is,

$$V_t^k(\mathbf{n} + \mathbf{e}^z) - V_t^j(\mathbf{n} + \mathbf{e}^z) \geq V_t^k(\mathbf{n}) - V_t^j(\mathbf{n}), \ \forall j, k \neq z \text{ and } j, k, z \in \mathcal{J}_{\mathbf{n}}. \tag{B.8}$$

(P2) Iteration $t+1$: We complete the proof by showing that property **P2** holds at iteration $t+1$. Using the definition of **P2**, we want to show:

$$V_{t+1}^k(\mathbf{n}+\mathbf{e}^z) - V_{t+1}^j(\mathbf{n}+\mathbf{e}^z) \geq V_{t+1}^k(\mathbf{n}) - V_{t+1}^j(\mathbf{n}),\ \forall j,k \neq z \text{ and } j,k,z \in \mathcal{J}_{\mathbf{n}}. \tag{B.9}$$

To prove that **P2** holds at iteration $(t+1)$, we consider three cases:

CASE 1: $n_j, n_k \neq 0$, $\forall j,k \in \mathcal{J}_{\mathbf{n}}$.

CASE 2: $n_j = 0$ and $n_k \neq 0$, $\forall j,k \in \mathcal{J}_{\mathbf{n}}$.

CASE 3: $n_j \neq 0$ and $n_k = 0$, $\forall j,k \in \mathcal{J}_{\mathbf{n}}$.

CASE 1: In this case, $n_j, n_k \neq 0$. We want to show

$$V_{t+1}^k(\mathbf{n}+\mathbf{e}^z) - V_{t+1}^j(\mathbf{n}+\mathbf{e}^z) \geq V_{t+1}^k(\mathbf{n}) - V_{t+1}^j(\mathbf{n}).$$

$$\begin{aligned} V_{t+1}^k(\mathbf{n}+\mathbf{e}^z) - V_{t+1}^j(\mathbf{n}+\mathbf{e}^z) = \frac{1}{\Lambda}\Bigg\{ & \sum_{i=1}^{M} \mu_i \left[V_t(\mathbf{n}+\mathbf{e}^z) - V_t(\mathbf{n}+\mathbf{e}^z)\right] \\ & + \mu_k \left[V_t(\mathbf{n}-\mathbf{e}^k+\mathbf{e}^z) - V_t(\mathbf{n}+\mathbf{e}^z)\right] \\ & + \mu_j \left[V_t(\mathbf{n}+\mathbf{e}^z) - V_t(\mathbf{n}-\mathbf{e}^j+\mathbf{e}^z)\right] \Bigg\} \end{aligned}$$

On the other hand,

$$\begin{aligned} V_{t+1}^k(\mathbf{n}) - V_{t+1}^j(\mathbf{n}) = \frac{1}{\Lambda}\Bigg\{ & \sum_{i=1}^{M} \mu_i \left[V_t(\mathbf{n}) - V_t(\mathbf{n})\right] \\ & + \mu_k \left[V_t(\mathbf{n}-\mathbf{e}^k) - V_t(\mathbf{n})\right] \\ & + \mu_j \left[V_t(\mathbf{n}) - V_t(\mathbf{n}-\mathbf{e}^j)\right] \Bigg\} \end{aligned}$$

According to Lemma 1, since **P2** holds at iteration t, we have

$$V_{t+1}^k(\mathbf{n}+\mathbf{e}^z) - V_{t+1}^j(\mathbf{n}+\mathbf{e}^z) \geq V_{t+1}^k(\mathbf{n}) - V_{t+1}^j(\mathbf{n})$$

This concludes the prove for Case 1.

CASE 2: In this case, $n_j = 0$ and $n_k \neq 0$. We want to show $V_{t+1}^k(\mathbf{n}+\mathbf{e}^z) - V_{t+1}^j(\mathbf{n}+\mathbf{e}^z) \geq V_{t+1}^k(\mathbf{n}) - V_{t+1}^j(\mathbf{n})$.

$$\begin{aligned} V_{t+1}^k(\mathbf{n}+\mathbf{e}^z) - V_{t+1}^j(\mathbf{n}+\mathbf{e}^z) = \frac{1}{\Lambda}\Big\{ & \sum_{i=1}^{M} \mu_i \left[V_t(\mathbf{n}+\mathbf{e}^z) - V_t(\mathbf{n}+\mathbf{e}^z)\right] \\ & + \mu_k \left[V_t(\mathbf{n}-\mathbf{e}^k+\mathbf{e}^z) - V_t(\mathbf{n}+\mathbf{e}^z)\right] \\ & + \mu_j \left[V_t(\mathbf{n}+\mathbf{e}^z) - V_t(\mathbf{n}+\mathbf{e}^z)\right] \Big\} \end{aligned}$$

On the other hand,

$$\begin{aligned} V_{t+1}^k(\mathbf{n}) - V_{t+1}^j(\mathbf{n}) = \frac{1}{\Lambda}\Big\{ & \sum_{i=1}^{M} \mu_i \left[V_t(\mathbf{n}) - V_t(\mathbf{n})\right] \\ & + \mu_k \left[V_t(\mathbf{n}-\mathbf{e}^k) - V_t(\mathbf{n})\right] \\ & + \mu_j \left[V_t(\mathbf{n}) - V_t(\mathbf{n})\right] \Big\} \end{aligned}$$

According to Lemma 2, since **P2** holds at iteration t, we have

$$V_{t+1}^k(\mathbf{n}+\mathbf{e}^z) - V_{t+1}^j(\mathbf{n}+\mathbf{e}^z) \geq V_{t+1}^k(\mathbf{n}) - V_{t+1}^j(\mathbf{n})$$

This concludes the prove for Case 2.

CASE 3: In this case, $n_j \neq 0$ and $n_k = 0$. We want to show

$$V_{t+1}^k(\mathbf{n}+\mathbf{e}^z) - V_{t+1}^j(\mathbf{n}+\mathbf{e}^z) \geq V_{t+1}^k(\mathbf{n}) - V_{t+1}^j(\mathbf{n}).$$

$$\begin{aligned} V_{t+1}^k(\mathbf{n}+\mathbf{e}^z) - V_{t+1}^j(\mathbf{n}+\mathbf{e}^z) = \frac{1}{\Lambda}\Big\{ & \sum_{i=1}^{M} \mu_i \left[V_t(\mathbf{n}+\mathbf{e}^z) - V_t(\mathbf{n}+\mathbf{e}^z)\right] \\ & + \mu_k \left[V_t(\mathbf{n}+\mathbf{e}^z) - V_t(\mathbf{n}+\mathbf{e}^z)\right] \\ & + \mu_j \left[V_t(\mathbf{n}+\mathbf{e}^z) - V_t(\mathbf{n}-\mathbf{e}^j+\mathbf{e}^z)\right] \Big\} \end{aligned}$$

On the other hand,

$$\begin{aligned} V_{t+1}^k(\mathbf{n}) - V_{t+1}^j(\mathbf{n}) = \frac{1}{\Lambda}\Big\{ & \sum_{i=1}^{M} \mu_i \left[V_t(\mathbf{n}) - V_t(\mathbf{n})\right] \\ & + \mu_k \left[V_t(\mathbf{n}) - V_t(\mathbf{n})\right] \\ & + \mu_j \left[V_t(\mathbf{n}) - V_t(\mathbf{n}-\mathbf{e}^j)\right] \Big\} \end{aligned}$$

According to Lemma 3, since **P2** holds at iteration t, we have

$$V_{t+1}^k(\mathbf{n}+\mathbf{e}^z) - V_{t+1}^j(\mathbf{n}+\mathbf{e}^z) \geq V_{t+1}^k(\mathbf{n}) - V_{t+1}^j(\mathbf{n})$$

This concludes the prove for Case 3. This completes the proof of Proposition 2.2.

□

PROOF OF THEOREM 2.2: (i) We prove by contradiction. Suppose property **P1** holds, but the optimal policy is state $\mathbf{n} \neq 0$ is to idle. Thus, according to the optimality

equation (2.1), we have

$$\sum_{i=1}^{M} \mu_i V(\mathbf{n}) < \min_{j \in \mathcal{J}_{\mathbf{n}}} \{\mu_j V(\mathbf{n} - \mathbf{e}^j) + \sum_{i=1, i \neq j}^{M} \mu_i V(\mathbf{n})\}$$

$$< \mu_j V(\mathbf{n} - \mathbf{e}^j) + \sum_{i=1, i \neq j}^{M} \mu_i V(\mathbf{n}) \quad \forall j \in \mathcal{J}_{\mathbf{n}}$$

Which means

$$\mu_j \left[V(\mathbf{n}) - V(\mathbf{n} - \mathbf{e}^j) \right] < 0 \quad \forall j \in \mathcal{J}_{\mathbf{n}}$$

$$\mathbf{D}_j V(\mathbf{n}) < 0 \quad \forall j \in \mathcal{J}_{\mathbf{n}} \quad \Longleftarrow \quad \text{Contradicts } \mathbf{P1}$$

Therefore, the optimal scheduling policy at states $\mathbf{n} \neq 0$ cannot be idling.

(ii) Theorem 2 (ii) implies that if it is optimal to serve a type-j customer in state $\mathbf{n} \in \mathcal{S}$, it is also optimal to serve a type-j customer in any state $\mathbf{n}' \in \mathcal{S}$, where $j \in \mathcal{J}_{\mathbf{n}'}$ and $\mathcal{J}_{\mathbf{n}'} \subseteq \mathcal{J}_{\mathbf{n}}$. To illustrate, suppose serving type-j customer is optimal in state $\mathbf{n}$ in which there exists at least one customer of each type. If Theorem 2 (ii) holds true, it is always optimal to serve type-j customer, regardless of the number of customers of other types in the system. This means that type-j customer has the highest priority. Now consider a state $\mathbf{n}'$ where there are no type-j customer but there exists at least one customer of other types. Without loss of generality, assume that serving type-k customer is optimal in state $\mathbf{n}'$. Thus, serving type-k customer is always optimal as long as there is no type-j customer in the system. This means that type-k customer has the second highest priority. Following the same line of argument, it is clear that the server scheduling policy is a static priority policy. To formally prove Theorem 2 (ii), we use contradiction.

Let us assume it is optimal to serve type-j customer in state $\mathbf{n}$, but the optimal action is to serve type-k customer ($k \neq j$ and $k \in \mathcal{J}_{\mathbf{n}'}$), in state $\mathbf{n}'$. Consider another state $\mathbf{n}''$, where $n_i'' = \min\{n_i, n_i'\}$ for all i.

There are three possible cases for the optimal action is state $\mathbf{n}''$:

CASE 1: Serving type-j customer is optimal in state $\mathbf{n}''$;

CASE 2: Serving type-k customer ($k \neq j$) is optimal in state $\mathbf{n}''$;

CASE 3: Serving type-z customer ($z \neq j, k$) or idling is optimal in state $\mathbf{n}''$;

CASE 1: Let $V^i(\mathbf{n})$ correspond to the value function if system is at state $\mathbf{n}$ and we decide to serve type-i customer. If serving type-j customer is optimal in state $\mathbf{n}''$, we have:

$$V^k(\mathbf{n}'') - V^j(\mathbf{n}'') \geq 0$$

On the other hand, according to **P2**, $V^k(\mathbf{n}'') - V^j(\mathbf{n}'') \geq 0$ is non-decreasing in n_i''. Thus, since $n_i' \geq n_i''$, we have $V^k(\mathbf{n}') - V^j(\mathbf{n}') \geq 0$, which implies that serving type-k customer cannot be optimal at state $\mathbf{n}'$. This is a contradiction.

CASE 2: If serving type-k customer is optimal in state $\mathbf{n}''$, we have:

$$V^j(\mathbf{n}'') - V^k(\mathbf{n}'') \geq 0$$

On the other hand, according to **P2**, $V^j(\mathbf{n}'') - V^k(\mathbf{n}'') \geq 0$ is non-decreasing in n_i''. Thus, since $n_i \geq n_i''$, we have $V^j(\mathbf{n}) - V^k(\mathbf{n}) \geq 0$, which implies that serving type-j customer cannot be optimal at state $\mathbf{n}$. This is a contradiction.

CASE 3: If serving type-z customer ($z \neq j, k$) or idling is optimal in state $\mathbf{n}''$, we have:

$$V^j(\mathbf{n}'') - V^z(\mathbf{n}'') \geq 0 \text{ and } V^k(\mathbf{n}'') - V^z(\mathbf{n}'') \geq 0$$

On the other hand, according to **P2**, $V^j(\mathbf{n}'') - V^z(\mathbf{n}'') \geq 0$ and $V^k(\mathbf{n}'') - V^z(\mathbf{n}'') \geq 0$ are non-decreasing in n_i''. Thus, since $n_i \geq n_i''$ and $n_i' \geq n_i''$, we have $V^j(\mathbf{n}) - V^z(\mathbf{n}) \geq 0$ and $V^k(\mathbf{n}) - V^z(\mathbf{n}) \geq 0$, which imply that serving type-j customer cannot be optimal at state $\mathbf{n}$ and serving type-k customer cannot be optimal at state $\mathbf{n}'$. This is a contradiction. Thus, in all three possible cases, we reached a contradiction. Therefore, it is also optimal to serve type-j customer in state $\mathbf{n}'$. This concludes the proof of Theorem 2.2. □

PROOF OF PROPOSITION 2.3

First, we present and prove following Lemma.

Lemma B.4. *If **P3** holds for $n_j > 1, n_k \geq 1$, then:*

Property **L1:** $\mathbf{D}_j V(\mathbf{n} - \mathbf{e}^j) - \mathbf{D}_j V(\mathbf{n} - \mathbf{e}^k) \geq 0,\ \forall j, k \in \mathcal{J}_\mathbf{n}$ and $j \neq k$.

PROOF OF LEMMA B.4: We use induction and value iteration algorithm to prove property **L1**.

(L1) Iteration 1: At iteration 0, $V_0(\mathbf{n}) = 0,\ \forall \mathbf{n} \in \mathcal{S}$. Therefore, $\mathbf{D}_j V_0(\mathbf{n} - \mathbf{e}^j) - \mathbf{D}_j V_0(\mathbf{n} - \mathbf{e}^k) \geq 0$.

(L1) Iteration t: We assume that property **L1** holds at iteration t. That is, for $n_j > 1$ and $n_k \geq 1$, we have:

$$\mathbf{D}_j V_t(\mathbf{n} - \mathbf{e}^j) \geq \mathbf{D}_j V_t(\mathbf{n} - \mathbf{e}^k),\ \forall j, k \in \mathcal{J}_\mathbf{n}. \tag{B.10}$$

(L1) Iteration $t+1$: We complete the proof by showing that property **L1** holds at iteration $t+1$. Using (B.5), for $n_j > 1$ and $n_k \geq 1$, we have:

$$\begin{aligned}
\mathbf{D}_j V_{t+1}(\mathbf{n}-\mathbf{e}^j) = \frac{1}{\Lambda}\Big\{ & b_j\theta_j + \sum_{i=1}^{M} \lambda_i \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j + \mathbf{I}^i_{\{n_i<N_i\}}) \\
& + \sum_{i=1,i\neq j}^{M} n_i\theta_i \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j - \mathbf{I}^i_{\{n_i>0\}}) + \sum_{i=1}^{M}(N_i-n_i)\theta_i \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j) \\
& + (n_j-1)\theta_j \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j-\mathbf{e}^j) + \mathbf{D}_j f_t(\mathbf{n}-\mathbf{e}^j)\Big\}
\end{aligned}$$

$$\begin{aligned}
\mathbf{D}_j V_{t+1}(\mathbf{n}-\mathbf{e}^k) = \frac{1}{\Lambda}\Big\{ & b_j\theta_j + \sum_{i=1}^{M} \lambda_i \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^k + \mathbf{I}^i_{\{n_i<N_i\}}) \\
& + \sum_{i=1,i\neq j}^{M} n_i\theta_i \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^k - \mathbf{I}^i_{\{n_i>0\}}) + \sum_{i=1}^{M}(N_i-n_i)\theta_i \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^k) \\
& + (n_j-1)\theta_j \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j-\mathbf{e}^k) + \mathbf{D}_j f_t(\mathbf{n}-\mathbf{e}^k)\Big\}
\end{aligned}$$

Thus, we have:

$$\begin{aligned}
& \mathbf{D}_j V_{t+1}(\mathbf{n}-\mathbf{e}^j) - \mathbf{D}_j V_{t+1}(\mathbf{n}-\mathbf{e}^k) \\
& \quad = \frac{1}{\Lambda}\Big\{ \sum_{i=1}^{M} \lambda_i \Big[\mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j + \mathbf{I}^i_{\{n_i<N_i\}}) - \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^k + \mathbf{I}^i_{\{n_i<N_i\}})\Big] \\
& \qquad + \sum_{i=1,i\neq j}^{M} n_i\theta_i \Big[\mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j - \mathbf{I}^i_{\{n_i>0\}}) - \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^k - \mathbf{I}^i_{\{n_i>0\}})\Big] \\
& \qquad + \sum_{i=1}^{M}(N_i-n_i)\theta_i \Big[\mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j) - \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^k)\Big] \\
& \qquad + (n_j-1)\theta_j \Big[\mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j-\mathbf{e}^j) - \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j-\mathbf{e}^k)\Big] \\
& \qquad + \Big[\mathbf{D}_j f_t(\mathbf{n}-\mathbf{e}^j) - \mathbf{D}_j f_t(\mathbf{n}-\mathbf{e}^k)\Big]\Big\}
\end{aligned}$$

According to the induction assumption (B.10), the first four terms are non-negative. Therefore, we only need to show:

$$\mathbf{D}_j f_t(\mathbf{n}-\mathbf{e}^j) - \mathbf{D}_j f_t(\mathbf{n}-\mathbf{e}^k) \geq 0,\ \forall j,k \in \mathcal{J}_{\mathbf{n}}.$$

There are two cases to consider: $n_j > 2$ and $n_j = 2$

CASE 1: $n_j > 2$ at state $\mathbf{n}$. In this case, when $j \in \mathcal{J}_{\mathbf{n}}$, using the result of Theorem 2 (ii) and since property **P3** holds at iteration $\mathbf{t}$, it is optimal to serve type-j customer at iteration t in all four states $\mathbf{n}-\mathbf{e}^j$, $\mathbf{n}-\mathbf{e}^k$, $\mathbf{n}-2\mathbf{e}^j$ and $\mathbf{n}-\mathbf{e}^k-\mathbf{e}^j$.

When $n_j > 2$ at state $\mathbf{n}$, using (B.3), we have:

$$\begin{aligned}
\mathbf{D}_j f_t(\mathbf{n}-\mathbf{e}^j) &= \mu_j V_t(\mathbf{n}-2\mathbf{e}^j) + \sum_{i=1,i\neq j}^{M} \mu_i V_t(\mathbf{n}-\mathbf{e}^j) \\
&\quad - \left[\mu_j V_t(\mathbf{n}-3\mathbf{e}^j) + \sum_{i=1,i\neq j}^{M} \mu_i V_t(\mathbf{n}-2\mathbf{e}^j)\right] \\
&= \mu_j \mathbf{D}_j V_t(\mathbf{n}-2\mathbf{e}^j) + \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j) \sum_{i=1,i\neq j}^{M} \mu_i
\end{aligned}$$

$$\begin{aligned}
\mathbf{D}_j f_t(\mathbf{n}-\mathbf{e}^k) &= \mu_j V_t(\mathbf{n}-\mathbf{e}^j-\mathbf{e}^k) + \sum_{i=1,i\neq j}^{M} \mu_i V_t(\mathbf{n}-\mathbf{e}^k) \\
&\quad - \left[\mu_j V_t(\mathbf{n}-2\mathbf{e}^j-\mathbf{e}^k) + \sum_{i=1,i\neq j}^{M} \mu_i V_t(\mathbf{n}-\mathbf{e}^j-\mathbf{e}^k)\right] \\
&= \mu_j \mathbf{D}_j V_t(\mathbf{n}-2\mathbf{e}^j) + \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j) \sum_{i=1,i\neq j}^{M} \mu_i
\end{aligned} \tag{B.11}$$

Therefore, by the induction assumption (B.10),

$$\begin{aligned}\mathbf{D}_j f_t(\mathbf{n}-\mathbf{e}^j) - \mathbf{D}_j f_t(\mathbf{n}-\mathbf{e}^k) &= \mu_j \mathbf{D}_j V_t(\mathbf{n}-2\mathbf{e}^j) + \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j) \sum_{i=1,i\neq j}^{M} \mu_i \\ &\quad - \left[\mu_j \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j-\mathbf{e}^k) + \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^k) \sum_{i=1,i\neq j}^{M} \mu_i\right] \\ &= \mu_j \left[\mathbf{D}_j V_t(\mathbf{n}-2\mathbf{e}^j) - \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j-\mathbf{e}^k)\right] \\ &\quad + \left[\mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j) - \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^k)\right] \sum_{i=1,i\neq j}^{M} \mu_i \geq 0\end{aligned}$$

CASE 2: $n_j = 2$ at state $\mathbf{n}$. In this case, serving type-j customer is still optimal in states $\mathbf{n}-\mathbf{e}^j$, $\mathbf{n}-\mathbf{e}^k$, and $\mathbf{n}-\mathbf{e}^k-\mathbf{e}^j$, but serving the customer of type $z \neq j$ is optimal at $\mathbf{n}-\mathbf{e}^j-\mathbf{e}^j$. Thus, $\mathbf{D}_j f_t(\mathbf{n}-\mathbf{e}^k)$ is still given by (B.11), but $\mathbf{D}_j f_t(\mathbf{n}-\mathbf{e}^j)$ after some algebra is as follows:

$$\begin{aligned}\mathbf{D}_j f_t(\mathbf{n}-\mathbf{e}^j) &= \mu_z \left[V_t(\mathbf{n}-\mathbf{e}^j) - V_t(\mathbf{n}-2\mathbf{e}^j-\mathbf{e}^z)\right] + \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j) \sum_{i=1,i\neq j,z}^{M} \mu_i \\ &= \mu_z \left[\mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j) + \mathbf{D}_z V_t(\mathbf{n}-2\mathbf{e}^j)\right] + \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j) \sum_{i=1,i\neq j,z}^{M} \mu_i \\ &= \mu_z \mathbf{D}_z V_t(\mathbf{n}-2\mathbf{e}^j) + \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j) \sum_{i=1,i\neq j}^{M} \mu_i\end{aligned}$$

Since serving type-z customer is optimal at state $\mathbf{n}-2\mathbf{e}^j$, we have $[f_t(\mathbf{n}-2\mathbf{e}^j)]_z \leq [f_t(\mathbf{n}-2\mathbf{e}^j)]_k$ for all $k \neq z$. Thus, using optimality equation (2.1), we have:

$$\mu_z \mathbf{D}_z V_t(\mathbf{n}-2\mathbf{e}^j) \geq \mu_k \mathbf{D}_k V_t(\mathbf{n}-2\mathbf{e}^j)$$

Thus, the rest of the proof for Case 2 follows exactly the same steps of the proof provided for Case 1. This concludes the proof of **L1**. □

Now, we prove Proposition 3. We use induction and value iteration algorithm to prove property **P3**.

(P3) Iteration 1: At iteration 0, $V_0(\mathbf{n}) = 0, \forall \mathbf{n} \in \mathcal{S}$. Therefore, $\mu_j \mathbf{D}_j V_0(\mathbf{n}) = \mu_k \mathbf{D}_k V_0(\mathbf{n})$.

(P3) Iteration t: We assume that property **P3** holds at iteration t. That is, for $n_j \geq 1$ and $n_k \geq 1$, we have:

$$\mu_j \mathbf{D}_j V_t(\mathbf{n}) \geq \mu_k \mathbf{D}_k V_t(\mathbf{n}), \ \forall j, k \in \mathcal{J}_{\mathbf{n}}. \tag{B.12}$$

(P3) Iteration $t+1$: We complete the proof by showing that property **P1** holds at iteration $t+1$. Using (B.5), for $n_j \geq 1$ and $n_k \geq 1$, we have:

$$\begin{aligned} \mu_j \mathbf{D}_j V_{t+1}(\mathbf{n}) = \frac{\mu_j}{\Lambda} \Big\{ & b_j \theta_j + \sum_{i=1}^{M} \lambda_i \mathbf{D}_j V_t(\mathbf{n} + \mathbf{I}^i_{\{n_i < N_i\}}) \\ & + \sum_{i=1}^{M} n_i \theta_i \mathbf{D}_j V_t(\mathbf{n} - \mathbf{I}^i_{\{n_i > 0\}}) + \sum_{i=1}^{M} (N_i - n_i) \theta_i \mathbf{D}_j V_t(\mathbf{n}) \\ & - \theta_j \mathbf{D}_j V_t(\mathbf{n} - \mathbf{e}^j) + \mathbf{D}_j f_t(\mathbf{n}) \Big\} \end{aligned} \tag{B.13}$$

$$\begin{aligned} \mu_k \mathbf{D}_k V_{t+1}(\mathbf{n}) = \frac{\mu_k}{\Lambda} \Big\{ & b_k \theta_k + \sum_{i=1}^{M} \lambda_i \mathbf{D}_k V_t(\mathbf{n} + \mathbf{I}^i_{\{n_i < N_i\}}) \\ & + \sum_{i=1}^{M} n_i \theta_i \mathbf{D}_k V_t(\mathbf{n} - \mathbf{I}^i_{\{n_i > 0\}}) + \sum_{i=1}^{M} (N_i - n_i) \theta_i \mathbf{D}_k V_t(\mathbf{n}) \\ & - \theta_k \mathbf{D}_k V_t(\mathbf{n} - \mathbf{e}^k) + \mathbf{D}_k f_t(\mathbf{n}) \Big\} \end{aligned} \tag{B.14}$$

There are 2 cases to consider: $n_j > 1$, and $n_j = 1$. Notice that it is clear that property **P3** holds for the case where $n_j = n_k = 1$.

CASE 1: $n_j > 1$ at state $\mathbf{n}$. In this case, when $j \in \mathcal{J}_{\mathbf{n}}$, by induction assumption (B.12), it is optimal to serve customer type-j at iteration t in all three states $\mathbf{n}$, $\mathbf{n}-\mathbf{e}^j$ and $\mathbf{n}-\mathbf{e}^k$.
CASE 2: $n_j = 1$ at state $\mathbf{n}$. In this case, serving type-j customer is optimal in states $\mathbf{n}$ and $\mathbf{n}-\mathbf{e}^k$, but serving the type-k customer is optimal at $\mathbf{n}-\mathbf{e}^j$.

<u>*CASE 1:*</u> When $n_j > 1$ at state $\mathbf{n}$, using (B.3), we have:

$$\begin{aligned}\mu_j \mathbf{D}_j f_t(\mathbf{n}) &= \mu_j \left[\mu_j V_t(\mathbf{n}-\mathbf{e}^j) + \sum_{i=1, i\neq j}^{M} \mu_i V_t(\mathbf{n})\right] \\ &- \mu_j \left[\mu_j V_t(\mathbf{n}-\mathbf{e}^j-\mathbf{e}^j) + \sum_{i=1, i\neq j}^{M} \mu_i V_t(\mathbf{n}-\mathbf{e}^j)\right] \\ &= \mu_j \mu_j \mathbf{D}_j V_t(\mathbf{n}-\mathbf{e}^j) + \mu_j \mathbf{D}_j V_t(\mathbf{n}) \sum_{i=1, i\neq j}^{M} \mu_i\end{aligned} \tag{B.15}$$

Similarly,

$$\begin{aligned}\mu_k \mathbf{D}_k f_t(\mathbf{n}) &= \mu_k \left[\mu_j V_t(\mathbf{n}-\mathbf{e}^j) + \sum_{i=1, i\neq j}^{M} \mu_i V_t(\mathbf{n})\right] \\ &- \mu_k \left[\mu_j V_t(\mathbf{n}-\mathbf{e}^j-\mathbf{e}^j) + \sum_{i=1, i\neq j}^{M} \mu_i V_t(\mathbf{n}-\mathbf{e}^j)\right] \\ &= \mu_j \mu_k \mathbf{D}_k V_t(\mathbf{n}-\mathbf{e}^j) + \mu_k \mathbf{D}_k V_t(\mathbf{n}) \sum_{i=1, i\neq j}^{M} \mu_i\end{aligned} \tag{B.16}$$

Considering induction assumption (B.12), when we compare (B.15) and (B.16), considering induction assumption (B.16), we can show that

$$\mu_j \mathbf{D}_j f_t(\mathbf{n}) \geq \mu_k \mathbf{D}_k f_t(\mathbf{n}),\ \forall j,k \in \mathcal{J}_{\mathbf{n}}. \tag{B.17}$$

Substituting $\mu_j \mathbf{D}_j f_t(\mathbf{n})$ with $\mu_k \mathbf{D}_k f_t(\mathbf{n})$ in (B.13) and (B.17), we get

$$\begin{aligned}\mu_j \mathbf{D}_j V_{t+1}(\mathbf{n}) \geq \frac{1}{\Lambda}\Big\{ & b_j \theta_j \mu_j + \sum_{i=1}^{M} \lambda_i \mu_j \mathbf{D}_j V_t(\mathbf{n} + \mathbf{I}^i_{\{n_i<N_i\}}) \\ & + \sum_{i=1, i\neq j}^{M} n_i \theta_i \mu_j \mathbf{D}_j V_t(\mathbf{n} - \mathbf{I}^i_{\{n_i>0\}}) + \sum_{i=1}^{M} (N_i - n_i)\theta_i \mu_j \mathbf{D}_j V_t(\mathbf{n}) \\ & + (n_j - 1)\theta_j \mu_k \mathbf{D}_k V_t(\mathbf{n} - \mathbf{e}^j) + \mu_k \mathbf{D}_k f_t(\mathbf{n})\Big\}\end{aligned} \tag{B.18}$$

Furthermore, using Lemma B.4 and considering induction assumption (B.12), we have:

$$\mu_j \mathbf{D}_j V_t(\mathbf{n} - \mathbf{e}^j) \geq \mu_j \mathbf{D}_j V_t(\mathbf{n} - \mathbf{e}^k) \geq \mu_k \mathbf{D}_k V_t(\mathbf{n} - \mathbf{e}^k)$$

Thus, we have

$$\begin{aligned}\mu_j \mathbf{D}_j V_{t+1}(\mathbf{n}) \geq \frac{1}{\Lambda}\Big\{ & b_j \theta_j \mu_j + \sum_{i=1}^{M} \lambda_i \mu_j \mathbf{D}_j V_t(\mathbf{n} + \mathbf{I}^i_{\{n_i<N_i\}}) \\ & + \sum_{i=1, i\neq j}^{M} n_i \theta_i \mu_j \mathbf{D}_j V_t(\mathbf{n} - \mathbf{I}^i_{\{n_i>0\}}) + \sum_{i=1}^{M} (N_i - n_i)\theta_i \mu_j \mathbf{D}_j V_t(\mathbf{n}) \\ & + (n_j - 1)\theta_j \mu_k \mathbf{D}_k V_t(\mathbf{n} - \mathbf{e}^k) + \mu_k \mathbf{D}_k f_t(\mathbf{n})\Big\}\end{aligned} \tag{B.19}$$

Therefore,

$$\begin{aligned}\mu_j \mathbf{D}_j V_{t+1}(\mathbf{n}) - \frac{b_j \theta_j \mu_j}{\Lambda} + \frac{\theta_j}{\Lambda} \mu_k \mathbf{D}_k V_t(\mathbf{n} - \mathbf{e}^k) \geq \frac{1}{\Lambda}\Big\{ & \sum_{i=1}^{M} \lambda_i \mu_j \mathbf{D}_j V_t(\mathbf{n} + \mathbf{I}^i_{\{n_i<N_i\}}) \\ & + \sum_{i=1}^{M} n_i \theta_i \mu_k \mathbf{D}_k V_t(\mathbf{n} - \mathbf{I}^i_{\{n_i>0\}}) \\ & + \sum_{i=1}^{M} (N_i - n_i)\theta_i \mu_j \mathbf{D}_j V_t(\mathbf{n}) \\ & + \mu_k \mathbf{D}_k f_t(\mathbf{n})\Big\}\end{aligned} \tag{B.20}$$

Similarly, using (B.16), we have

(B.21)

$$\mu_k \mathbf{D}_k V_{t+1}(\mathbf{n}) - \frac{b_k \theta_k \mu_k}{\Lambda} + \frac{\theta_k}{\Lambda} \mu_k \mathbf{D}_k V_t(\mathbf{n} - \mathbf{e}^k) = \frac{1}{\Lambda} \Big\{ \sum_{i=1}^{M} \lambda_i \mu_k \mathbf{D}_k V_t(\mathbf{n} + \mathbf{I}^i_{\{n_i < N_i\}}) + \sum_{i=1}^{M} n_i \theta_i \mu_k \mathbf{D}_k V_t(\mathbf{n} - \mathbf{I}^i_{\{n_i > 0\}}) + \sum_{i=1}^{M} (N_i - n_i) \theta_i \mu_k \mathbf{D}_k V_t(\mathbf{n}) + \mu_k \mathbf{D}_k f_t(\mathbf{n}) \Big\}$$

Comparing (B.20) and (B.21), and considering (B.12), we conclude

$$\mu_j \mathbf{D}_j V_{t+1}(\mathbf{n}) - \frac{b_j \theta_j \mu_j}{\Lambda} + \frac{\theta_j}{\Lambda} \mu_k \mathbf{D}_k V_t(\mathbf{n} - \mathbf{e}^k) \geq \mu_k \mathbf{D}_k V_{t+1}(\mathbf{n}) - \frac{b_k \theta_k \mu_k}{\Lambda} + \frac{\theta_k}{\Lambda} \mu_k \mathbf{D}_k V_t(\mathbf{n} - \mathbf{e}^k)$$

$$\text{or } \mu_j \mathbf{D}_j V_{t+1}(\mathbf{n}) - \mu_k \mathbf{D}_k V_{t+1}(\mathbf{n}) \geq \frac{1}{\Lambda} \Big\{ b_j \theta_j \mu_j - b_k \theta_k \mu_k + (\theta_k - \theta_j) \mu_k \mathbf{D}_k V_t(\mathbf{n} - \mathbf{e}^k) \Big\}$$

Therefore, to prove **P3** holds at iteration $t+1$ in Case 1, we only need to show that

(B.22) $$b_j \theta_j \mu_j - b_k \theta_k \mu_k + (\theta_k - \theta_j) \mu_k \mathbf{D}_k V_t(\mathbf{n} - \mathbf{e}^k) \geq 0$$

If **C1** holds for customer-types j and k, then we have two cases:

(1) $b_j \theta_j \mu_j > b_k \theta_k \mu_k$ and $\theta_k > \theta_j$. In this case, since $\mathbf{D}_k V_t(\mathbf{n} - \mathbf{e}^k) \geq \frac{b_k \theta_k}{\Lambda} > 0$, we have $b_j \theta_j \mu_j - b_k \theta_k \mu_k + (\theta_k - \theta_j) \mu_k \mathbf{D}_k V_t(\mathbf{n} - \mathbf{e}^k) \geq 0$.

(2) $b_j \theta_j \mu_j > b_k \theta_k \mu_k$ and $\theta_k \leq \theta_j$. In this case, using **P1**, we replace $\mathbf{D}_k V_t(\mathbf{n} - \mathbf{e}^k)$ in (B.22) with its lower bound b_k, and considering the fact that $(\theta_k - \theta_j)$ is non-positive in this case, we have

$$b_j \theta_j \mu_j - b_k \theta_k \mu_k + (\theta_k - \theta_j) \mu_k \mathbf{D}_k V_t(\mathbf{n} - \mathbf{e}^k) \geq b_j \theta_j \mu_j - b_k \theta_k \mu_k + (\theta_k - \theta_j) \mu_k b_k > 0$$

since if $b_j\mu_j > b_k\mu_k$, then $b_j\theta_j\mu_j - b_k\theta_k\mu_k + (\theta_k - \theta_j)\mu_k b_k = \theta_j(b_j\mu_j - b_k\mu_k) > 0$.

If **C2** holds between the customer of types j and k, then $b_j\theta_j\mu_j \leq b_k\theta_k\mu_k$, $\theta_j \leq \theta_k$ and $b_j\theta_j\mu_j \geq b_k\theta_k\mu_k\left[1 - \frac{\theta_k-\theta_j}{\Lambda}\right]$. Note that,

$$b_j\theta_j\mu_j \geq b_k\theta_k\mu_k\left[1 - \frac{\theta_k - \theta_j}{\Lambda}\right] \quad \Rightarrow \quad b_j\theta_j\mu_j - b_k\theta_k\mu_k + (\theta_k - \theta_j)\mu_k(\frac{b_k\theta_k}{\Lambda})$$

Using **P1**, we have $\mathbf{D}_k V_t(\mathbf{n} - \mathbf{e}^k) \geq \frac{b_k\theta_k}{\Lambda}$ and therefore

$$b_j\theta_j\mu_j - b_k\theta_k\mu_k + (\theta_k - \theta_j)\mu_k\mathbf{D}_k V_t(\mathbf{n} - \mathbf{e}^k) \geq b_j\theta_j\mu_j - b_k\theta_k\mu_k + (\theta_k - \theta_j)\mu_k(\frac{b_k\theta_k}{\Lambda}) \geq 0$$

CASE 2: When $n_j = 1$ at state $\mathbf{n}$. In this case, serving type-j customer is optimal in states $\mathbf{n}$ and $\mathbf{n} - \mathbf{e}^k$, but serving the customer of type $z \neq j$ is optimal at $\mathbf{n} - \mathbf{e}^j$. Thus, $\mu_k\mathbf{D}_k V_t(\mathbf{n})$ is still given by (B.16), but $\mu_j\mathbf{D}_j V_t(\mathbf{n})$ we have

$$\begin{aligned}
\mu_j\mathbf{D}_j f_t(\mathbf{n}) &= \mu_j\left[\mu_j V_t(\mathbf{n} - \mathbf{e}^j) + V_t(\mathbf{n})\sum_{i=1, i\neq j}^{M}\mu_i\right] \\
&\quad - \mu_j\left[\mu_z V_t(\mathbf{n} - \mathbf{e}^j - \mathbf{e}^z) + V_t(\mathbf{n} - \mathbf{e}^j)\sum_{i=1, i\neq z}^{M}\mu_i\right] \\
&= \mu_j\mu_z\left[V_t(\mathbf{n}) - V_t(\mathbf{n} - \mathbf{e}^j - \mathbf{e}^z)\right] + \mu_j\mathbf{D}_k V_t(\mathbf{n})\sum_{i=1, i\neq j,z}^{M}\mu_i \\
&= \mu_j\mu_z\left[\mathbf{D}_k V_t(\mathbf{n}) + \mathbf{D}_k V_t(\mathbf{n} - \mathbf{e}^j)\right] + \mu_j\mathbf{D}_k V_t(\mathbf{n})\sum_{i=1, i\neq j,z}^{M}\mu_i \\
&= \mu_j\mu_z\mathbf{D}_z V_t(\mathbf{n} - \mathbf{e}^j) + \mu_j\mathbf{D}_k V_t(\mathbf{n})\sum_{i=1, i\neq j}^{M}\mu_i
\end{aligned} \tag{B.23}$$

Since serving type-z customer is optimal at state $\mathbf{n} - \mathbf{e}^j$, we have $[f_t(\mathbf{n} - \mathbf{e}^j)]_z \leq [f_t(\mathbf{n} - \mathbf{e}^j)]_k$ for all $k \neq z$. Thus, using the optimality equation (2.1), we

have:

$$\mu_j \mu_z \mathbf{D}_z V_t(\mathbf{n} - \mathbf{e}^j) \geq \mu_j \mu_k \mathbf{D}_k f_t(\mathbf{n} - \mathbf{e}^j)$$

Using the induction assumption (B.12) and comparing (B.23) and (B.16), we have

$$\mu_j \mathbf{D}_j f_t(\mathbf{n}) \geq \mu_k \mathbf{D}_k f_t(\mathbf{n}), \ \forall j, k \in \mathcal{J}_{\mathbf{n}} \text{ and } j \neq k.$$

Thus, the rest of the proof for Case 2 follows exactly the same steps of the proof provided for Case 1. This concludes the proof of Proposition 2.3. □

PROOF OF THEOREM 2.3: By Theorem 2.1, we know that there exists a stationary average-cost optimal policy for the MDP problem presented with the optimality equation (2.1) and Idling is not optimal in a nonempty system. We prove by contradiction that if property **P3** holds, then type-j customer always has a higher priority over type-k customer. Suppose that **P3** holds, but it is optimal to give priority to type-k customer. Thus considering (1) and the fact that idling is not optimal, we have

$$\mu_k V(\mathbf{n} - \mathbf{e}^k) + \sum_{i=1, i\neq k}^{M} \mu_i V(\mathbf{n}) < \mu_j V(\mathbf{n} - \mathbf{e}^j) + \sum_{i=1, i\neq j}^{M} \mu_i V(\mathbf{n})$$

$$\mu_j \mathbf{D}_j V(\mathbf{n}) < \mu_k \mathbf{D}_k V(\mathbf{n}) \quad \Longleftarrow \quad \text{Contradicts } \mathbf{P3}$$

Therefore, if property **P3** holds, then type-j customer always has a higher priority over type-k customer. This concludes the proof of Theorem 2.3. □

APPENDIX C

Appendix of Chapter 3

Section 1

NORTHWESTERN EMERGENCY MEDICINE QUALITY SURVEY

Emergency Department Operational Use Only

Data and Time: *"auto-fill"*

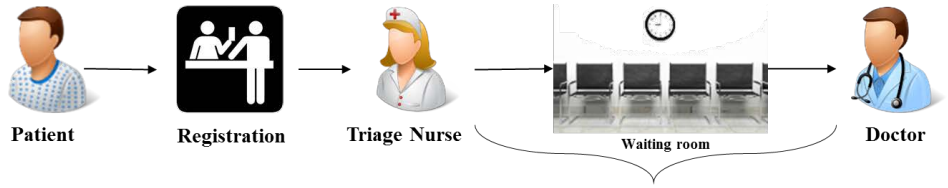
Section 2

NORTHWESTERN EMERGENCY MEDICINE QUALITY SURVEY

Your answers will help us make the experience for patients better in Emergency Department (ED). All your answers are confidential; your name will not be associated with anything you say.

When you arrived at this Emergency Department, the **triage nurse** evaluated your symptoms, vital signs and medical history. We would like to ask you about your waiting experiences only between seeing the triage nurse and seeing the doctor.

Patient's typical progress at ED



Who is filling out this survey? Patient Parent Family/Friend/Colleague Other

Section 3 (Waiting Experience)

Q1: How long did you wait in the waiting room after seeing the triage nurse and before seeing the doctor? Provide the best estimate of your wait ……
Less than half an hour
Half an hour to 1 hour
1 hour to 2 hours
2 hours to 3 hours
3 hours to 4 hours
More than 4 hours
I do not remember

Q2: Did you wait … you expected?
Much longer than
Longer than
Almost the same as
Shorter than
Much shorter than

Q3: On a scale of 1 to 7, how satisfied are you with your wait-time in the waiting room (after seeing the triage nurse and before seeing the doctor)?
1 Least Satisfied
2
3
4
5
6
7 Most Satisfied

Q4: On a scale of 1 to 7, how satisfied are you with your overall experience at our Emergency Department?
1 Least Satisfied
2
3
4
5
6
7 Most Satisfied

Q5: What was the estimated wait-time announced to you? Provide the best estimate of the wait announced to you ……
Less than half an hour
Half an hour to 1 hour
1 hour to 2 hours
2 hours to 3 hours
3 hours to 4 hours
More than 4 hours
Do not remember
No Announcement

Q6: How was your wait-time compared to the last time you were here?
Shorter than last time
Almost the same as last time
Longer than last time
I do not remember
This is my first time in Northwestern ED.

Q7: On a scale of 1 to 7, how did you feel about the order at which patients receive care in Emergency Departments?
1 Very unfair
2
3
4
5
6
7 Very fair
Did not notice the order

ED Wait Time Predictor

The growth in the number of hospitals publishing their ED wait-times leads to the rise of severals studies to develop methods for predicting the ED wait-times. From using a rolling average, the method currently used by some hospitals (Dong et al. 2015), or Quantile regression, which has the property of monotone equivariance and is less sensitive to outliers (Sun et al. 2012), to use a more accurate, widely-applicable method called "Q-Lasso" (Ang et al. 2015), which combines fluid model estimators and statistical learning. All developed models, even though insightful and somewhat applicable, make some large errors, sometimes in order of hours. As explained in section 4.3, we use a variety of statistical and machine learning techniques, introduced in the literature to estimate Triage-To-Doctor (TTD) time (i.e., the time between seeing the triage nurse and seeing the doctor). We, then, compare the accuracy of these models using Mean Squared Error (MSE) and select the one with the highest MSE. After checking for regulatory assumptions, we developed a generalized linear model, with the factors found to be a significant predictor of TTD, to estimate TTD for each patients visiting the ED under study. In this appendix, we first, describe the data and processing of collecting the data and then we introduce all the prediction method used and compare their performance.

C.0.1. Description of the ED Under Study, Data and Process

Our data comes a from the ED of an urban hospital in downtown Chicago. This study employs two years of de-identified data of all 177,831 patients treated in the ED from January 1, 2016 to December 31, 2017. The data set contains patient-level information including, but not limited to, the following: the patient's time of arrival and departure,

triage time, bed assignment and doctor time, LOS, ESI level, attending Physicians and Nurses, and ED Census. List of available data with the description is available in appendix 3. We exclude almost 7% of the overall sample which include missing data and patients who left without being seen. We start with a preliminary analysis of the data.

Figure C.1. Acuity Pie Chart

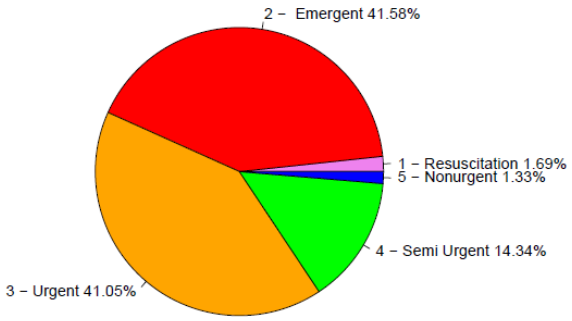


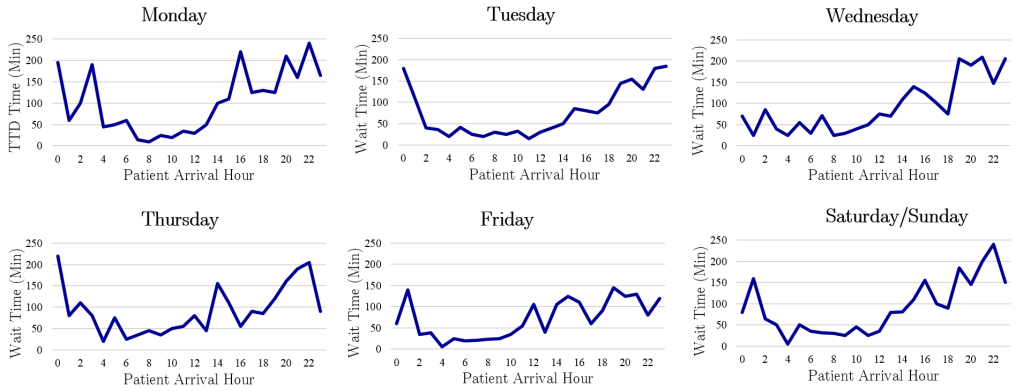
Figure C.1 demonstrates the pie chart of patients with different acuities, excluding unidentified acuity cases. This figure illustrates that more than 80% of the patient treated in NMH ED are acuity 2 (emergent) or 3 (urgent) patients and they are the patients who usually wait the longest in the waiting room. Around 2% of patients do not have any acuity level assigned, who were excluded from this analysis. The percentage of male patients are 45% (excluding the $< 0.1\%$ "Unknown" or "Not available" cases). The average age of patients is 48 years old with standard deviation of 19. Table C.1 shows TTD time for all priority levels.

As shown in Table C.1, acuity level 3 (i.e., urgent patients) waits on average the longest to see a doctor in this ED. However, the TTD time patterns are so different between mornings and evenings. Table C.2 shows a typical daily TTD time changes for different days of week.

Table C.1. Basic Descriptive Statistics

Priority (acuity) level	N	Mean	St. Dev.
1 - Resuscitation	2,794	18.260	103.972
2 - Emergent	68,571	64.259	85.780
3 - Urgent	67,689	77.706	88.171
4 - Semi Urgent	23,654	66.390	89.138
5 - Nonurgent	2,199	74.523	119.300

Figure C.2. A Typical Daily TTD Time in the ED



As shown in Table C.2, the TTD time of patients are non-stationary and much smaller in the mornings and increases in the afternoons. This trend is observed in all days of the week, even though the figures are slightly different throughout the week. This observed daily seasonality of TTD time will be carefully considered in all our prediction methods.

C.0.2. Wait Time Prediction Models

In this section, we present all models considered in our study. Table C.2 summarizes the methods used to predict patients wait-times.

In the following sections, we briefly introduce each method, how they generally work and explain why they are selected to predict patients' wait-time.

Table C.2. Prediction Models Grouped by Prediction Method Categories

Model Categories	Models
Time Series Models	Rolling Average (K last patients)
	Rolling Average (W last hours)
	Holt-Winter's Method
Regression Based Models	Quantile Regression
	Stepwise Linear Regression
	Generalized Linear Models
	Q-Lasso
Machine Learning Models	Boosted Regression

C.0.2.1. Time Series Models. This section summarizes the time-series models used for wait-time prediction.

Rolling Average

Moving averages (or rolling averages) are a variety of models commonly used with time-series data to conduct forecasting. The rolling average method is often used to smooth time-series and isolate longer term trends. They take set sized subsets of the data and calculate the average value of the number of interest. The subset then "moves" to next observation, drops the farthest observation on the other end of the subset and calculate the new average. The size of the subsets are defined by a given value N, which sets the number of observations per subset. We can define the subsets based on a fixed-time-window W and then take the average of all observations occurred during that time-window. These models aim to reduce the effect of short-term fluctuations and reveal long term trends.

This approach seemed attractive due to the seasonality exhibited of wait times between weekdays and between operating hours of each day. The calculations for a moving average involve a rather simple averaging of the variable of interest. In this study, we used both forms of the rolling average model approach and changed the parameters N and W to identify the candidate rolling average and parameter level that achieves the minimum

MSE. We took the averages for each priority class separately. The best number of observations that works the best for the ED under study was $N^* = 9$ and the time-window that works the best was $W^* = 3$ hours.

Holt-Winters Model

Holt-Winters method, also known as Triple-Exponential smoothing, is used to forecast data points in a seasonal time series. Three parameters control smoothing: α, β, and γ, each of which provides an additional forecast. These parameters take on values between 0 and 1, with values closer to 0 placing less value on the more recent data points, and are calculated to minimize the total MSE. We computed the values of these parameters through a cross-validation process. This method provides a seasonal factor, the length of which varies on the type of data. For detailed introduction to this method, see **?**. In the ED under study, we observed seasonal trends by the day and week. As shown in Figure C.2, there is a rise and fall of incoming patients throughout different hours of the day and days of the week. Thus, this smoothing method can be used to forecast waiting times based on the hour of day and/or day of week. Advantages of this model include being able to apply and provide accurate forecasts, including any seasonal peaks. However, because it is a smoothing method, daily variation may not be taken into account appropriately, resulting in wide confidence intervals when charted out.

C.0.2.2. Regression-Based Models. This section summarizes the regression-based models.

Stepwise Regression Method

The stepwise method utilizes the linear regression method with an additional element of

feature selection. The method has three different varieties: forward, backward and both. The forward method starts with a model with no predictors and sequentially adds more one by one based on which predictor improves the model based on a set error measure (e.g., AIC, BIC, R^2). This method penalizes having more predictors, but is still based on the maximum likelihood function. The feature selection aspect of the stepwise model makes it useful for determining the most important predictors in our model. In this study, we used the *step* function in R to run our stepwise regression model. This function takes two models as inputs. One defines the "upper bound" or the model with the maximum number of predictors. The second model defines the "lower bound" or the model with the minimum number of predictors. Our model was set to use bidirectional elimination with AIC as the error measure. Our model determined that significant predictors in our model were patient acuity, day of the week, patient age, time of the day, number of patients already in the ED, patient gender and several interaction between these predictors.

Generalized Linear Models

The generalized linear models are flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The model still assumes a linear relationship between the predictors and response variable. There is no element of feature selection as every predictor available is used. This can often lead to several insignificant predictors being used in the model. In this study, we used a model produced by the gamma distribution, since we observed that the TTD time distribution best fits by gamma distribution. Mathematically, the GLM with the set to use the gamma family attempts to fit a regression in the form

$f(y;\lambda,\alpha) = \frac{1}{\Gamma(\alpha)}\lambda^{\alpha}e^{-\lambda}y^{\alpha-1}$ for $y \geq 0$, where $f(y)$ is the response variable value, α is the dispersion or scale parameter and λ is the parameter of interest.

The coefficient values are selected to min $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ where $\hat{y}_i$ is the predicted value and y_i is the actual values. In other words, GLM determines which line minimizes the distance between predicted values and actual values. Our selection of predictors to use in the GLM model was based off of conversations with providers and managers at the ED under study as well as wait time predictors previously found to be significant in the medical literature. For example, the number of patient in the ED at each acuity level which was used be Sun et al. (2012) and therefore we used patients' acuity (priority) level as one of the predictors. The *GLM* function in R indicates significant predictors based on the results of t-tests at several significance levels. In our model, we found that patient acuity, day of the week, patient age, time of the day, number of patients already in the ED and patient gender were all significant predictors of a patient's wait time.

Quantile Regression

Quantile regression is used to estimate and conduct inference about conditional quantile functions. Similar to how classic linear regression methods based on minimizing sums of squared residuals enable one to estimate models for conditional mean functions, quantile regression methods offer a mechanism for estimating models for the conditional median (50^{th} percentile) function, and other conditional quantile (e.g., 1^{st}, $20^{th}, 90^{th}$ ...) functions. By allowing for the estimation of conditional quantile functions, such as the median, quantile regression gives the capability to thoroughly examine the stochastic relationship among random variables. Mathematically, quantile regression differs from multiple linear

regression which presents a numerical linear algebra problem to solve. Quantile regression, rather, uses a linear programming approach that is often solved using the Simplex method. The formulation of this optimization function is as follows.

$$\min_{\beta^+,\beta^-,u^+,u^-\in\mathbb{R}^{2k}\times\mathbb{R}^{2n}_+}\{\tau 1_n' u^+ + (1-\tau)1_n' m u^- | X(\beta^+ - \beta^-) + u^+ - u^- = Y\},$$

where $\beta_j^+ = \max(\beta_j, 0), \beta_j^- = -\min(\beta_j, 0), u_j^+ = \max(u_j, 0), u_j^- = -\min(u_j, 0)$.

We used the *rq* function found in the *quantreg* library in R to implement our quantile regression model. In our model, setting $\tau = 0.5$, we construct a model that produced an estimate of median values, using all previous predictors.

Lasso Regression

The Lasso method falls under the family of shrinkage models. Methods in this family aim to reduce the magnitude of predictor coefficients which can dramatically reduce the variance of the coefficient values. In exchange for the decrease in variance the Lasso method yields higher bias. What sets the Lasso method apart from other shrinkage methods is it feature selection qualities. While other methods usually shrink non-significant predictor coefficients to a value close to zero they do not actually set the value to zero. The Lasso method, with the right shrinkage parameter λ, is more capable selecting a subset of important predictors. The Lasso method is useful in our case due to the dual benefits of variance reduction and feature selection. Plambeck et al. (2014) use Lasso combined with queueing related predictors to forecast wait-times in EDs. Here, we follow their procedure to

predict wait-time. The objective of the lasso is to solve for $\min_{\beta_0,\beta_1} \frac{1}{N} \sum_{i=1}^{N}(y_i - \beta_0 x_i^T \beta_1)^2$ subject to $\sum_{j=1}^{p} \beta_j \leq t$.

In our study, the penalty term was determined through cross validation with the *cv.glmnet* function. The model was created using the *glmnet* function in R and the penalty term was determined to be 0.0105.

C.0.2.3. Machine Learning Models. A key issue in statistical research is the development of algorithms for model building and variable selection (Hastie et al. 2009). This issue is due to the fact that classical techniques for model building and variable selection (such as generalized linear modeling with stepwise selection) are known to be unreliable or might even be biased. Hofner et al. (2014) consider component- wise gradient boosting (Friedman 2001), which is a machine learning method for optimizing prediction accuracy and for obtaining statistical model estimates via gradient descent techniques. A key feature of the method is that it carries out variable selection during the fitting process without relying on heuristic techniques such as stepwise variable selection. Moreover, gradient boosting algorithms result in prediction rules that have the same interpretation as common statistical model fits. This is a major advantage over machine learning methods such as random forests (Breiman 2001) that result in non-interpretable "black-box" predictions. The idea of boosting models is to start with several simple models and combine them in an adaptive way that leads to a stronger predictor. The boosted regression method works well in our case because of the highly complex and non-linear relationship between our predictors and response variable. In cases such as this, boosting can offer significant improvements in terms of prediction accuracy. We used the algorithm presented in Bühlmann and Hothorn (2007) to predict patients' wait-times in the ED.

C.0.3. Comparison of Statistical Learning Models

To select the best prediction method, we use the following selection criteria.

Mean Squared Error (MSE) - Mean squared error is often used as measure of fit for prediction models. It is calculated by dividing the squared difference between the predicted value and observed value by the number of observations.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

In this study, we use MSE as the primary criteria for model selection.

Standard Error (SE)- Standard Error is an estimate of the standard deviation in a sample. It allows us to take note of the variance in our sample.

$$SE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

where N is the number of observations, Y is the variable value and $\bar{Y}$ is the variable mean.

Overestimates/Underestimates Mean- Overestimates mean represents the average overestimation of wait time by the model. It is calculated by taking all overestimates and dividing them by number overestimate predictions. Underestimates Mean is of interest because of the greater emotional weight patients place on predictions that underestimate the actual wait compared to overestimates. Similar to the overestimates mean it is calculated by taking all underestimates and dividing them by number overestimate predictions.

Our models were trained on a partition of the data and then tested on the remaining data. Below the error statistics for each model for both in-sample and out-sample error are reported. We compare the performance of described statistical learning models, using out-of-sample and in-sample errors in Table C.3.

Table C.3. Comparison of wait-time prediction models

	In Sample (min)				Out-of-Sample (min)			
Prediction method	MSE	SE	Overestimates Mean	Underestimates Mean	MSE	SE	Overestimates Mean	Underestimates Mean
Rolling Average (K last patients)	4191	48	40	-52	4572	48	39	-52
Rolling Average (W last hours)	3984	53	37	-47	4367	53	36	-47
Quantile Regression (tau = 0.5)	3974	35	29	-54	6695	35	43	-76
Stepwise Linear Regression	3678	45	35	-52	3596	45	35	-52
GLM	3809	45	35	-53	3660	45	36	-52
Boosted GLM	3174	42	33	-49	3249	40	32	-53
Q-Lasso	4156	31	37	-60	4069	31	37	-61

Definition of Variables

Table C.4. Summery Definition of Variables

Name	Source	Description	Data Type
		Dataset used for wait-time prediction	
arrival_time	Enterprise Data Warehouse (EDW)	Patient Arrival Time to ED (recorded at T1 check-in)	Date/Time
triage_time	Enterprise Data Warehouse (EDW)	Patient Triage Time (recorded at the beginning of triage)	Date/Time
bed_assignment_time	Enterprise Data Warehouse (EDW)	Patient Bed Assignment Time	Date/Time
doctor_initial_contact_time	Enterprise Data Warehouse (EDW)	Patient Doctor Time	Date/Time
acuity_dsc	Enterprise Data Warehouse (EDW)	Patient Acuity (1:Resuscitation, 2: Emergent, 3: Urgent, 4: Semi-Urgent, 5:Nonurgent	Integer
primary_doc_id	Enterprise Data Warehouse (EDW)	Primary doctor assigned to the patient ID number	Text
primary_nurse_id	Enterprise Data Warehouse (EDW)	Primary nurse assigned to the patient ID number	Text
is_lwbs_flg	Enterprise Data Warehouse (EDW)	If the patient is LWBS 1, otherwise 0.	Binary
patient_gender	Enterprise Data Warehouse (EDW)	Patient's gender (male/female/unknown)	Text
patient_age	Enterprise Data Warehouse (EDW)	Patient's Age	Integer
chief_complaint_ed	Enterprise Data Warehouse (EDW)	Patient's chief complaint	Long Text
ED_census	Enterprise Data Warehouse (EDW)	Number of all patients in the ED at the hour of arrival (In waiting room or in beds)	Integer
inpatient_status_census	Enterprise Data Warehouse (EDW)	Number of patients in the inpatient unit	Integer
obs_status_census	Enterprise Data Warehouse (EDW)	Number of patients in the observation unit	Integer
ED_pts_waiting_for_inpt_bed	Enterprise Data Warehouse (EDW)	Number of patients discharged but waiting to be admitted	Integer
acuity_nonurgent_null_at_arrv	Enterprise Data Warehouse (EDW)	Number of acuity 5 patients in ED at the time of arrival	Integer
acuity_semiurgent_at_arrv	Enterprise Data Warehouse (EDW)	Number of acuity 4 patients in ED at the time of arrival	Integer
acuity_urgent_at_arrv	Enterprise Data Warehouse (EDW)	Number of acuity 3 patients in ED at the time of arrival	Integer
acuity_emergent_at_arrv	Enterprise Data Warehouse (EDW)	Number of acuity 2 patients in ED at the time of arrival	Integer
acuity_resuscitation_at_arrv	Enterprise Data Warehouse (EDW)	Number of acuity 1 patients in ED at the time of arrival	Integer
ed_team_name	Enterprise Data Warehouse (EDW)	ED Team that treated the patient	Text
DTT	Computed	Door-To-Triage Time = arrival_time – triage_time (minutes)	Non-negative Real Number
DTB	Computed	Door-To-Bed Time = bed_assignment_time – arrival_time (minutes)	Non-negative Real Number
DTD	Computed	Door-To-Doctor Time = doctor_initial_contact_time - arrival_time (minutes)	Non-negative Real Number
TTB	Computed	Triage-To-Bed Time = bed_assignment_time – triage_time (minutes)	Non-negative Real Number
TTD	Computed	Triage-To-Doctor Time = doctor_initial_contact_time - triage_time (minutes)	Non-negative Real Number
LOS	Computed	Length-Of-Stay = ed_departure_time – arrival_time(minutes)	Non-negative Real Number
month	Computed	Month the patient arrived	Text
day_of_week	Computed	Day of the week the patient arrived	Text
daily_shift	Computed	Shift of the day the patient arrived (8 shifts:7:30,9:30,11:30,15:30,19:30,22,23:30,3:30)	Integer
CAAT	Computed	If the patient treated at Care-At-Arrival 1, otherwise 0.	Binary
		Press Ganey survey data	
overall (F68)	Press Ganey Survey	Overall rating of care received during your visit	Integer (1-5)
LTR (F4)	Press Ganey Survey	Likelihood of your recommending our emergency department to others	Integer (1-5)
kept_informed (F1)	Press Ganey Survey	How well you were kept informed about delays	Integer (1-5)
wait_before_staff_noticed (A86)	Press Ganey Survey	Waiting time before staff noticed your arrival	Integer (1-5)
wait_before_treatment (A4)	Press Ganey Survey	Waiting time before you were brought to the treatment area	Integer (1-5)
wait_treatment_area (C1)	Press Ganey Survey	Waiting time in the treatment area, before you were seen by a doctor	Integer (1-5)
wait_radiology (D3)	Press Ganey Survey	Waiting time for radiology test	Integer (1-5)
waiting_area_comfort (A5)	Press Ganey Survey	Comfort of the waiting area	Integer (1-5)
		Point of care survey data	
perceived_wait_time	Point of care survey	Question 1 on survey: "How long did you wait in the waiting room after seeing the triage nurse and before seeing the doctor? Provide the best estimate of your wait . . . . . "	Non-negative Real Number
wait_time_gap	Point of care survey	Question 2 on survey: "Did you wait . . . you expected?"	Text
wait_satisfaction	Point of care survey	Question 3 on survey: "On a scale of 1 to 7, how satisfied are you with your wait-time in the waiting room (after seeing the triage nurse and before seeing the doctor)?"	Integer (1-7)
overall_satisfaction	Point of care survey	Question 4 on survey: "On a scale of 1 to 7, how satisfied are you with your overall experience at our Emergency Department?"	Integer (1-7)
perceived_announced_wait_time	Point of care survey	Question 5 on survey: "What was the estimated wait-time announced to you? Provide the best estimate of the wait announced to you . . . . . "	Non-negative Real Number
last_visit_wait_time	Point of care survey	Question 6 on survey: "How was your wait-time compared to the last time you were here? "	Notice/not notice
perceived_fairness	Point of care survey	Question 7 on survey: "On a scale of 1 to 7, how did you feel about the order at which patients receive care in Emergency Departments?"	Integer (1-7)
High Priority pct.	Computed	The percentage of patients with ESI level 2 or higher in our data	Percentage
First time visit pct.	Computed	The percentage of patients who visited the ED under study for the first time	Percentage
Larger wait-time gap pct.	Computed	The percentage of patients who choose "Longer than" or "Much longer than" for Question 2 on the survey	Percentage

Robustness Tests Summary Results

This section summarizes the result of robustness checks introduced in Section 3.7. Table C.5 shows the robustness checks for the model introduced to test Hypothesis 2B (i.e., loss aversion) and Table C.6 shows the robustness checks for the model introduced to test Hypothesis 2. Even though, observations made earlier, somewhat repeated in the provided models, still they may be some concerns about the definition of the wait-time gap. To address this issue, we used latent variable analysis, discussed in Section 3.7, where we assume the wait-time gap is unobserved and set of structural equations used to the test the robustness of our findings.

Table C.5. Models for Loss Aversion for Different Wait-time Gaps

	Dependent variable:		
	$A - W$	$\bar{W} - W$	$A - \dot{W}$
	(1)	(2)	(3)
1. Wait-time Gap	0.026^{***}	0.017^{***}	0.013^{***}
	(0.004)	(0.002)	(0.002)
2. Wait-time Gap $\times I_{\text{Wait-time Gap}>0}$	-0.019^{***}	-0.007^{*}	-0.004^{*}
	(0.005)	(0.005)	(0.004)
Constant	4.096^{***}	5.196^{***}	4.092^{***}
	(0.166)	(0.157)	(0.158)
Observations	286	286	286
Adjusted R^2	0.291	0.402	0.261

Notes. Parentheses contain robust standard errors. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively (two-tailed tests). Note that A, W, $\bar{W}$ and $\dot{W}$ denote announced wait-time, actual wait-time, average wait-time and perceived wait-time, respectively.

Table C.6. Models for Large Wait-time for Different Wait-time Gaps

	Dependent variable:		
	$A - W$	$\bar{W} - W\dagger$	$A - \dot{W}$
	(1)	(2)	(3)
1. Wait-time Gap	0.012^{***}	0.009	0.0097^{***}
	(0.005)	(0.011)	(0.002)
2. Wait-time Gap (> Third Quartile)	-0.02^{***}	-0.002	-0.0114^{***}
	(0.011)	(0.005)	(0.004)
Constant	4.009^{***}	5.325^{***}	4.237^{***}
	(0.146)	(0.335)	(0.158)
Observations	211	63	170
Adjusted R^2	0.045	0.027	0.086

Notes. Parentheses contain robust standard errors. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively (two-tailed tests). Note that A, W, $\bar{W}$ and $\dot{W}$ denote announced wait-time, actual wait-time, average wait-time and perceived wait-time, respectively. † The number of observations dropped by this definition of the reference point and therefore the results did not find to be significant.