Carnegie Mellon
Tepper
SCHOOL OF BUSINESS

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
INDUSTRIAL ADMINISTRATION
(OPERATIONS MANAGEMENT)

Titled

"INNOVATIVE MODELS IN SERVICE OPERATIONS"

Presented by

Leela A. Nageswaran

Accepted by

Alan Scheller-Wolf 4/17/18

Chair: Prof. Alan Scheller-Wolf **Date**

Approved by The Dean

Robert M. Dammon 4/23/18

Dean Robert M. Dammon **Date**

INNOVATIVE MODELS IN SERVICE OPERATIONS

BY

LEELA A. NAGESWARAN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Operations Management
at Tepper School of Business, Carnegie Mellon University, 2018

Pittsburgh, Pennsylvania

Doctoral Committee:

Soo-Haeng Cho
Mor Harchol-Balter
Aliza Heching
Alan Scheller-Wolf (Chair)
Sridhar Tayur

April 22, 2018

To my parents Nageswaran and Ulagammal,
and to my husband Tom.

Acknowledgments

First and foremost, I would like to thank my adviser Professor Alan Scheller Wolf, for his support, patience, and mentorship throughout my graduate studies. Alan introduced me to the wonderful world of queueing theory in the fall of 2013, and his enthusiasm made many discussions tackling challenging problems thoroughly enjoyable. I also thank him for his sage advice on everything from navigating challenges to communicating effectively.

I am indebted to Professor Soo-Haeng Cho, who has been instrumental in developing my skills as a researcher, writer, and presenter. I have benefited a great deal from his extensive knowledge and steadfast attention to detail, and I particularly appreciate his help during my job search. I am grateful to Professor Sridhar Tayur for his valuable advice on how to be successful; his enthusiasm for great work has shaped the way I perceive research. I would like to thank Professor Mor Harchol-Balter for providing guidance on research and for organizing the SQUALL lunches. I would also like to thank Dr. Aliza Heching for her mentorship and for kindly serving on my dissertation committee. Several faculty members have provided me with guidance at various stages of my PhD: I thank Professor Nicola Secomandi for teaching me a great deal about the academic lifestyle, Professor Joseph Xu for his advice on giving effective talks, and Professor Param Singh for sharing his knowledge and expertise with me.

I am fortunate to have the company of many wonderful colleagues at Tepper. I am grateful to former students Sherwin Doroudi, Xin Wang, Vince Slaugh, and Ying Xu for their friendship, advice, and support throughout my PhD. I am particularly thankful to Siddharth Singh and Wenting Yu for several engaging and inspiring conversations. I would like to thank Lawrence Rapp and Laila Lee for their prompt and considerate help that made the PhD experience enjoyable.

Most importantly, I would like to extend my deepest gratitude to my family. My parents, Nageswaran and Ulagammal, have always believed in me and provided me with encouragement and motivation throughout my life. Their passion for math, research, and hard work is truly contagious, and I owe all my successes to them. My in-laws, Judd and Sandra, have been a source of encouragement and love, and I am fortunate to receive their advice on numerous occasions. I would also like to thank my sister, Suju, for all the fun and laughter; it is nice to know that at least one other person constantly looks to me for advice. And most importantly, I am thankful to my husband, Tom, my biggest supporter. I am forever indebted to him for his encouragement, advice, patience, support, and unconditional love; he is my strength, my inspiration, and my world.

Contents

Chapter 1

Introduction

Many factors, including regulatory changes and the ubiquity of information technology, have spurred the growth of innovative ways of delivering services. As a result, the operations management community has been presented with new challenges. This dissertation seeks to model and provide insights for three such problems in the healthcare and retail industries.

The widely reported liver transplant of former Apple CEO Steve Jobs "in Tennessee, some 2,000 miles from Jobs' home in northern California"[1] posed new questions for organ transplantation: Critics of multiple listing – wherein patients awaiting organ transplants register in more than one region – claim that the wealthy enjoy unfair access to organs at the expense of others, while defenders argue that it helps alleviate geographic disparity. In Chapter 2 we model multiple listing using a two-server queueing system with two classes of customers, one of which is redundant and joins both lines while the other joins a single line. Prior work has focused on the improvements to overall system efficiency due to redundancy, leaving open the question of fairness – whether this benefit is at the expense of *non-redundant* (or singly-listed) customers.

A crucial aspect of this work is the ability to realistically study the fairness of multiple listing by considering the approach that multi-listed patients are likely to resort to in its absence: Specifically, we show that singly listed patients actually benefit under redundancy of the other class compared to the case in which redundancy is prohibited, and the 'wealthier' class using their resources to travel, if necessary, joins the shortest queue – i.e., redundancy is *fair* – if the queues are symmetric in their arrival rates. But, if the queues are asymmetric, redundancy can be unfair; we find situations when singly listed patients are provably worse off under redundancy than when the other class utilizes JSQ. To prove the former result, we developed a novel coupling method that allows one to process state transitions out of temporal order and showed that the queue length under the redundancy policy satisfies a specific ordering. We also extend some of these results to include customer abandonments and unequal service rates analytically, allowing us to more closely model the organ transplantation application.

Redundancy shows up in other settings as well: Supermarkets with multiple checkout lines exhibit a similar phenomenon when a customer and a companion each join a line, and abandon the "slower" line as soon as either of them reaches the cashier. While singly listed patients join a transplant center in whichever region they live in, non-redundant customers in other applications

[1] http://www.cnn.com/2009/HEALTH/06/24/liver.transplant.priority.lists/

often need to pick a line. Thus, we also model more general redundancy settings by allowing the non-redundant customers to have a queue-joining policy; for example, joining the shortest queue in a supermarket. And we consider two variants of redundancy differentiated by when the redundant request is dropped from the system – when the required service is complete or when it begins – thus modeling a diverse set of applications. The question now is how should non-redundant customers choose which queue(s) they join to minimize their wait. We show that joining the shortest queue does not always result in the lowest delay for a non-redundant customer if complete system state information (including which customers are redundant) is available, but that it *is* optimal if only queue length information is observable. By analyzing different policy choices for the non-redundant customer, this work provides fundamental insights on fairness, optimal queue-joining policies, and the value of information in redundancy systems, from the non-redundant customers' perspective.

Consumers in today's e-commerce era are extremely savvy, and they interact with retailers across multiple channels. Surveys show that their purchasing and return behavior is affected by the retailer's return policy; for example, charging fees for returns via mail or allowing free returns to physical store locations. As a result, the omnichannel strategy – wherein a firm utilizes its physical stores and digital assets to provide a convenient experience – is quickly becoming the new norm among retailers. In Chapter 3 we study how an omnichannel firm makes return policy and pricing decisions when faced with heterogeneous customers who purchase, and possibly return misfit items, using either of its two (online or in-store) channels. Existing research focuses either on return policies of single channel firms or on fulfillment strategies of omnichannel firms without considering return policies: The primary contribution of this work is to marry the two streams to fully understand the characteristics of an omnichannel firm that influence its preference for different return policies observed in practice, such as full refunds, in-store returns and partial refunds.

We incorporate two important aspects of omnichannel operations that have largely been ignored in prior work – identical prices across channels and cross-channel returns. In addition to being consistent with practice, the identical price assumption highlights an important constraint experienced by a firm due to its omnichannel nature. By modeling cross-channel returns, we are able to provide one explanation for the generous refunds common in practice. We show that when customers return online, the refund chosen by an omnichannel firm is more generous than suggested by prior work that considers only single channel firms.

This is the first study to characterize the optimal return policy in an omnichannel context: We demonstrate that all three return policies seen in the marketplace – full refunds, online returns for a fee, and free in-store returns – may be prescribed by a profit-maximizing omnichannel firm. The optimal policy helps explain why some omnichannel firms choose full refunds while others charge a fee for returns: We find that firms with good salvage partners (e.g., Nordstrom) as well as those with more store-based customers (e.g., Macy's) should offer full refunds. In contrast, firms who enjoy better in-store salvage opportunities (e.g., J.C. Penney) should charge a fee for online returns in order to prompt free in-store returns. We also find that firms should be cautious in making the returns process overly convenient or improving accessibility to their stores, as these policies, combined with the wrong return policy, could reduce the firm's overall profit.

The ease of disseminating information online has made it possible to rate nearly every service, from movies (IMDB) to physicians (RateMDs) to professors (RateMyProfessors). Recently, the Centers for Medicare and Medicaid Services (CMS) published quality scores for hospices on various dimensions (e.g., pain assessment, treatment preferences) throughout the US: The goal is to provide more transparency for patients and encourage providers to improve the quality of their services. In Chapter 4 we study a hospice manager's problem of controlling quality of care in light of these regulatory changes.

Hospice providers have questioned whether such programs actually improve quality. We investigate this claim analytically by studying a hospice manager's incentives under mandated public reporting of quality. A key modeling feature captured in this work is the impact of quality on the patient census – the number of patients cared for by a hospice. On the one hand, quality impacts the rate at which patients leave the hospice; for example, high hospital re-admissions are considered to be correlated with poor quality. On the other hand, publicly available measures enable easy comparison of hospices, thus directly impacting the joining decision of new patients. We use a discrete-time Markov decision process (MDP) model to study the weekly quality control problem of the hospice.

By analyzing the MDP, we show that a simple constant quality policy is optimal in the absence of reporting; but, the optimal policy under reporting is, in general, quite complex. Thus, we designed a closed-form heuristic policy that is easy to compute. Using realistic parameter values that represent a typical hospice setting, we conducted several experiments to investigate the heuristic's performance numerically. We find that the heuristic does extremely well against the optimal policy: The average quality under the heuristic policy is within 3%, and the profit is within 0.05%, of the optimal policy across all experiments. And the structural properties of the quality level under the heuristic – high at low patient census and bounded below by the constant level in the absence of reporting – translate to the optimal policy for realistic parameter settings.

By considering two scenarios – one before and one after public reporting is in effect – we show that quality always improves under reporting. But not all hospices are likely to obtain a larger profit in the presence of reporting: Those that enjoy a large share of the market in the absence of reporting will be less inclined to report quality due to the risk of losing customers. We find that the "pay-for-reporting" structure currently implemented by CMS is one approach to incentivize reporting; but, those hospices that strategically choose not to report their quality metrics will actually reduce their quality level as a result of the penalty on their reimbursement rate.

We conclude by summarizing the contributions of this dissertation and by outlining directions for future research based on the work in this dissertation in Chapter 5.

Chapter 2

Queues with Redundancy: Is Waiting in Multiple Lines Fair?

2.1 Introduction

While customers dislike waiting in general, they particularly dislike being stuck in a slow queue. Thus, a party with multiple persons might have members join multiple lines (even though they need service at only a single server) to reduce the risk of getting "stuck"; this creates *redundant* requests or jobs. The most prevalent and highly relatable application of this sort of redundancy is at the supermarket: To potentially reduce the checkout time, a customer and a companion may each join a line, abandoning the "slower" line as soon as either one of them reaches the cashier. Given that not all customers do this (because of differences in their patience levels or due to visiting the supermarket alone), one may wonder how this partial redundancy affects those customers who join only one queue, and whether the simple rule of joining the shortest queue (JSQ) will always be these "non-redundant" customers' best policy. Moreover, it is not clear if such redundancy harms the non-redundant customers in any way i.e., if it is *unfair*. (Anecdotal evidence suggests that customers who joined a single line are often annoyed when others indulge in redundancy.)

A more critical application, and our primary motivation for the majority of this chapter, is organ transplants. For example, there are over 100,000 patients waiting for deceased donor kidney transplants for End-Stage Renal Disease (ESRD) in the US (Organ Procurement and Transplantation Network 2015). The unadjusted median wait times for adults transplanted in 2011 varied from 0.8 to 3.8 years in different states (United States Renal Data System 2013). Owing to this variability, patients have found that they can potentially improve their chances of getting a transplant by listing at two or more transplant centers simultaneously i.e., *multiple listing*: Studies show that transplantation rates are significantly higher (88% more access, Merion et al. 2004), and the expected transplantation waiting time is reduced (White et al. 1997), for multiple listed patients. Critics of multiple listing claim that it provides unfair access to organs for patients based on wealth (e.g., Blumstein 1989, Merion et al. 2004), since it is not feasible for all patients to multiple list due to the need for evaluations at every additional transplant center and potential travel and lodging expenses incurred if the patient receives an organ outside of her local area (United Network for Organ Sharing 2015). As a result, few patients are multiple listed (5.1% of patients on the kidney transplant list as of January 31, 2009, Ardekani and Orlowski 2010). Startups, such as Organ-

Jet Corporation, have emerged to try to alleviate this through the use of time-shared private jets (Lamas 2014).

Redundancy finds use in computer systems as well: Large-scale distributed storage systems in computing, such as the Google File System and the Amazon Elastic Block Store, have begun to replicate content over multiple disks. Recent computer science papers analyze the use of redundancy to allow content, stored on a set of disks, to be reconstructed by reading a subset of these disks (Joshi et al. 2014), potentially enabling different jobs of a single customer to be in service at the same time (Shah et al. 2016). There is some practical evidence that redundant jobs help reduce latency (Liang and Kozat 2014).

Finally, in general service systems, redundancy can be particularly beneficial when customers face uncertainty around resource availability. For example, a delay-sensitive firm may place requests for orders from several suppliers and purchase from whomever delivers first. Similarly, Ticketmaster, the ticket sales and distribution company, suggests that customers use more than one computer while booking tickets for popular shows. And, families interested in adoption often apply to multiple agencies in order to reduce the wait time to be matched.

We model redundancy using a two-server queueing system with two classes of customers:

1. Non-redundant customers join a single queue (according to some policy) upon arrival.
2. Redundant customers send jobs to both queues upon arrival.

Redundant systems may differ based on when redundant requests of a customer are dropped:

1. *Cancel upon service* (CUS) removes redundant requests when *any* one of them *completes service*. Thus multiple copies of the same customer could be in service simultaneously.
2. *Cancel upon entry* (CUE) permits only *one* request of each customer to be in service at any time. Thus when any request *enters service*, the others are removed from the queue.

The supermarket and distributed storage applications fall naturally into the CUE and CUS settings respectively, whereas multiple listing may be modeled using CUS by defining being "in service" as waiting at the head of the list for the next organ to become available.

All of the service systems mentioned above are built around redundancy; yet, several fundamental questions about redundancy remain unanswered. First, whereas the improvement in overall system performance due to redundancy has largely been the focus of prior work (Ata et al. 2016, Gardner et al. 2017, 2016), the question of fairness – whether this improvement is at the expense of non-redundant customers – has not been definitively addressed. This question is particularly important in the context of organ transplants: Identifying if multiple listing is unfair to singly listed patients and, if so, finding an alternative policy that is fair is crucial. Second, non-redundant customers may have the option to choose which queue(s) to join to try to minimize their wait. While singly listed patients join a transplant center in whichever region they live in, non-redundant customers in the supermarket, distributed storage and adoption applications need to pick a line according to some policy. What policies they should use, and how different levels of information affect these decisions, is an open question. Finally, does redundancy help or hurt overall system performance, and what role does the customers' joining policy have on the extent of this effect? We provide answers to these questions.

Our main contributions are summarized below:

- *Fairness:* We are the first to provide analytical results on fairness: Are singly listed patients necessarily harmed by multiple listing? We assume that the redundant class either utilizes multiple listing (i.e., redundancy) or joins the shorter queue (if multiple listing is not allowed). We prove that singly listed patients actually *benefit* under redundancy of the other class (compared to JSQ) when the queues are symmetric. But, if the queues are asymmetric, there are situations when singly listed customers are provably worse off under redundancy. By incorporating the approach that multi-listed patients are likely to resort to in the absence of multiple listing i.e., JSQ, we are able to realistically compare their impact on singly listed customers and provide insights as to when multiple listing is unfair.
- *Queue-joining policy:* We analyze the queue-joining decision for non-redundant customers who wish to minimize their delay. We show that the optimal policy depends on the extent of state information – queue lengths and customers' class – available to them: JSQ is optimal when only queue lengths are known; but, if additional information regarding the class of each customer is known, it may be better to join the *longer* queue.
- *Value of information:* Given that a complex state-dependent policy is optimal when complete state information is revealed to non-redundant customers, we quantify the value of this customer class information, finding that the (class-blind) JSQ policy is a very good heuristic.
- *Optimality of 'smallest workload' policy:* We show that the mean overall system time under CUE is at most that under JSQ when there are two customer classes, one of which is flexible and one that joins a queue randomly (with equal probability). Foss (1989) showed a similar result for what he called the "smallest workload" policy (equivalent to CUE), for the special case where only a single (flexible) class is present. We develop a new coupling approach and a method for keeping track of state information under redundancy to show our result.
- *Overall system performance:* We find that redundancy could be an important tool to improve the overall response times especially in less efficient systems such as when non-redundant customers join a dedicated server, systems operating at higher loads, and systems with asymmetric loads.
- *Model extensions:* We extend our optimal joining results and some of our fairness results to systems with more than two queues. We also extend some of our fairness results to include customer abandonments and unequal service rates, allowing us to more closely model organ transplantation.

The rest of the chapter is organized as follows. In §2.2, we review the relevant literature and describe how it differs from our present work. We formalize our queueing model in §2.3. In §2.4 we address the question of fairness. In §2.5 we study the optimal routing problem under different information regimes, and quantify the value of revealing complete state information. We study the effect of redundancy on the overall system performance under different service policies in §2.6. We present several extensions in §2.7 and finally, in §2.8, we conclude and discuss future research directions.

2.2 Literature review

Our work falls in the area of redundancy models: These consist of n parallel queues such that each incoming customer sends jobs to r queues and departs when any subset $k \leq r$ of them are served.

Most related to our work are Gardner et al. (2016) and Ata et al. (2016), both of which consider class-based r. Gardner et al. (2016) obtain the limiting distribution for CUS under the assumption that each class is associated with a fixed subset of queues to which a customer sends her jobs – restricting themselves to only one queue-joining policy, the Dedicated policy. We analytically compare the performance for several queue-joining policies for the non-redundant customers such as Dedicated, JSQ, and the optimal joining policy. In follow-up work, Gardner et al. (2017) provide a method to find the response time for systems in which a customer (i.e., single class) makes copies at a constant number of servers.

A key result in these papers is that redundancy (CUS, in particular) improves mean *overall* system time compared to when there is no redundancy. We show that the overall system time under redundancy (*both* CUE and CUS) is lower *compared to when redundant customers follow JSQ*, if the queues are symmetric. Furthermore, we are the first to provide analytical results on the fairness of redundancy, identifying situations when redundancy (compared to JSQ) benefits *non-redundant* customers. While Gardner et al. (2016) use numerical simulation to note that redundancy is "often preferred" by non-redundant customers, we analytically provide sufficient conditions for this to be true, as well as conditions for the opposite: There are parameters for which redundancy is *not* preferred i.e., it is unfair to non-redundant customers.

Ata et al. (2016) use fluid and diffusion approximations to analyze their overloaded queueing model for organ transplantation. While they find that multiple listing (i.e., by wealthier patients) can lead to improvements in geographic equity compared to when all patients join the closest regional transplant center, we show that multiple listing is beneficial to singly-listed patients compared to when these (wealthier) patients, prevented by multiple listing, join the transplant center with the shorter wait list. Thus, we are arguably able to compare multiple listing to a more realistic alternative. Finally, we extend our fairness results to include key aspects of organ transplantation, such as patient abandonment and unequal service rates, that also appear in their model.

This research stream has gained attention as redundancy has begun to be utilized to improve latency in large-scale distributed storage systems: For a single class of customers Joshi et al. (2014) and Shah et al. (2016) note that waiting for only a subset of disks (i.e., $k < r$) improves latency. Kumar et al. (2014b) consider class-based k and obtain bounds on the response time for the CUS system. While they demonstrate through simulation that increasing redundancy for *all* classes reduces the system time for all classes, they do not study the impact of more redundant jobs on less redundant ones, or the performance of different routing policies for non-redundant jobs. Guo and Hassin (2016) consider the strategic behavior of customers who, upon arrival to a two-server queueing system, may choose either to place redundant orders, place a single order, or balk, and analyze the equilibrium thereof.

Some papers consider redundancy within a central queue setting, removing the question of routing policies, which differentiates this line of work from ours. For the CUS system Chen et al. (2014) identify the optimal policy for allocating identical servers to customers. Sun et al. (2015) study the performance of various scheduling policies under non-exponential distributions, and Lee et al. (2015) incorporate the overhead of canceling redundant requests. Koole and Righter (2008) identify conditions ensuring that it is optimal to replicate all jobs to all servers.

There are several other areas that are closely related to our model, namely fork-join queues, flexible server systems, and duplicate orders. The fork-join queue is used to model a system where jobs requiring multiple resources may be executed in parallel. Jobs arrive in batches of k tasks and each of them is served by one of n servers. The job is complete when the last of the k tasks finishes service. This is different from our model because fork-join queues have no redundancy – all of the tasks must be completed. Exact solutions for the stationary distribution have been obtained by Flatto and Hahn (1984) for the case of two servers. Li et al. (2016) use mean-field analysis to derive the stationary distribution when jobs are sent to the k least-loaded servers among n servers.

In flexible server systems jobs may be of different classes; each server is capable of serving some subset of these classes. The CUE model in which the non-redundant customers arrive to each queue according to independent Poisson processes (i.e., they follow the Dedicated policy) can be modeled as a flexible server system. Visschers et al. (2012) derive the stationary distribution for this model under specific assumptions on the assignment of flexible (i.e., redundant) jobs to idle servers. And, for the two server case, McDonald and Turner (2000) use large deviations methods to show that having flexible servers outperforms other policies, such as routing the flexible class to the shorter queue, for minimizing mean system time. But neither of the aforementioned papers provide any discussion of optimal policies for the non-redundant customers, or evaluate whether redundancy is unfair to non-redundant customers, as we consider.

During a product stockout a delay-sensitive customer may place her order at many firms, obtain the product from whomever produces it first, and cancel other outstanding orders. Li (1992) notes that such “duplicate ordering” is a common pheonomenon in the semiconductor industry, and studies its welfare implications in the presence of lead-time competition between firms. Considering the perspective of a manufacturer selling through two independent distributors, Armony and Plambeck (2005) study the demand estimation and capacity investment implications of duplicate orders. These papers differ from our model in that redundancy exists in the above settings only when the product is out of stock.

Table 2.1 provides a categorization of the theoretical redundancy literature (excluding overloaded

Table 2.1: Brief overview of redundancy literature.

Paper	Type	Class-based	Service policy	Non-FCFS scheduling	Info. regimes	Fairness	General service
Chen et al. (2014)	CUS		Central	✓			
Gardner et al. (2016)	CUS	✓	Dedicated			✓	
Gardner et al. (2017)	CUS		Dedicated				
Joshi et al. (2014)	CUS		Dedicated				
Koole and Righter (2008)	CUS		Central	✓			✓
Kumar et al. (2014b)	CUS	✓	Dedicated	✓			
Lee et al. (2015)	CUS		Central				
Shah et al. (2016)	CUS		Central, Dedicated				✓
Sun et al. (2015)	CUS	✓	Central	✓			✓
Visschers et al. (2012)	CUE	✓	Dedicated				
Joshi et al. (2015)	Both		Dedicated				✓
This work	Both	✓	Dedicated, JSQ, Optimal		✓	✓	

systems, which are analytically distinct) based on the main aspects considered in each paper. Our work alone considers the following aspects together:

1. We study class-based redundancy under CUS and CUE, thus encompassing a great variety of applications.
2. We analyze alternate queue-joining policies for non-redundant customers, such as JSQ and the optimal policy, in addition to Dedicated.
3. We identify the optimal delay minimizing policy for non-redundant customers under different information regimes, and quantify the value of complete system information to them.
4. We provide analytical results on fairness and identify when redundancy harms or benefits non-redundant customers, as compared to other policies such as JSQ.

2.3 Model

We consider a system with two identical servers, denoted $j \in \{1, 2\}$, each with its own queue, and two classes of customers, denoted $i \in \{1, 2\}$. The arrival process for class $i \in \{1, 2\}$ is assumed to be Poisson with rate λ_i. Class 2 customers constitute a fraction $r \in [0, 1]$ of the total arrival stream (arriving at rate λ), so that $\lambda_2 = r\lambda$ and $\lambda_1 = (1 - r)\lambda$.

A customer of class 2 replicates into two jobs upon arrival, with each job joining a queue. Let $\pi_2 \in \{\text{CUS, CUE}\}$ denote the specific (redundancy) policy used by class 2. (Later, in §2.4, we will investigate $\pi_2 = \text{JSQ}$ as well.) A customer of class 1 does not replicate; she joins a queue immediately upon arrival according to policy $\pi_1 \in \{\text{Ded, JSQ, Opt}\}$. Under the Dedicated policy ($\pi_1 = \text{Ded}$) class 1 customers form independent Poisson streams to either queue: They arrive to queue 1 at rate $\lambda_1 p$ and queue 2 at rate $\lambda_1(1-p)$, where $p \in [0.5, 1]$ denotes the fraction of class 1 customers joining queue 1 under this policy. This policy arises naturally in many unobservable queueing models that deal with strategic customer behavior (Edelson and Hilderbrand 1975). The case of $p = 0.5$ is of particular interest: It is the equilibrium strategy when customers choose to be redundant or not (due to replication costs), as in Guo and Hassin (2016). In addition, Dedicated is often used to model arrivals in related work (see Table 2.1), for example independent arrivals of patients to different regional transplant centers. Under the JSQ policy ($\pi_1 = \text{JSQ}$) class 1 customers join the shorter queue upon arrival; when the queue lengths are equal the customer joins queue 2 with some probability, say η. We also examine the individually optimal policy ($\pi_1 = \text{Opt}$) that minimizes *the class 1 customers' delay.* By definition the optimal policy provides superior performance, but may be quite complex. The JSQ policy is simpler and yet is known to perform well in a variety of settings when queues are observable. In all cases we assume jockeying is prohibited.

The service time at each server is independent and distributed exponentially with rate μ. We define the system load to be $\rho = \lambda/2\mu$. A redundant customer arriving to an empty system under $\pi_2 = \text{CUE}$ chooses server 2 with probability η. Within each queue the service discipline is first-come first-serve (FCFS). Our assumptions of exponential service time and Poisson arrivals are empirically justified in our motivating kidney transplants application (Davis et al. 2013). Furthermore, using real traces from Amazon S3, Chen et al. (2014) find that exponential service times are an accurate approximation for distributed storage systems.

We focus on two servers in the majority of this chapter for two reasons: (i) This is assumed in several related papers due to its (relative) ease of analysis: Guo and Hassin (2016) consider two parallel queues with strategic redundancy, and Gardner et al. (2016) derive performance measures in closed-form for simple two or three server systems; and (ii) It captures the primary dynamics of redundancy and highlights important trade-offs: Gardner et al. (2017) show that the biggest improvement due to redundancy occurs when customers replicate to two servers, and Ata et al. (2016) allow customers to list at only one other location. Nevertheless, we extend our results to larger systems in §2.7, where we also allow customer abandonments and unequal service rates.

The state $\mathbf{S} = \{\mathbf{S}_1; \mathbf{S}_2\}$ of the system is given by vectors $\mathbf{S}_j, j \in \{1, 2\}$ denoting the types of different jobs at server j in the order of their arrival; the first (left-most) element denotes the job in service at that server. Exact analysis of this system is prohibitively complex, except in certain special cases for $\pi_1 = \text{Ded}$ (Gardner et al. 2016, Visschers et al. 2012).

We denote the class i mean response time – which is defined as the average time spent in the system (including in service) – under joining policies π_1 and π_2 as $T_i^{(\pi_1,\pi_2)}$, and the mean overall response time (taken over both classes) as $T^{(\pi_1,\pi_2)}$, for $\pi_1 \in \{\text{Ded, JSQ, Opt}\}$ and $\pi_2 \in \{\text{CUS, CUE, JSQ}\}$. It immediately follows that $T^{(\pi_1,\pi_2)} = rT_2^{(\pi_1,\pi_2)} + (1-r)T_1^{(\pi_1,\pi_2)}$. We assume that $\rho < 1$, and additionally $\lambda_1 p < \mu$ when $\pi_1 = \text{Ded}$, for stability (Gardner et al. 2016, Visschers et al. 2012).

2.4 Fairness

This section addresses the question of fairness – are non-redundant customers adversely affected by redundancy of the other class? Suppose we have two systems $\mathcal{X}$ and $\mathcal{Y}$ that differ only in the policy adopted by class 2 customers, say x and y in $\mathcal{X}$ and $\mathcal{Y}$ respectively. Then, we define system $\mathcal{X}$ as less "fair" than $\mathcal{Y}$ if $T_1^{(\pi_1,x)} \geq T_1^{(\pi_1,y)}$.

Since this question is of particular interest for organ transplantation due to the extant discussion around the (seemingly) unfair advantage for multiple listed patients (Blumstein 1989, Merion et al. 2004), we consider $\pi_1 = \text{Ded}$ and $\pi_2 \in \{\text{CUS, CUE, JSQ}\}$ in §2.4.1. We then consider the JSQ policy ($\pi_1 = \text{JSQ}$) in §2.4.2.

2.4.1 Fairness when class 1 uses the Dedicated policy

In organ transplantation, singly listed patients are likely constrained (e.g., due to cost) to listing in their home geographic region. Thus, we assume $\pi_1 = \text{Ded}$. In contrast, if redundancy were prohibited, potentially in the name of fairness, class 2 customers would likely list at whichever transplant center has the *shorter* wait list at the time of listing: They have the resources to travel, and it is easy to see that this is their optimal action in this regime. Thus, we assume that they will join the shorter queue (i.e., $\pi_2 = \text{JSQ}$) in the absence of redundancy.

Throughout this analysis we will focus on class 1 response times weighted by which queue they joined i.e., as an average of the response time of class 1 customers who join server 1 (weighted by p) and those who join server 2 (weighted by $1-p$). This is because we are interested in the class of singly listed customers' preference as a whole within the context of multiple listing. Later, in Figure 2.6, we will consider the impact of redundancy on class 1 customers at each server separately.

Before we study the impact of class 2 customers' joining policy $\pi_2 \in \{\text{JSQ, CUE, CUS}\}$ on class 1 customers' mean response time, we first specify its impact on the mean *overall* response time i.e., the response time taken over all customers, in Theorem 2.1. In addition to being of independent interest, we will use this result in the proof of Theorem 2.2 when we consider fairness.

Theorem 2.1. *When class 1 customers follow the Dedicated policy with $p = 0.5$, $T^{(Ded,CUS)} \leq T^{(Ded,CUE)} \leq T^{(Ded,JSQ)}$, $\forall r \in [0,1]$.*

All proofs appear in Appendix A.1.

Theorem 2.1 states that when the two queues are symmetric, the mean overall response time is smallest when class 2 utilizes CUS, and largest when class 2 utilizes JSQ, among the policy choices available to class 2. Redundant customers improve server utilization by getting served (queueing) at whatever turns out to be the "faster" server in the CUS (CUE) model, effectively balancing queues, thus leading to a decrease in the mean response time taken over all customers.

When class 2 customers utilize CUE, it is easy to see that their routing policy is equivalent to the smallest workload (SW) queue policy. Thus, the second inequality in Theorem 2.1 extends the classical result – the policy that routes all customers (i.e., a single class) to the SW queue results in a smaller overall response time compared to other policies such as JSQ (Foss 1989) – to two classes of customers, one of which is flexible (can be routed either according to the SW queue policy or the JSQ policy) while the other class forms two independent arrival streams (with the same arrival rate) to either queue. While McDonald and Turner (2000) use large deviations methods to show a similar result, our result has a more general scope and is true at all loads.

Our proof of the second inequality in Theorem 2.1 adopts a new coupling approach that allows us to process state transition events out of order, as needed, and a novel counting method to keep track of redundant jobs. Using these tools, we show that the queue length vector (including any customers in service) when π_2 = CUE is weakly submajorized by that when π_2 = JSQ; in other words, CUE results in lower *overall* number of customers in system and queues that are *more balanced.* (See Appendix A.1 for details.) We use the response times from Visschers et al. (2012) and Gardner et al. (2016) for CUE and CUS respectively, to show the first inequality in Theorem 2.1.

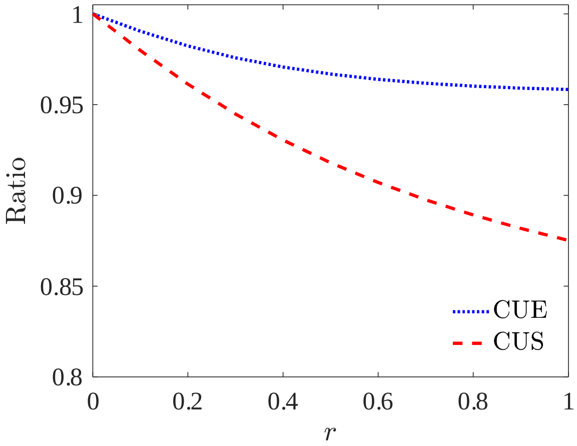
Next, we consider the impact of class 2 customers' joining policies on non-redundant customers i.e., when redundancy is *fair.*

Theorem 2.2. *When class 1 customers follow the Dedicated policy with $p = 0.5$, they prefer CUS over CUE over JSQ for class 2, i.e., $T_1^{(Ded,CUS)} \leq T_1^{(Ded,CUE)} \leq T_1^{(Ded,JSQ)}$, $\forall r \in (0, 1)$. Furthermore, in this case $T_1^{(Ded,CUS)}$ and $T_1^{(Ded,CUE)}$ are decreasing in r.*

Theorem 2.2 states that when $p = 0.5$, class 1 customers prefer redundancy (CUS more so than CUE) no matter what the proportion r or the load ρ is, i.e., redundancy is *fair.* Moreover, they prefer that more jobs be redundant. Figure 2.1 shows class 1 response time under redundancy indexed to π_2 = JSQ when $p = 0.5$. The two curves never exceed 1 which reiterates class 1 customers' preference for redundancy.

To understand the benefit of redundancy, suppose a "tagged" redundant customer enters service at server 2 but would have joined queue 1 if she were non-redundant. Then, non-redundant customers joining queue 1 after the tagged job in the current busy period experience one of the following two "benefits": (i) Under CUE, there is one fewer *real* job ahead of them; (ii) Under CUS, there is one fewer job ahead of them if the tagged job never reaches service at server 1. If the tagged job does enter service at server 1, it is served in half the time (in expectation). But, when $p = 0.5$ the two queues are symmetric, so, intuitively, any such expected benefit experienced by

Figure 2.1: Ratio of $T_1^{(\text{Ded},\pi_2)}$ when $\pi_2 \in \{\text{CUS, CUE}\}$ indexed to $\pi_2 = \text{JSQ}$ for $p = 0.5$ and $\rho = 0.7$.



class 1 customers in queue 1 due to redundancy is offset by the expected pain to class 1 customers in queue 2. Nevertheless, since redundant customers improve server utilization and overall mean system time (Theorem 2.1), the resulting reduction in the length of the busy period of the servers leads class 1 customers to prefer redundancy of class 2 over JSQ.

Theorem 2.2 states that redundancy is fair when $p = 0.5$. But can we say anything for general p? We answer in the affirmative in the following theorem, which shows how class 1 customers' preference changes with the extent of redundancy.

Theorem 2.3. *Suppose class 1 customers follow the Dedicated policy, and the limit of their wait time at either server exists as r approaches 1. Then, the limit of $T_1^{(Ded,\pi_2)}$ as r approaches 1 is given by $\frac{1}{\mu} + \frac{N^{\pi_2}}{2\mu}, \forall p$, where N^{π_2} is the mean number of customers when all customers follow policy $\pi_2 \in \{CUS,\ CUE,\ JSQ\}$. Moreover, in this case class 1 customers prefer CUS over CUE over JSQ.*

Theorem 2.3 presents the limiting value of class 1 response time, and states that redundancy is more fair than JSQ, as r approaches 1, if this limit exists[1]. When $r = 1$ there are no class 1 customers, but, contingent on the limit's existence, in a small neighborhood around $r = 1$, we can analyze fairness: Class 1 customers' preference for redundancy is sustained when the proportion r is sufficiently high. The reason is that if there are *enough* redundant customers, the improvement they bring to system utilization benefits class 1 customers no matter what p is.

So, is redundancy *always* fair? Theorem 2.4 answers this in the negative:

Theorem 2.4. *When class 1 customers follow the Dedicated policy, they prefer JSQ over CUE over CUS for class 2, i.e., $T_1^{(Ded,JSQ)} \leq T_1^{(Ded,CUE)} \leq T_1^{(Ded,CUS)}$, if $p = \eta = 1$ and r approaches 0.*

[1] The existence of the limit depends upon the convergence of the response time as $r \to 1$; conditions guaranteeing this are beyond the scope of this work.

Theorem 2.4 is a limiting result: When $r = 0$ there are no class 2 customers and class 1 response times are independent of π_2, but there is a small neighborhood around $r = 0$ such that class 1 prefers that class 2 follow JSQ as opposed to redundancy. In other words, redundancy becomes "unfair" as the queues become imbalanced and redundant jobs become scarce. To see the intuition assume the queues are highly asymmetric, i.e., p is very high. In the much more common case of queue 2 being shorter, a redundant customer will join both queues and could be served at queue 1 with non-zero probability, thus hurting class 1 customers at the more heavily loaded queue 1 (compared to JSQ). On the other hand, if queue 1 is shorter, class 1 customers at queue 1 could benefit from redundancy (compared to JSQ) since the customer may end up getting served at queue 2 if she were redundant. Since p is high, the former scenario is much more likely to occur than the latter. Thus the net effect is a preference for JSQ among class 1 customers at queue 1, who comprise the overwhelming majority of class 1 customers as $p \to 1$. As before, there may also be a reduction in the busy period of the servers due to redundancy, which *benefits* class 1, but this is outweighed for p close to 1 and r close enough to 0.

Overall, depending on which of the effects is stronger, class 1 may prefer JSQ over redundancy, or vice-versa. Intuitively, preference for JSQ is amplified when (i) p is high, due to increased asymmetry; (ii) ρ is high, due to increased load (and therefore sensitivity of jobs at queue 1 to interference); and (iii) r is low, due to a higher proportion of class 1 customers, which also heightens the asymmetry.

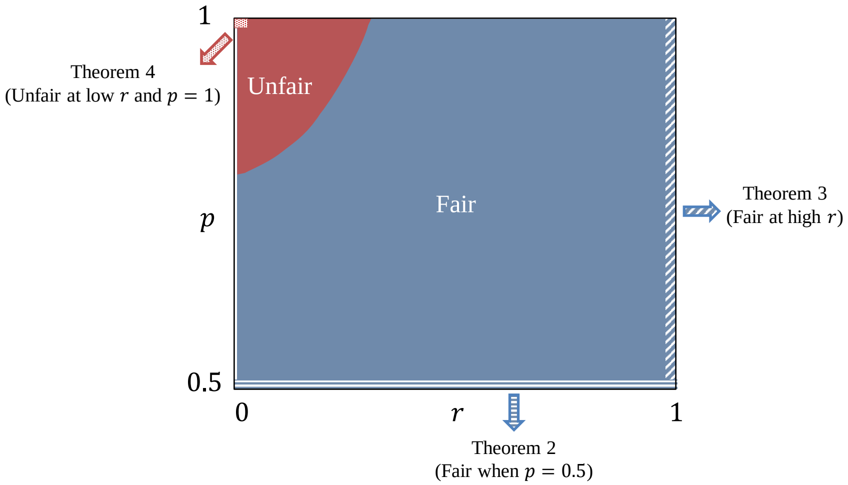
Within redundancy, class 1 customers prefer CUS over CUE when $p = 0.5$ (Theorem 2.2) while the reverse is true when $p = 1$ and r is low (Theorem 2.4). When the two queues are symmetric, class 1 customers prefer the more efficient system i.e., CUS. On the other hand, when $p = 1$ and r is infinitesimally small, the CUE policy is virtually equivalent to the JSQ policy, and thus preference for CUE over CUS is similar to that for JSQ over redundancy described above.

Quantifying Unfairness

We utilize numerical computations to evaluate fairness at general values of p, ρ and r. Specifically, we compare the response time for class 1 under CUS calculated using Gardner et al. (2016), to that under JSQ calculated using Matrix-Analytic methods (Neuts 1981) and formulated as in van Houtum et al. (2001). We assume $\eta = 0.5$ for simplicity throughout this subsection.

Figure 2.2 depicts the combinations of p and r for which CUS is fair (or unfair) for a fixed $\rho = 0.6$ and summarizes our findings: Analytically, we show that redundancy is fair when $p = 0.5$ (i.e., bottom boundary) in Theorem 2.2, and for $r \to 1^-, \forall p$, if the limit exists, in Theorem 2.3 (i.e., right boundary), and that redundancy is unfair when $p = 1$ and r is sufficiently small (i.e., the top left corner) in Theorem 2.4. Our numerical computations corroborate our findings, and in addition, we find that there is a boundary where the preference switches from JSQ to redundancy. We also see that redundancy is increasingly preferred at high r or low p, confirming our intuition based

Figure 2.2: Fairness, comparing CUS against JSQ, when $\pi_1 =$ Ded and $\rho = 0.6$.



Note: "Unfair" region indicates parameters for which redundancy is unfair i.e., $T_1^{(\text{Ded,JSQ})} \leq T_1^{(\text{Ded,CUS})}$. Our analytical results are overlaid with the shaded regions, and the corresponding theorems are also indicated.

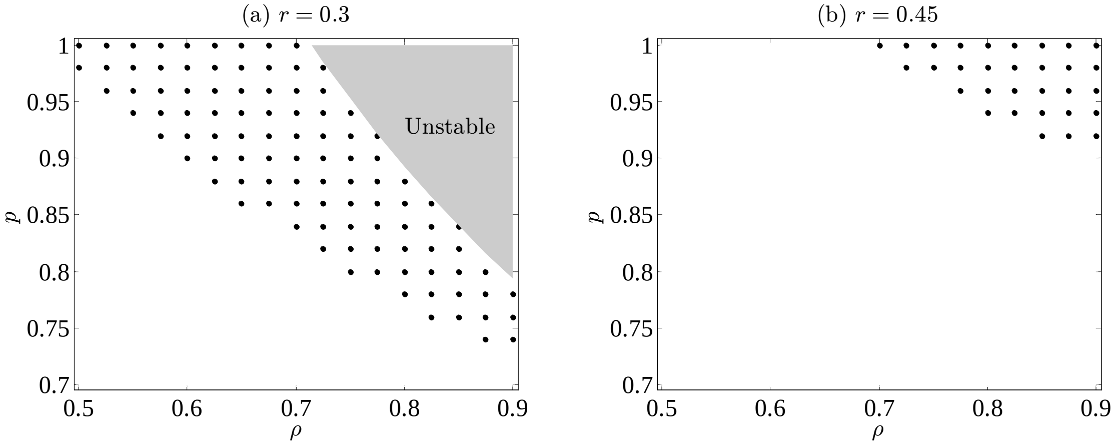
on Theorems 2.2-2.3: Redundancy becomes fair either when r is increased for fixed p (increasing redundancy drives greater efficiency), or when p is decreased for fixed r (greater symmetry balances relative pain and gain, leading to improvements, again, due to increased efficiency).

To see how class 1 customers' preference is impacted by load, Figure 2.3 depicts the combinations of p and ρ for which CUS is unfair, thus revealing class 1 customers' preference along two vertical axes of Figure 2.2 – at $r = 0.3$ (Figure 2.3a) and at $r = 0.45$ (Figure 2.3b) – for varying ρ. At low r (Figure 2.3a) we find a larger range of parameters where JSQ is preferred due to increased asymmetry, whereas when the proportion of the two classes is comparable (Figure 2.3b), JSQ is preferred only at high p and ρ, reiterating our intuition based on Theorem 2.4.

To quantify the magnitude of this preference, we compare the class 1 response time when $\pi_2 \in \{\text{CUS, JSQ}\}$ in Table 2.2. To focus on situations when redundancy is likely to be unfair, we consider $p = 1$ and large ρ. For $r = 0.45$, the response time for moderate values of ρ is comparable for the two policies – 0.2% increase for $\rho = 0.7$; the gap in response time, in both absolute and relative terms, is larger at higher values of ρ – a 1.2% increase for $\rho = 0.85$. There is a drop in the relative difference at very high load. Here, the absolute difference is increasing in ρ, but not as fast as the increase in the mean response times due to increasing load.

Additional experiments indicate that when $p \neq 1$ CUS tends to perform better, so that the difference is not as large. Thus, even though class 1 customers do prefer JSQ over redundancy at certain parameters, the difference is likely not enough to be significant unless delay is very costly, as potentially in the case of organ transplants.

Figure 2.3: Parameters for which class 1 prefers $\pi_2 = \text{JSQ}$ over $\pi_2 = \text{CUS}$, when $\pi_1 = \text{Ded}$.



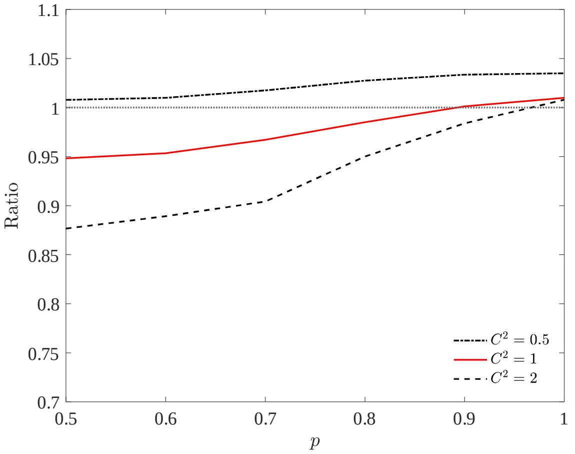
Note: The circles denote points where $T_1^{(\text{Ded},\text{JSQ})} < T_1^{(\text{Ded},\text{CUS})}$. The grey area in (a) represents the region where the system is unstable.

Table 2.2: Class 1 response time $T_1^{(\text{Ded},\pi_2)}$ for $p = 1, r = 0.45$, and $\mu = 1$.

ρ	Preference	$T_1^{(\text{Ded},\pi_2)}$		
		$\pi_2 = \text{CUS}$	$\pi_2 = \text{JSQ}$	Difference (%)
0.65	CUS	4.16	4.17	0.32
0.7	JSQ	5.2	5.19	-0.20
0.75	JSQ	6.86	6.82	-0.69
0.8	JSQ	9.94	9.83	-1.08
0.85	JSQ	17.78	17.57	-1.19
0.9	JSQ	104.01	103.57	-0.42

Note: The percentage difference in column 5 is computed as $(T_1^{(\text{Ded},\text{JSQ})} - T_1^{(\text{Ded},\text{CUS})})/T_1^{(\text{Ded},\text{CUS})}$.

Figure 2.4: $T_1^{(\text{Ded, CUS})}$ indexed to $T_1^{(\text{Ded, JSQ})}$ under general service time distributions when $r = 0.3$ and $\rho = 0.6$.

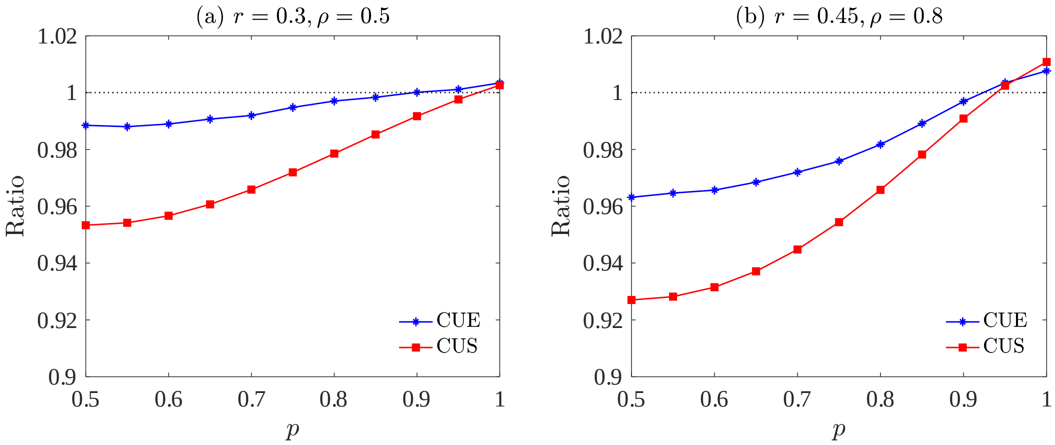


Note: A ratio smaller than 1 indicates that CUS is fair. The curves correspond to service time $S \sim$ Erlang with squared coefficient of variation $C^2 = 0.5$, $S \sim$ Exp with $C^2 = 1$, and $S \sim H_2$ with $C^2 = 2$. The mean service time $E[S]$ is set to 1 in all cases.

Fairness under general service distribution. We conjecture that the benefit from redundancy is larger when the service times have higher variability than an exponential distribution, and in this case, non-redundant customers may prefer redundancy *more often* than suggested by our model. Indeed, we observe this to be true from Figure 2.4, which shows the response time under CUS relative to JSQ for distributions that are more (Hyperexponential with $C^2 = 2$) and less (Erlang with $C^2 = 0.5$) variable as compared to the exponential distribution (i.e., $C^2 = 1$).

Fairness of CUE. Theorems 2.2 and 2.4 together suggest that the preference for class 1 customers between CUE and JSQ of class 2 is similar to that between CUS and JSQ. Figure 2.5 shows class 1 response times (indexed to $\pi_2 =$ JSQ) for two parameter combinations: $r = 0.3, \rho = 0.5$ in Figure 2.5(a) and $r = 0.45, \rho = 0.8$ in Figure 2.5(b). We see that the fairness of CUE is qualitatively similar to that under CUS: CUE is fair when $p = 0.5$, but is less so when the queues are imbalanced. A key difference is that class 1 customers' preference switches from redundancy to JSQ at lower values of p when $\pi_2 =$ CUE than when $\pi_2 =$ CUS. The CUS system tends to outperform CUE due to an additional benefit in the form of the minimum of two service times, when only one class 2 job is present. But, as Theorem 2.4 shows, CUE can be preferred to CUS: The CUE curve dips below the CUS curve in Figure 2.5(b) at $p = 1$. In this case the benefit through improved server utilization is not as dominant as the load is quite high, and class 1 prefers that class 2 not be served at (highly congested) server 1.

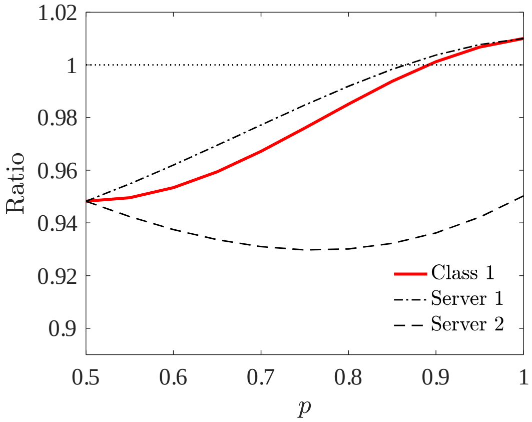
Figure 2.5: Class 1 response time $T_1^{(\text{Ded},\pi_2)}$ when $\pi_2 \in \{\text{CUS, CUE}\}$ indexed to $\pi_2 = \text{JSQ}$.


(a) $r = 0.3, \rho = 0.5$

(b) $r = 0.45, \rho = 0.8$

Fairness at each server. Up until now we have looked at aggregate metrics. But of course class 1 customers at queue 1 (and, likewise, those at queue 2) care much more about how redundancy affects them than class 1 as a whole. Therefore, Figure 2.6 shows the class 1 response time taken over all class 1 customers i.e., $T_1^{(\text{Ded,CUS})}$, as well as that at each server, for varying values of p. (The only value at which all three coincide is $p = 0.5$; when $p = 1$, $T_1^{(\text{Ded,CUS})}$ coincides with that at server 1 because all class 1 customers join server 1.) From Figure 2.6 we see that those class 1 customers at the more congested server increasingly prefer JSQ as the queues become more asymmetric, and they do so at lower values of p than when the entire class would: When $\rho = 0.6$ and $r = 0.3$, class 1 customers at server 1 switch their preference to JSQ closer to $p = 0.85$, whereas the entire class switches at approximately $p = 0.9$. In other words multiple listing may be considered *even more unfair* in a congested region.

Figure 2.6: Class 1 response time – taken over all class 1 and at each server – when $\pi_2 = \text{CUS}$ indexed to $\pi_2 = \text{JSQ}$, for $\pi_1 = \text{Ded}$, $\rho = 0.6$ and $r = 0.3$.



Implications for Multiple Listing

Our model suggests that when the arrival and service rates at two transplant centers are close, multiple listing likely benefits not only those patients who are able to do so, but also singly listed patients as a class and, by extension, the aggregate patient population. Thus, it is *fair.* However, when the two transplant centers are significantly asymmetric, singly listed patients (in aggregate) may be worse off under multiple listing. (Although singly listed patients at one of the centers may be better off.) Thus, it may be more fair to force patients who are capable of multiple listing to join the transplant center with the shorter wait list instead.

2.4.2 Fairness when class 1 uses the JSQ Policy

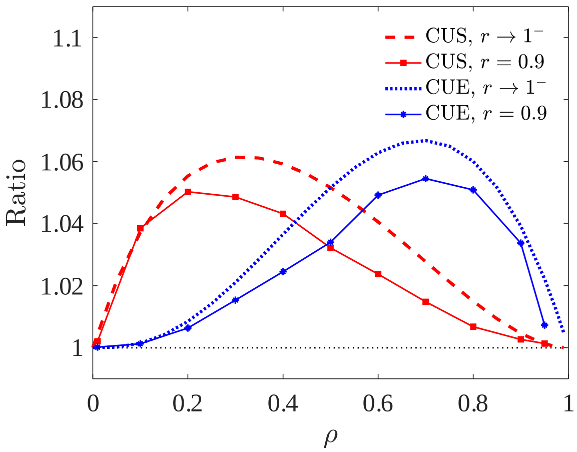
We now assume that the non-redundant customers join queues according to the JSQ policy, i.e., π_1 = JSQ; class 2 customers can potentially list at both ($\pi_2 \in$ {CUS, CUE}) or follow JSQ (π_2 = JSQ). The following theorem specifies conditions under which class 1 *prefers* that one of the two types of redundancy, CUS or CUE, be followed by class 2 versus *the other type of redundancy.*

Theorem 2.5. *Suppose class 1 customers follow the JSQ policy. Then, the limit of* $T_1^{(JSQ,\pi_2)}$ *as* r *approaches 1, when it exists, is given by* $\frac{1}{\mu} + \frac{\rho^2(2-\rho)}{\mu(1-\rho^2)}$ *and* $\frac{1}{\mu} + \frac{\rho}{2\mu(1-\rho)}$ *for* $\pi_2 = CUE$ *and* $\pi_2 = CUS$ *respectively. They prefer* $\pi_2 = CUE$ *if* $\rho < 0.5$ *and* $\pi_2 = CUS$ *if* $\rho > 0.5$ *(they are indifferent for* $\rho = 0.5$*).*

Theorem 2.5 is also a limiting result as r approaches 1 (similar to Theorem 2.3). When π_1 = JSQ, class 1 customers' preference between the two types of redundancy for class 2 customers depends on the load – at low load they prefer π_2 = CUE to π_2 = CUS. To see the intuition behind this, consider the waiting time for an arriving class 1 customer (who utilizes JSQ) when she sees only one customer in either system. In the CUS system, the redundant customer is in service at both servers, so the wait time is $1/2\mu$, whereas in the CUE system the redundant customer is in service at *one* of the two servers, so the wait time is zero. On the other hand, at high loads the number of customers, say n, in either system tends to be large, and the wait time is $n/2\mu$ in both. Since the expected number of customers in a single M/M/1 working at rate 2μ is less than that in an M/M/2 where each server works at rate μ (Wolff 1989), class 1 prefers CUS at high loads.

To evaluate class 1 customers' preference between the JSQ policy and redundancy of class 2 customers, Figure 2.7 shows the ratio of class 1 response times when class 2 customers follow redundancy to that of JSQ i.e., $T_1^{(\text{JSQ},\pi_2)}/T_1^{(\text{JSQ, JSQ})}$, $\pi_2 \in$ {CUS, CUE}, where $T_1^{(\text{JSQ},\pi_2)}$ is given by Theorem 2.5 for $r \to 1^-$ and using simulation for $r = 0.9$, and $T_1^{(\text{JSQ, JSQ})}$ (independent of r) is computed using van Houtum et al. (2001). Since the graphs of CUS and CUE lie above one, class 1 prefers JSQ, i.e., redundancy is unfair, in the limit $r \to 1^-$. Moreover, comparing $r = 0.9$ to $r \to 1^-$, we find that class 1 customers' preference is qualitatively identical to that in the limit: Both curves lie above one and cross each other at $\approx \rho = 0.5$. Thus, class 1 customers prefer JSQ

Figure 2.7: Class 1 response time $T_1^{(\text{JSQ},\pi_2)}$ when $\pi_2 \in \{\text{CUS, CUE}\}$ indexed to $\pi_2 = \text{JSQ}$.



Note: $T_1^{(\text{JSQ,CUS})}/T_1^{(\text{JSQ,JSQ})}$ and $T_1^{(\text{JSQ,CUE})}/T_1^{(\text{JSQ,JSQ})}$ are shown with square and asterisk markers respectively.

over redundancy, and within redundancy, they prefer CUE (CUS) at low (high) loads.

Comparing these results on fairness when $\pi_1 = \text{JSQ}$ to $\pi_1 = \text{Ded}$ and $p = 0.5$ from the previous section, we note two major differences:

1. Class 1 customers prefer $\pi_2 \in \{\text{CUS, CUE}\}$ to $\pi_2 = \text{JSQ}$ under the Dedicated policy, and the opposite is true under the JSQ policy; and
2. Under the Dedicated policy class 1 customers prefer $\pi_2 = \text{CUS}$ to $\pi_2 = \text{CUE}$, while their preference under JSQ switches from $\pi_2 = \text{CUE}$ to $\pi_2 = \text{CUS}$ at $\rho = 0.5$.

The former effect arises because when class 1 customers follow JSQ, they no longer prefer the system with the least total number of jobs as in the Dedicated case; they only care about the least number of jobs in the *shorter* queue. Since redundant jobs balance load across servers, thus lengthening the shorter queue, class 1 customers prefer $\pi_2 = \text{JSQ}$. The intuition behind class 1 customers' preference for $\pi_2 = \text{CUE}$ under the JSQ policy at low loads is similar to the intuition presented after Theorem 2.5 – JSQ enables them to enjoy zero queueing time if they find one server free; they may not experience this advantage if they follow the Dedicated policy. Thus, class 1 customers experience their JSQ benefit only in the CUE system.

2.5 Optimal queue-joining policy

We remove the restriction of class 1 joining policy to $\pi_1 \in \{\text{Ded, JSQ}\}$, examining now class 1 customers' *optimal* queue-joining decision. Specifically, let $\pi_1 = \text{Opt}$ denote the optimal queue-joining policy i.e., the one that minimizes their *individual* system time. (Class 2 customers do not face this question since they join both queues upon arrival.) We consider two possible regimes of state information that arriving class 1 customers may use to make their decision:

1. Fully Observable (*FO*): The customer has information about the class and position of each of the jobs at each server; including which replicated jobs are "paired up."
2. Partially Observable (*PO*): The customer has information about only the *number* of jobs at each server, not any classes.

While it is relatively easy to keep track of queue lengths (the *PO* regime), this may not be true for the *FO* regime. For example, a centralized list of multi-listed candidates, and where they are listed, does not currently exist for transplant patients. On the other hand, in a distributed storage system, *FO* information is readily available and may be utilized to improve performance (Sun et al. 2015 allow scheduling policies to take into account both the number of requests served so far as well as those in service); however, it would involve storing a large amount of data. Thus, the move from a *PO* to a *FO* regime would likely be costly, making the question of the incremental benefit an important one.

For the case with no redundancy (i.e., all jobs are class 1), Winston (1977) showed that JSQ is optimal among policies that can observe queue lengths but not job sizes. This being a special case of our model when $r = 0$, it is logical to consider whether this holds true under redundancy as well. We evaluate the optimality of JSQ in §2.5.1, and then compare its performance against the optimal policy for those cases when JSQ is suboptimal in §2.5.2.

2.5.1 Is JSQ Optimal?

The following proposition answers our question about JSQ's optimality for the *FO* regime.

Proposition 2.1. *In the Fully Observable case – i.e., when an arriving class 1 customer has information about the class of each of the jobs at each server – the optimal policy is state-dependent. In particular, JSQ is not always optimal.*

The suboptimality of JSQ in some *FO* states arises because – depending on the position of the jobs – some jobs in the longer queue may be more likely to vanish before they enter service, effectively decreasing the queueing delay. Joining the longer queue has been shown to be optimal in other settings that are very different from our present model. In a game where customers arriving to two service providers choose strategically – based on unknown service values and observable queue lengths – Veeraraghavan and Debo (2009) show that customers may purchase from the provider with the longer queue. Likewise, Whitt (1986) presents counterexamples to show that there are service time distributions for which it is optimal to join the longer queue.

This brings up the question of optimality of JSQ in the *PO* regime.

Theorem 2.6. *In the Partially Observable case – i.e., if an arriving class 1 customer can observe only the number of jobs at each server – JSQ is optimal: For PO state* $n = (n_1, n_2), n_1 \geq n_2,$

joining queue 2 results in no more mean system time than queue 1, where $n_j \geq 0$ represents the number of jobs, in queue and in service, at server $j \in \{1, 2\}$.

Theorem 2.6 states that, in contrast to the *FO* regime, the optimality of JSQ is preserved when information about only the number of jobs is shared with the arriving customer[2]. Together with Proposition 2.1, this suggests that even though there may exist some underlying *FO* states in which joining the longer queue is delay minimizing, the mass of the JSQ-optimal states dominates those states in which JSQ is sub-optimal. In the next subsection we will quantify this difference.

2.5.2 Value of *FO* Information

We want to determine the value of revealing all state information by evaluating how effective JSQ is as a heuristic for the state-dependent optimal *FO* policy. To do so, we first formulate the problem as a continuous-time Markov chain (CTMC). We then compare $T_1^{(\pi_1,\pi_2)}$ under the optimal ($\pi_1 = \text{Opt}$) and JSQ ($\pi_1 = \text{JSQ}$) policies. We demonstrate this approach for the CUS model (i.e., $\pi_2 = \text{CUS}$). The approach for CUE is similar, and is omitted.

Recall that the state of the system is denoted by $\mathbf{S} = \{\mathbf{S}_1; \mathbf{S}_2\} \in \mathcal{S}$ where $\mathbf{S}_j$ denotes the vector of *classes* of each of the jobs, and let $h_j(\mathbf{S}) \in \{1, 2\}$ be the first element of $\mathbf{S}_j$ for $|\mathbf{S}_j| > 0$, i.e., the class of the job currently in service at server $j \in \{1, 2\}$. Without loss of generality we consider states such that queue 2 is at most as long as queue 1, or $|\mathbf{S}_1| \geq |\mathbf{S}_2|$, reindexing states as necessary. Let $c_i(\mathbf{S})$ denote the number of class $i \in \{1, 2\}$ customers in the system in any state $\mathbf{S} \in \mathcal{S}$.

Suppose a policy selects queue $a \in \{1, 2\}$ to join for every state $\mathbf{S}$.

Below, $\mathbf{S}_a^j$ represents the next state due to arrivals, $\mathbf{S}_d^j, j \in \{1, 2\}$ the next state due to departures, the operator $\mathbf{J} \setminus \{k\}$ removes the first instance of k from vector $\mathbf{J}$, and $\mathbf{I}_{(\cdot)}$ represents the indicator function. With this notation, the transition rates are given by:

1. $p\{\mathbf{S}, \mathbf{S}_a^1 | a\} = \lambda(1 - r)$, $\mathbf{S}_a^1 = \mathbf{S} \circ \{1\} \triangleq \mathbf{I}_{a=1}\{(\mathbf{S}_1, 1); \mathbf{S}_2\} + \mathbf{I}_{a=2}\{\mathbf{S}_1; (\mathbf{S}_2, 1)\}$. Class 1 arrivals join one of the queues according to policy $a \in \{1, 2\}$.
2. $p\{\mathbf{S}, \mathbf{S}_a^2 | a\} = \lambda r, \forall a$, $\mathbf{S}_a^2 = \mathbf{S} \circ \{2\} \triangleq \{(\mathbf{S}_1, 2); (\mathbf{S}_2, 2)\}$. Class 2 arrivals join both queues.
3. $p\{\mathbf{S}, \mathbf{S}_d^1 | a\} = \mu, \forall a$ if $|\mathbf{S}_1| > 0, h_1(\mathbf{S})h_2(\mathbf{S}) \neq 4$, $\mathbf{S}_d^1 = \mathbf{I}_{h_1(\mathbf{S})=1}\{\mathbf{S}_1 \backslash \{1\}; \mathbf{S}_2\} + \mathbf{I}_{h_1(\mathbf{S})=2, h_2(\mathbf{S})=1}\{\mathbf{S}_1 \setminus \{2\}; \mathbf{S}_2 \setminus \{2\}\}$. If there is a service completion at queue 1 and the jobs in service are not both class 2, remove the completed job from queue 1 and its copy if the completed job is of class 2.
4. $p\{\mathbf{S}, \mathbf{S}_d^2 | a\} = \mu, \forall a$ if $|\mathbf{S}_2| > 0, h_1(\mathbf{S})h_2(\mathbf{S}) \neq 4$, $\mathbf{S}_d^2 = \mathbf{I}_{h_2(\mathbf{S})=1}\{\mathbf{S}_1; \mathbf{S}_2 \backslash \{1\}\} + \mathbf{I}_{h_2(\mathbf{S})=2, h_1(\mathbf{S})=1}\{\mathbf{S}_1 \setminus \{2\}; \mathbf{S}_2 \setminus \{2\}\}$. If there is a service completion at queue 2 and the jobs in service are not both

[2] The optimality of JSQ is preserved even if, in addition to the number of jobs at each server, the total number of customers of each class in the system (queue) is revealed in the CUS (CUE) model. In this "Class Observable" (*CO*) regime the state would be (n_1, n_2, n_r), where $n_j, j \in \{1, 2\}$ represents the number in system at server j as in *PO*, and n_r represents the total number of redundant *customers* in system (queue) in CUS (CUE). It follows that $0 \leq n_r \leq n_2$ in CUS and $0 \leq n_r \leq (n_2 - 1)^+$ in CUE. A similar induction as in the proof of Theorem 2.6 (Appendix A.1) – now three-step (on n_1, n_2 and n_r) with boundary states of the form $(i, 0, 0)$ and (i, i, j) – will reveal that JSQ is optimal.

class 2, remove the completed job from queue 2 and its copy if the completed job is of class 2.

5. $p\{\mathbf{S}, \mathbf{S}_d^1 | a\} = 2\mu, \forall a$ if $|\mathbf{S}_1| > 0, |\mathbf{S}_2| > 0, h_1(\mathbf{S})h_2(\mathbf{S}) = 4$, $\mathbf{S}_d^1 = \{\mathbf{S}_1 \setminus \{2\}; \mathbf{S}_2 \setminus \{2\}\}$. If the jobs in service are both class 2, any service completion leads to the departure of both jobs.

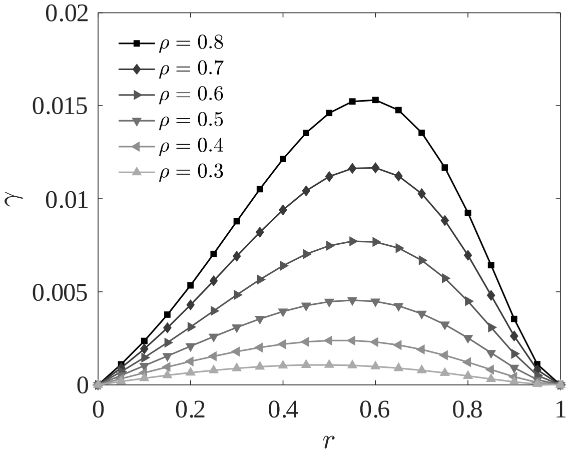
We apply uniformization as in Puterman (1994) with rate $\Lambda \triangleq \lambda + 2\mu$, and consider the discrete time equivalent of the CTMC. Thus, the transition probability from state $\mathbf{S}$ to state $\mathbf{S}' \neq \mathbf{S}$ is $p(\mathbf{S}, \mathbf{S}', a)/\Lambda$ and that from state $\mathbf{S}$ to itself is $1 - \sum_{\mathbf{S}' \neq \mathbf{S}} p(\mathbf{S}, \mathbf{S}', a)/\Lambda$.

Since the number of *FO* states increases combinatorially in the length of the queues, we must limit ourselves to a smaller, more manageable state space. Inspired by the approximation in Nahmias (1977) we propose a modified version of our model where, in addition to a finite number of *FO* states corresponding to at most k jobs in the longer queue, we have a central queue of finite, possibly larger, size M. We use this central queue to track arrivals at the boundary (when there are k jobs in the longer queue) so that we can assign a class – and a joining choice if the job is of class 1 – to these jobs when slots become available in the *FO* state. Since we are potentially delaying the queue-joining decision for the class 1 jobs, our modified model gives us a lower bound on the response time of the original model. Our goal is to compute the gap between the response time under the optimal and the JSQ policies; using the same modified model to evaluate JSQ gives us an approximation for this gap. Therefore for our modified model, the state of the system is given by $(\mathbf{S}, m)$, where $\mathbf{S} \in \mathcal{S}$ represents the *FO* state and $m \in \{0, M\}$ the number of jobs in the central queue, and the transitions $\tilde{p}((\mathbf{S}, m), \cdot | a)$ are provided in Appendix A.2.

For any policy $\mathbf{a} \equiv a(\mathbf{S}, m)$ we can numerically evaluate the stationary probability $\pi^{\mathbf{a}}(\mathbf{S}, m), \forall \mathbf{S} \in \mathcal{S}, m \in \{0, M\}$. Using Little's law we can compute the mean response time for class i as $T_i^{(\mathbf{a},\text{CUS})} = N_i^{\mathbf{a}}/\lambda_i$ where $N_1^{\mathbf{a}} = \sum_{\mathbf{S}\in\mathcal{S}, m\in\{0,M\}} \pi^{\mathbf{a}}(\mathbf{S}, m)\big(c_1(\mathbf{S}) + m(1-r)\big)$ and $N_2^{\mathbf{a}} = \sum_{\mathbf{S}\in\mathcal{S}, m\in\{0,M\}} \pi^{\mathbf{a}}(\mathbf{S}, m)(c_2(\mathbf{S}) + mr)$. The overall response time is therefore $T^{(\mathbf{a},\text{CUS})} = (1-r)\, T_1^{(\mathbf{a},\text{CUS})} + rT_2^{(\mathbf{a},\text{CUS})}$. In order to numerically evaluate the optimal policy (and separately, JSQ) we fix $k = 5$. We allow M to be a function of the arrival rate λ so that virtually all of the stationary probability is contained within M slots of the central queue: For a given value of λ, we start with $M = 0$ and iteratively increase M until the difference between the resulting overall response time for $r = 1$ and the exact response time (M/M/1) is at most $\epsilon = 0.0001$. To identify the individually optimal policy $\pi_1 = \text{Opt}$ we compute the mean wait time for an arriving class 1 customer at either queue for all *FO* states (as in the proof of Proposition 2.1 in Appendix A.1).

Let $\gamma \equiv T_1^{(\text{JSQ},\text{CUS})} - T_1^{(\text{Opt},\text{CUS})}$ denote the gap between our approximation for the class 1 response time for the optimal and JSQ policies. Figure 2.8 shows γ for different values of ρ. If the proportion of one of the two classes is much higher than the other the two policies are very similar in their performance – the gap for $\rho = 0.5$ is 0.001 when $r = 0.1$ or $r = 0.9$. When the proportion of both the classes is comparable i.e., $r \in [0.4, 0.7]$, JSQ experiences the greatest deviation; the difference achieves a maximum of 0.015 at $\rho = 0.8$ and $r = 0.6$. Intuitively, in order for JSQ to be non-trivially suboptimal for an arriving class 1 customer, we need a staggered arrangement of class

Figure 2.8: Difference γ in class 1 response time $T_1^{(\cdot,\text{CUS})}$ under the optimal and JSQ policies ($\mu = 1$).



2 jobs (such as in the proof of Proposition 2.1) which makes those in the longer queue more likely to depart prior to reaching service. To ensure this we need a sufficient number of class 2 customers to be present; thus, γ skews right of $r = 0.5$. Finally, at higher loads these arrangements are more likely, and any such deviation between the two policies gets amplified.

Since $\gamma \leq 0.015$ the average difference is small, but this does not tell the entire story; the difference in any particular state could be quite significant. For example, the mean wait times are 1.25 and 1.75 in queues 1 and 2 respectively, for a class 1 customer arriving to the state $\mathbf{S}_1 = (1, 2), \mathbf{S}_2 = (2, 1)$; under JSQ, the customer would experience a wait time of 1.5. Thus, revealing "full" information could result in 17% savings in wait time for this particular customer; but, the probability of being in this state is only 0.012 – about one-fifth that for state $\mathbf{S}_1 = (1, 2), \mathbf{S}_2 = (2)$ – when $r = \rho = 0.6$.[3] These examples provide concrete values for how large the value of information can be *in a given state*; however, finding tight bounds on γ remains an open question.

Finally, since the system time obtained for the JSQ policy using the modified model is on average only 0.03% off from that obtained from simulating the *original* model, we believe the modified model provides a very close lower bound for the JSQ policy. In fact, the gap between the modified model's optimal policy (i.e., a *lower bound* for the optimal policy) and the upper 95% confidence interval (c.i.) bound on the original model for JSQ from simulation is less than 3%, suggesting that JSQ is a very good heuristic in general. Combined with the fact that there may be overhead costs associated with tracking state information, this suggests that it is advisable to use JSQ except in those situations in which information is inexpensive and/or delay is extremely costly. Note that,

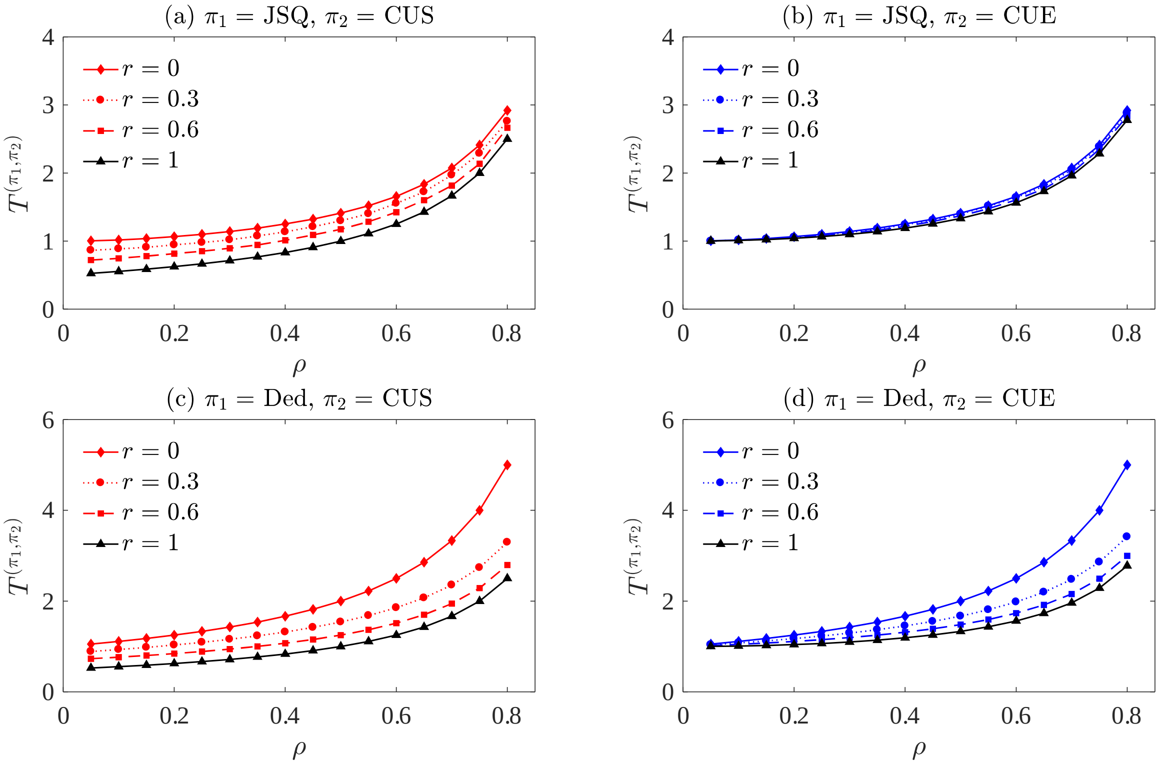
[3] We provide two examples – one with very few customers and one with infinitely many in Appendix A.2 – to demonstrate settings in which this absolute difference is unbounded, and their ratio approaches constants between 1.46 and 1.67. These provide concrete values for how large the value of information can be *in a given state*; however, finding tight bounds on γ remains an open question.

arguably, organ transplantation may very well fall into this latter case.

2.6 System performance

In this section we numerically evaluate the benefit to the overall system from redundancy: We compare the overall response time $T^{(\pi_1,\pi_2)}$ under different policies π_1 and π_2 to identify when this benefit is significant. Our analysis focuses primarily on low to high loads; other techniques may be available to study the system in heavy traffic.

Figure 2.9: Overall system time when class 2 customers constitute a fraction r of all arrivals.



Note: $p = 0.5$ for π_1 = Ded.

Figure 2.9 compares $T^{(\pi_1,\pi_2)}$ when class 2 customers form a fraction $r \in \{0, 0.3, 0.6, 1\}$ of the total arrival stream. When class 1 customers follow π_1 = JSQ, Figure 2.9(a)-(b) shows that redundancy improves the overall system time in both π_2 = CUS and π_2 = CUE settings; however, the improvement is considerably smaller in CUE. The extremes of the CUE model – JSQ at no redundancy ($r = 0$) and M/M/2 at full redundancy ($r = 1$) – have comparable performance at all loads, whereas the extremes for CUS, JSQ ($r = 0$) and M/M/1 serving at rate 2μ ($r = 1$), have a significantly wider dispersion. In contrast, when class 1 customers follow the Dedicated policy (π_1 = Ded) with $p = 0.5$, Figure 2.9(c)-(d) shows that there is considerable improvement in the overall system time under both CUS and CUE, particularly at higher loads, since both M/M/2 and M/M/1 queues serving at rate 2μ ($r = 1$) are more efficient than two M/M/1 queues serving at rate μ each ($r = 0$).

To explore the benefits of partial redundancy more closely, Table 2.3 shows how much of the benefit of full redundancy is captured by having some fraction of redundant customers. Let $t_r \equiv T^{(\pi_1,\pi_2)}$ be the overall system time when the proportion of class 2 is r. Specifically, Table 2.3 shows

$$\tau(r) \equiv \frac{t_0 - t_r}{t_0 - t_1}, \tag{2.1}$$

calculated analytically for π_1 = Ded and via simulation (with 95% c.i.) for π_1 = JSQ.

Table 2.3: Benefit $\tau(r)$ (%) from having a fraction r of redundant customers.

π_2	ρ	π_1	r				
			0.1	0.3	0.5	0.7	0.9
CUS	0.6	Ded	22	52	71	85	96
		JSQ	11 (10-12)	30 (30-31)	49 (48-50)	69 (68-70)	89 (88-90)
	0.9	Ded	53	81	91	96	99
		JSQ	14 (11-17)	35 (33-38)	57 (54-60)	71 (68-74)	94 (91-97)
CUE	0.6	Ded	24	55	75	88	96
		JSQ	9 (6-12)	31 (27-34)	53 (50-56)	72 (69-74)	87 (84-90)
	0.9	Ded	55	83	92	97	99
		JSQ	17 (12-22)	46 (41-52)	66 (61-71)	81 (76-86)	96 (91-100)

Note: Benefit $\tau(r)$ is defined in (2.1). The values for π_1 = JSQ correspond to the mean, and the 95% c.i. in parenthesis, obtained via simulation. $p = 0.5$ for π_1 = Ded.

When π_1 = Ded and $p = 0.5$ the reduction in system time is dramatic; for instance, attaining 53% of the benefit with only 10% redundancy at $\rho = 0.9$ for CUS. The corresponding benefit at $\rho = 0.6$ is 22%, indicating that the higher the load, the greater the benefit. Furthermore, as r increases the % benefit likewise increases rapidly. The % benefit is slightly larger for the CUE system, but recall from Figure 2.9 that these are larger percentages of a relatively smaller total benefit. In further experiments (not shown), we find that the % benefit is significantly greater when the queues are asymmetric under the Dedicated policy: 56% and 93% of the benefit are attained at 10% and 50% redundancy respectively, when $\rho = 0.6$ and $p = 0.75$ in the CUS model.

Table 2.3 also shows that when class 1 jobs follow JSQ instead of Dedicated, the impact of redundancy is not as dramatic; as class 1 jobs already experience the benefits of JSQ, redundancy does not benefit them as much. Redundancy is thus an important lever to improve response times, particularly in less efficient systems (such as Dedicated) as compared to JSQ, and even more so at higher loads and more asymmetric systems.

Redundancy and JSQ are similar in that they both make a joining decision to minimize some form of delay – actual system time for CUS, actual queueing time for CUE, and expected queueing and system time for JSQ. This observation invokes a comparison of our results to van Houtum et al. (2001) where they consider two classes of jobs: 1) common jobs (fraction r) that can be served by all servers and are assigned to queues according to JSQ; and 2) specialized jobs that form a dedicated stream to each server. Their reported "realized reduction" in the gap (similar to (2.1)) – 54% at $r = 0.1$, 82% at $r = 0.3$, and 91% at $r = 0.5$ for $\rho = 0.9$ – is very similar to ours under the Dedicated policy, implying that the underlying dynamics are similar i.e., the presence of few delay minimizing "common" or redundant jobs improves the overall response time significantly.

2.7 Extensions

2.7.1 Larger Systems

We now consider a system with m identical servers, denoted $j \in \{1, \ldots, m\}$, each with its own queue, two classes of customers, denoted $i \in \{1, 2\}$, and "s-Redundancy (sR)" in which a class 2 customer replicates into $s \leq m$ jobs upon arrival so that each job joins the *same* dedicated subset of queues, say $k \in \{1, \ldots, s\}$.[4] Let π_i be the queue-joining policy adopted by class i customers, p_j be the proportion of class 1 customers joining queue $j \in \{1, \ldots, m\}$ under $\pi_1 =$ Ded, and let η_j be the probability that a class 2 customer joins queue j when more than one of the queue lengths are equal under $\pi_2 =$ JSQ, or when more than one of the servers are idle under $\pi_2 =$ CUE.

Our main structural results from §§2.4-2.5 continue to hold for these larger systems.

Theorem 2.7. *Suppose class 1 customers follow the Dedicated policy, $s = m$, and the limit of their wait time at any server exists as r approaches 1. Then the limit of $T_1^{(Ded,\pi_2)}$ as r approaches 1 is given by $\frac{1}{\mu} + \frac{N^{\pi_2}}{m\mu}, \forall p_j$, where N^{π_2} is the expected number of customers in a system of m parallel queues where all customers follow policy $\pi_2 \in \{CUS, CUE, JSQ\}$. Moreover, in this case class 1 customers prefer CUS over CUE over JSQ.*

As before, the preference is due to the reduction in the length of the busy periods.

Unfortunately, our proof fails for the extension of Theorem 2.2; while we have not been able to construct a counterexample where the coupling approach in the proof of Theorem 2.1 (which is

[4] There are other ways to extend our model to larger systems: one option is to allow redundant customers to join any (randomly chosen) $s < m$ queues upon arrival. This is a significantly harder problem, and extending our results to this particular type of redundancy would be non-trivial.

used in Theorem 2.2) breaks, we have also not been successful in extending the argument. We have been able to generalize Theorem 2.4:

Theorem 2.8. *When class 1 customers follow the Dedicated policy, they prefer JSQ and CUE over CUS when $p_1 = 1, \eta_1 = 0$ and r approaches 0.*

Theorem 2.8 suggests that in the extended model – as in the two-server model – redundancy may be unfair, particularly if the queues are asymmetric.

We now turn to the optimal queue-joining policy for class 1 customers: As in the two-server model, we can easily construct counterexamples to show that JSQ is not always optimal in the *FO* regime. But JSQ continues to be optimal in the *PO* regime:

Theorem 2.9. *In the FO regime, JSQ is not always optimal. In the PO regime, JSQ is optimal if $s = m$, but it is not always optimal if $s < m$.*

2.7.2 Abandonments

In this subsection we will add abandonments to our model. This is particularly important in the context of organ transplants where patients may leave the wait list before receiving an offer because they become too sick to receive a transplant, they receive a living donor organ, or due to death. We assume that each customer (regardless of her class) will leave the system after some uncertain time, which is exponentially distributed with rate α. (A redundant customer has a single deadline.) We allow abandonments to occur from the system; thus, a customer may leave even if she is in service[5]. This is justified if we consider service time as the time between consecutive organs becoming available.

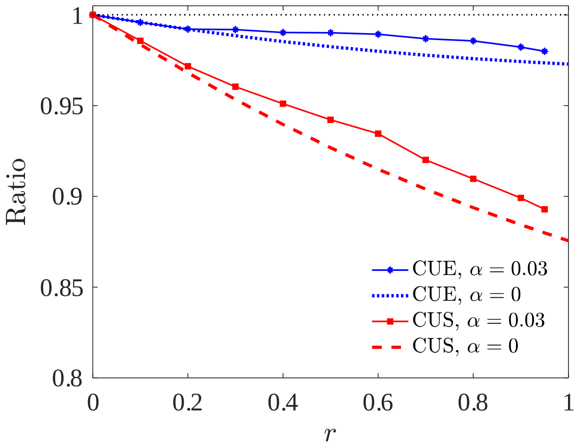
In addition to showing that exponential service times and Poisson arrivals are valid assumptions in the context of organ transplants, Davis et al. (2013) also show that individual abandonment time is also exponentially distributed. The only related work in redundancy that models abandonments is Ata et al. (2016), who study an overloaded queueing system using fluid and diffusion limits. To the best of our knowledge there is no other work that considers the fairness of redundancy in the presence of abandonments.

Let $A^{(\pi_1,\pi_2)}$ denote the mean number of abandonments from the system when class i uses policy $\pi_i, i \in \{1, 2\}$. Note that $T^{(\pi_1,\pi_2)}$ ($T_i^{(\pi_1,\pi_2)}$) now denotes the overall (class 1) average time in system – either before service is complete or until the customer abandons[6]. The following theorem partially extends one of our main results to include abandonments.

[5] One may consider deadlines to the beginning of service so that customers do not abandon if they are in service; however, this typically imposes additional conditions (such as $\alpha < \mu$) for similar stochastic results to hold (Akgun et al. 2011, Sparaggis and Towsley 1994).

[6] We use this definition for two reasons: 1) It allows us to derive an ordering on the system times; and 2) Other work that consider customer abandonment such as Whitt (2005) and Akgun et al. (2012) use the same definition.

Figure 2.10: Class 1 response time $T_1^{(\text{JSQ},\pi_2)}$ when $\pi_2 \in \{\text{CUS, CUE}\}$ indexed to $\pi_2 = \text{JSQ}$ when $\rho = 0.5$ and $p = 0.5$, in the presence ($\alpha = 0.03$) and absence ($\alpha = 0$) of customer abandonments.



Theorem 2.10. *Let each customer abandon the system after an exponentially distributed time at rate α. When class 1 customers follow the Dedicated policy with $p = 0.5$, $T^{(Ded,CUE)} \leq T^{(Ded,JSQ)}$ and $A^{(Ded,CUE)} \leq A^{(Ded,JSQ)}$, $\forall r \in [0, 1]$.*

When the queues are symmetric, Theorem 2.10 states that CUE-redundancy results in lower mean system times – and mean number of customers in system by Little's Law – as compared to JSQ, for a system with abandonments. To show this result, we provide a novel coupling approach that includes abandonments (see Appendix A.3). Importantly, Theorem 2.10 also shows that redundancy results in fewer mean number of abandonments, which implies that the number of service completions is higher when $\pi_2 = \text{CUE}$ as well.

Since we no longer have a closed-form expression for CUS under abandonments to compare against CUE, and our coupling approach does not extend to CUS[7], we cannot show the corresponding result $T^{(\text{Ded,CUS})} \leq T^{(\text{Ded,CUE})}$ to be true.

With respect to fairness, it is no longer immediate that class 1 customers prefer a more balanced system: For example, when $\pi_1 = \text{Ded}, p = 0.5$ and $\alpha > 0$, the state $(3, 1)$ results in a smaller response time for an arriving class 1 customer than the more balanced state $(2, 2)$. This is because while each individual customer has the same rate of abandonment from the system, longer queues are likely to witness more abandonments, so that abandonments effectively "shorten" long queues. Thus, it is not trivial to extend Theorem 2.2, i.e., $T_1^{(\text{Ded, CUE})}$ versus $T_1^{(\text{Ded, JSQ})}$, purely from Theorem 2.10 using Little's law and PASTA as we did in §2.4.

We numerically find that Theorem 2.2 holds when the abandonment rate is small[8]: Figure 2.10

[7] This is because when the two copies of a class 2 customer are both in service, the weak submajorization ordering used in the proof of Theorem 2.1 breaks.

[8] When α or ρ is large we conjecture that redundancy can be unfair for $p = 0.5$. In this case the preference of class 1 for unbalanced systems (described above) is likely to dominate the overall system efficiency effect in Theorem 2.10.

shows class 1 response time under redundancy (CUS and CUE) indexed to JSQ for $p = 0.5$, both when abandonments are considered ($\alpha = 0.03$) and when they are absent ($\alpha = 0$). The gap between the $\alpha = 0$ and $\alpha = 0.03$ curves is precisely due to the decrease in response times owing to abandonment. The indexed class 1 response times lie below one, indicating that redundancy can be fair under both CUS and CUE, if $p = 0.5$ and α is small.

This being said, we can still show that redundancy can be unfair to class 1.

Theorem 2.11. *Let each customer abandon the system after an exponentially distributed time at rate α. When class 1 customers follow the Dedicated policy, they prefer JSQ over CUE over CUS for class 2, i.e., $T_1^{(Ded,JSQ)} \leq T_1^{(Ded,CUE)} \leq T_1^{(Ded,CUS)}$ and $A_1^{(Ded,JSQ)} \leq A_1^{(Ded,CUE)} \leq A_1^{(Ded,CUS)}$, if $p = \eta = 1$ and r is sufficiently close to 0.*

Theorem 2.11 extends Theorem 2.4: Even though customers abandon from the system, the fundamental issue of redundant customers sometimes being served at the congested server, thus impacting non-redundant customers at that server, continues to exist. When the queues are asymmetric enough, this effect causes redundancy to be less fair compared to JSQ.

The implication for multiple listing is that redundancy is better on two levels when queues are symmetric: (i) It reduces the number of patients leaving the wait list before receiving an organ; and (ii) Patients, on average, spend less time on the wait list. But, when queues are asymmetric, it may be more fair to have flexible patients join the shorter queue instead of multiple listing.

2.7.3 Unequal Service Rates

We have considered equal service rates at the two servers throughout this chapter. This is a common assumption in related work, such as Sparaggis and Towsley (1994) and Akgun et al. (2011, 2012), that involve comparisons of performance measures using stochastic ordering or optimality of intuitive routing policies, as we do. Nevertheless, we now extend our system to consider unequal service rates. Similar to §2.7.2, this may be justified in organ transplantation models because different regions are likely to have different rates of organ availability. Let μ_j be the service rate at server $j \in \{1, 2\}$.

We need to modify our queue-joining under policies $\pi_2 \in \{\text{JSQ, CUE}\}$ appropriately: If only one server is idle, a class 2 customer will enter service at that server. Under $\pi_2 = \text{JSQ}$, a class 2 customer will join the shorter queue i.e., the queue with *fewer* customers. Under these assumptions, the following theorem shows that redundancy may be unfair.

Theorem 2.12. *Let the service rate at server $j \in \{1, 2\}$ be μ_j. When class 1 customers follow the Dedicated policy, they prefer JSQ over CUE over CUS for class 2, i.e., $T_1^{(Ded,JSQ)} \leq T_1^{(Ded,CUE)} \leq T_1^{(Ded,CUS)}$, if $p = \eta = 1$ and r approaches 0.*

If the arrival rates to the two queues are very different, Theorem 2.12 shows that class 1 customers prefer JSQ to redundancy when the service rates are unequal. The intuition behind this result is similar to that of Theorem 2.4. This confirms that our results on the unfairness of multiple listing are robust in the more realistic scenario where the service speeds in the two regions is different.

From the proof of Theorem 2.12 (Appendix A.3), we know that the additional pain when $\pi_2 =$ CUS (compared to when $\pi_2 =$ CUE or $\pi_2 =$ JSQ) is $\frac{1}{\mu_1+\mu_2}$ to (potential) class 1 customers, and this pain occurs with probability $\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^n$, in any state with n customers at server 1. Therefore, the unfairness of CUS is likely to be decreasing in μ_2 for constant μ_1. In other words, if the less congested transplant center is also the one with the slower rate of organ arrivals (which is likely given that population likely impacts both the arrival and the service rates in the same direction), then singly listed patients are likely to be *more* worse off than suggested by our base model with equal service rates.

It is worth noting that, even with exponentially distributed service times and a single class of (flexible) customers, the policy that minimizes individual mean system time is not optimal for the system as a whole, when the service rates are unequal: Krishnan (1987) shows that an alternate 'separable rule' performs better than the individually optimal policy. Identifying the individually and socially optimal routing policies – for either customer class – in the presence of redundancy and unequal service rates remains an open problem, and it will be interesting to investigate whether Theorems 2.1 and 2.2 extend (albeit with some modifications) as well.

2.8 Conclusion

We study a system of two parallel queues with two classes of customers: redundant customers join both queues and non-redundant customers join a single queue according to some policy. We allow for two different variants of our model, differentiated by when the redundant jobs are deleted from the system: when the required service is complete, or when it begins, thereby modeling a diverse set of applications. By incorporating policy choices of class 1 customers and considering class-based redundancy, our model provides fundamental insights on the fairness of redundancy, the optimal queue-joining policy for non-redundant customers, and the value of information.

We show that non-redundant customers forming independent Poisson streams to each queue are *always* better off when the other class is allowed to be redundant, as opposed to utilizing JSQ, if the queues are symmetric. However, if the queues are asymmetric there may be situations in which the non-redundant class is worse off under redundancy. This implies that in settings such as kidney transplantation, multiple listing may benefit not only those patients who are able to multi-list but also singly listed patients, when the transplant centers are symmetric (or nearly so). But such multi-listing may harm singly listed patients if centers are asymmetric – implying that it might be more fair to have patients who possess the means to multi-list join the transplant center with the shorter wait list instead. Our analysis helps identify when, and by how much, multiple listing may be "unfair," and as such could be a useful tool for policy makers.

Our analytical results on the optimal queue-joining policy show that different policies may be optimal under different information regimes: If an arriving non-redundant customer can observe only the total *number* of jobs at each server, JSQ is optimal, but if the customer can observe the *class* of each of the jobs at each server, JSQ is not always delay minimizing. But, even in this latter case, JSQ is near-optimal. Thus, considering the potential costs of gathering such information, it may not be worthwhile to do anything but JSQ, unless the setting is one in which delay is very costly, such as waiting for a kidney transplant.

There are several interesting directions for future research. Customers may incur a cost for generating redundant jobs; for example, evaluation costs associated with being listed at more than one transplant center, or communication costs for sending jobs to nodes. How does the presence of such costs affect the decision to be redundant under CUE and CUS? Our simplifying assumptions of Poisson arrivals and exponential service times, while enabling us to derive structural properties of this complex problem, may not always hold true in practice. One may also consider other scheduling disciplines instead of FCFS, or strategic redundancy instead of an exogenous proportion of redundant customers, particularly if redundancy comes with a cost. And, even though our model can be used to describe the dynamics of a kidney transplant center with multiple listing at a high level, there are many aspects of this problem such as the complex matching system assessing the compatibility of a donor kidney, organ allocation policies, and post-transplant survival outcomes that would be important additions to our model.

Chapter 3

Consumer Return Policies in Omnichannel Operations

3.1 Introduction

The retail landscape is rapidly evolving in response to changing customer preferences: While same-store sales have been declining due to decreasing store traffic (Bloomberg 2016), online sales have been growing steadily (Centre for Retail Research 2017). Fortunately, the omnichannel strategy – providing an integrated shopping experience through physical stores and digital channels – allows a firm to effectively participate in both channels, yielding advantages over single-channel peers. An omnichannel firm has advantages over conventional brick-and-mortar retailers, since it is appealing to customers who value the convenience of online shopping. Likewise, an omnichannel firm enjoys a competitive edge over pure-play online retailers – it can leverage its store assets to cater to customers who value the ability to test the "fit" of an item before purchasing.

This latter advantage is significant – many customers feel that the biggest drawback of shopping online is the inability to accurately assess the fit of an item before purchase (Body Labs 2016). To overcome this drawback customers must either visit a retail location or order the item and check the "fit" at home.[1] The prevalence of this latter strategy is borne out by the fact that a third of all Internet sales get returned (WSJ 2013). It is not surprising then that customers value the option of returning, with a majority (66% according to UPS 2014) reviewing the return policy *before* making their purchasing decision. But, handling these returns is a costly affair for retailers – the average cost of processing a returned item is \$30–\$35 – *and* these items may need to be salvaged at a loss; for example, third party outlet stores buy returned items for only 10-20% of their value (Stock et al. 2006).

This leads to yet another advantage omnichannel firms enjoy compared to their single-channel competitors: Omnichannel firms can offer cross-channel returns. Again, customers value this flexibility: UPS (2015) finds that over 50% of shoppers would not complete their purchase if "return to store" were not allowed.

Interestingly, we see a wide range of return policies offered by omnichannel firms, from the "no questions asked" full refunds for all returns policy (e.g., Nordstrom and Macy's) to charging a fixed

[1] Customers may indulge in showrooming i.e., buying online after assessing "fit" in-store. Some firms also offer physical showrooms – stores set up for display purposes where customers can try on items – and virtual showrooms – programs that allow them to try on items virtually.

return fee (e.g., $5.99 offered by H&M) for returns by mail, to the no returns policy (e.g., Gap and J.Crew for "clearance" items). And, while most omnichannel firms offer the additional "free returns to store" option of returning (non-clearance items) to their physical stores for a full refund, the convenience of this option ranges widely, depending on the number of stores and the ease of the in-store return process. For example, while some retailers offer store credit instead of refunds for in-store returns – which may be seen as a significant inconvenience – others (e.g., J.C. Penney) seem to encourage in-store returns as "the quickest and easiest way to return."[2]

Unfortunately, research on return policies has largely focused on single channel firms; the return policy decision in an omnichannel setting remains an open problem. In fact, our analysis reveals that while some results obtained by prior research on single channel firms apply to the omnichannel firm, others do not. Thus, our main contribution in this chapter is to model and analyze how an omnichannel firm makes return policy (and pricing) decisions when serving customers who interact across its channels. We specifically ask the following questions:

- Do the different return policies we see in practice, such as full refunds, partial refunds, and in-store returns, arise as the optimal policy of a profit-maximizing omnichannel firm? In particular, are the generous refunds seen in practice advisable?
- What are the key characteristics of the omnichannel firm that influence its preference for key return policies such as full refunds and incentivizing in-store returns?
- What impact does improving customers' accessibility to stores (e.g., by opening more physical locations) and the convenience of in-store returns (e.g., not requiring a receipt or having a dedicated area for processing returns) have on the firm's overall profit?

To answer these questions, we consider an omnichannel firm that sells a product to customers who face uncertainty around their valuation for the product. These customers differ in how they resolve their uncertainty – some evaluate the product in person at a physical store before deciding to purchase, while others purchase online and may or may not return misfit items depending on the firm's return policy. Keeping with the practices of most omnichannel firms, the firm sets identical prices across its channels and allows free returns to its stores.

Our main contributions are summarized below:

- We demonstrate that all three return policies – full refunds, online returns for a fee, and free in-store returns – may be prescribed by a profit-maximizing omnichannel firm. Prior work considering *single-channel* firms concluded that the refund must be set equal to the salvage value. In contrast, we show that when customers return online, the optimal refund chosen by an *omnichannel* firm is more generous: It is *larger* than the salvage value, and it may even be a *full* refund. We attribute these high refunds to two key drivers not considered in prior work – the mixed customer base (in-store and online) and the convenience of returning items in-store. Thus, we are able to explain the negligible or zero fees for online returns often offered by omnichannel firms in practice.

[2] Source: `https://www.jcpenney.com/dotcom/jsp/customerservice/serviceContent.jsp?pageId=pg100026`.

- Our results help omnichannel firms decide when to offer full refunds or charge a fee for online returns. We show that an omnichannel firm should offer full refunds whenever it is able to salvage online returns at mild discounts, or if a majority of its customers choose to visit its stores before purchasing. The former result extends prior work considering *single-channel* firms, and can explain why firms that have their own outlet stores, e.g., Nordstrom and Nordstrom Rack, offer full refunds. But, through the latter result, we are able to explain why firms such as Macy's that use third-party outlet stores for salvaging (e.g., T.J. Maxx, likely at low values) offer full refunds as well: We attribute it to Macy's share of store sales being more significant than Nordstrom.
- We recommend that omnichannel firms who enjoy better salvage capabilities in-store, and a sufficiently wide physical store network, charge a fee for online returns in order to prompt customers to utilize free in-store returns. For example, J.C. Penney claims to have strong in-store cross-selling capabilities and possesses a wide store network. And, in line with our results, they offer partial refunds online, but advertise free in-store returns, and reportedly over 90% of their (online-purchased) returns take place in-store.
- By introducing two model elements not considered together in prior work – the convenience of returning in-store and the share of customers who choose to shop online – we are able to explore the impact of improving customers' *accessibility* to stores. Prior work empirically finds that an omnichannel retailer who offers full refunds sees increased revenue as a result of opening a new store. This is in line with our results for a firm offering full refunds, and we attribute it to either more customers visiting the physical store before purchasing or to more customers finding "free returns to store" convenient. In addition, we find that this increase in revenues may not prove to be true in general, and we caution that, depending on the return policy in place, measures that improve accessibility or the convenience of returns may actually hurt an omnichannel firm.

The rest of the chapter is organized as follows. In §3.2, we review the relevant literature and describe how it differs from our present work. We formalize our model in §3.3 and study the two extreme policies of full refunds and no returns, that are commonly considered in prior work, in §3.4. We derive the optimal policy for general refunds, and illustrate our results using realistic parameter values that represent a typical omnichannel setting, in §3.5. We explore several extensions in §3.6 and finally, in §3.7, we conclude and discuss future research directions.

3.2 Literature review

Our work is primarily related to two streams of research: 1) return policies of *single-channel* retailers; and 2) operational problems of *omnichannel* firms. The literature on consumer returns largely highlights the positive sales effects of a lenient return policy. Early research by Davis et al. (1995) and Che (1996) studied two extreme policies – full and no refunds – to identify when a firm benefits from money-back guarantees. Moorthy and Srinivasan (1995) show that full refunds may also signal quality, and Petersen and Kumar (2009) find that a satisfactory return experience results in higher levels of customer trust.

More recent papers study optimal return policies of *single-channel* firms: Su (2009) considers a firm selling to ex-ante homogeneous customers who realize their valuation after purchase, and shows that the optimal refund is *equal* to the salvage value. Moreover, he identifies two extensions wherein the firm offers generous refunds i.e., the optimal refund exceeds the salvage value: (i) when customers are heterogeneous in their uncertainty before purchase; or (ii) if customers are boundedly rational (or, forgetful) and naively believe that they will return items as originally planned. In contrast to Su (2009) we show that an omnichannel firm offers generous refunds in the absence of these additional features, thus providing an alternate explanation for the same. Su (2009) also finds the optimal inventory to stock and the impact of return policies on supply chain performance. In a similar vein, Hsiao and Chen (2014) assume that customers have ex-ante heterogeneous valuations, but they allow the firm to decide the "hassle" for returns, i.e., how inconvenient it is to return an item, and compare two special policies – full refunds with optimally chosen hassle and hassle-free policies that provide partial refunds.

Matthews and Persico (2007) find a similar result to Su (2009) i.e., optimal refund is equal to salvage value, even when customers can choose to acquire costly information about the product on their own; however, the firm may promise a larger refund to dissuade customers from seeking this information. In earlier work Matthews and Persico (2005) consider two customer types such that one type has negligible costs of obtaining this information, which is mathematically similar to assuming that customers shop across different channels (as we do). Their work differs from ours in two ways: (i) They allow the firm to offer a menu of contracts, and in particular, two different prices and refund policies to the two customer types, which is not common in omnichannel settings; and (ii) They do not model cross-channel returns, or the inconvenience of doing so, since they consider a single-channel setting.

Several papers add other dimensions such as customer behavior and competition to this line of work. Shulman et al. (2009) study the optimal policy in a two-product setting, thus adding the exchange option to the return vs. keep decision. They find that the optimal refund can be more or less than the salvage value. They also identify conditions under which it may be optimal for a seller to provide more information about product fit to prevent misfit. Shulman et al. (2011) study the equilibrium strategies when firms compete; they find that restocking fees may be higher in a duopoly than that in a monopolistic setting. Altug and Aydinliyim (2016) consider strategic customers who incorporate the product's availability in the clearance period into their decision to purchase at the full price. The authors find that the optimal refund must be *less* than the salvage value, contrary to Su (2009). In contrast to both these papers, we find that an *omnichannel* firm may optimally choose a refund that is *larger* than the salvage value.

The issue of opportunistic behavior – wherein consumers purchase with no intention of keeping the product – has also been studied: Chu et al. (1998) show that partial refunds are preferable in the presence of opportunistic customers, and Hess et al. (1996) show that one way to limit them is to impose a non-refundable fee.

Overall, this stream of research provides valuable insights that help understand the role of consumer return policies, but it does not shed light on the challenges (and potential advantages) of a typical *omnichannel* firm. In particular, it ignores cross-channel purchasing and return behavior and the impact of price being linked between the two channels.

The second stream of related literature considers the unique problems associated with operating more than one channel, especially with interactions between the two; see Brynjolfsson et al. (2013) and Bell et al. (2014) for a broad survey. Huang and Swaminathan (2009) study the optimal pricing strategies of a firm selling through two channels. The dual channel nature also creates complex information flows: For example, Lal and Sarvary (1999) show that the online channel affects consumers' search behavior and may result in higher prices.

Recent papers empirically evaluate different omnichannel fulfillment strategies: Gallino and Moreno (2014) consider the impact of buy-online, pickup-in-store (BOPS) on a firm's sales, and Gallino et al. (2016) consider the effect of introducing the ship-to-store functionality, wherein customers have an out of stock product shipped to their local store for free. Bell et al. (2017) investigate the impact of opening physical showrooms on consumers' channel choices. Others approach these problems from a theoretical perspective: Gao and Su (2016a) study the implications of BOPS on customers' purchasing decisions, whereas Gao and Su (2016b) study the impact of three different approaches to sharing information – physical showrooms as in Bell et al. (2017), virtual showrooms, and providing information on in-store product availability. In related work Karp et al. (2017) study the order fulfillment strategies for online orders considering that inventory counts may be inaccurate, due to depletion by an in-store sale before the item is picked for shipment. Harsha et al. (2016) develop an approach to optimize the cross-channel fulfillment inventory at an omnichannel firm. And, Ofek et al. (2011) examine how physical store assistance levels change as a result of adding an online channel. While some of these papers (Bell et al. 2017, Gao and Su 2016b, Ofek et al. 2011) acknowledge that customers may return misfit items, they do not model the return policy decision of the firm.

In fact, there has been very limited work on customer return policies in an omnichannel setting. An empirical study by Kumar et al. (2014a) investigates the impact of opening a new store on revenue for an omnichannel firm that offers a full refund policy. The authors find that the revenue increased as a result of the store opening, and they attribute this, partially, to customers buying more online owing to the reduced cost of returning in-store. Our model is able to capture this effect of opening a new store at a high level, and our results corroborate their finding when the firm offers full refunds. In addition, we find that the impact of improving the convenience of returns crucially depends on the return policy in place, suggesting that the conclusions may be different for an omnichannel firm that offers partial refunds.

The closest paper to ours is Chen and Chen (2017), who investigate when a firm should offer full refunds versus no returns, as well as when to operate in both channels. But they (and Ofek et al. 2011) model the rate of product returns using an exogeneous parameter (i.e., a fraction of sales); in contrast, we allow the firm to influence the return decision of its customers by controlling the price and refund amount (as in the single-channel model of Su 2009). Our approach of endogenizing the return decision is more realistic as several surveys (summarized in §3.1) find that customers do make their choices for returns by considering the firm's policy. In addition, Chen and Chen (2017) allow the price to be set independently in both channels, and they do not consider cross-channel effects, as we do. Finally, they only consider two extreme policies (full refunds and no returns), whereas we derive the *optimal* return policy.

The primary contribution of our work is in marrying the above two streams of literature to fully investigate the return policy decisions of an omnichannel firm. We incorporate two common practices – identical prices across channels and cross-channel returns – that have largely been ignored in prior work, and we are the first to characterize the optimal return policy in an omnichannel context. In addition to being consistent with practice in most omnichannel settings, the identical price assumption highlights an important tension experienced by a firm due to its omnichannel nature i.e., the inability to charge a higher price to customers who visit their stores. By modeling the cross-channel return option, we are able to explain why some firms (e.g., J.C. Penney) experience more online purchases returned in-store (and why they advertise this option). Finally, through the optimal policy, we are able to explain which commonly observed policies are advisable, and in particular, when an omnichannel firm prefers one return policy over another (e.g., when to offer full refunds over in-store returns). The novel aspects of our model enable us to generate valuable insights on the decisions of an omnichannel firm, but they also present analytical challenges; for example, the solution to the optimal return policy involves optimizing over a two-dimensional space wherein some optimal policies do not have a simple characterization and are not amenable to a straightforward comparison.

3.3 Model

We consider an omnichannel firm that sells a product through two channels – online and physical store. The firm can decide whether to allow mail-in product returns (referred to as online returns), and if so, the refund amount r offered to the customer for any items purchased online that are shipped back. In addition, the firm may allow online purchases to be physically returned to a store in exchange for a full refund as is commonly observed in practice – "free returns to store" (referred to as in-store returns).

The firm faces a unit cost c of procurement, a salvage value s_e for products returned online, and a salvage value s_p for products returned to the physical store. The values of c, s_e and s_p represent how efficient the firm's supply chain is and how effectively it can resell through its (own or external)

Table 3.1: Table of notation.

Symbol	Definition
c	Firm's marginal cost of production
s_e	Firm's salvage value for an item returned online
s_p	Firm's salvage value for an item returned to its physical store
h	Online-savvy customer's hassle cost of returning to the physical store
θ	Fraction of store-savvy customers
V	Random variable denoting a customer's valuation for the product
p	Price charged for the product
r	Refund offered for online returns

secondary channels. We assume that s_e and s_p are smaller than or equal to c.[3] We also assume that the firm has enough production capacity and inventory to meet its demand. This allows us to isolate insights on the pricing and return policy decisions, which are the main focus of this work. Table 3.1 summarizes our notation.

The firm also decides the price p to charge for the product. Keeping with the practices of most omnichannel firms, the price is assumed to be identical across channels. Cavallo (2017) finds that 70% of retailers have identical pricing across channels – even more so (92%) for clothing products – and that differences, if any, are negligible. In fact, customers expect this: 82% expect a retailer's prices to be the same in-store and online (Accenture 2015), and increasingly so (up from 69% in the previous year) as it becomes easier to check prices online and to express dissatisfaction regarding price differences through social media and the retailer's website.

Finally, omnichannel firms have a key advantage over online-only competitors when customers utilize their store-return policy – the ability to sell items in-store to them (UPS 2016b, WSJ 2016). To capture this effect we assume that the firm may earn an additional (cross-selling) profit from each customer who returns an item to their store.[4] Given that such cross-selling is more effective when returning in-store compared to returning online – UPS (2016b) finds that 25% more customers made an additional purchase while returning in-store – we assume that the additional profit is realized only when the "free returns to store" option is exercised. This implies that $s_e \leq s_p$.

Customers face uncertainty in their valuation V for the product, which is drawn from a uniform distribution: $V \sim U[0, 1]$. They differ in how they resolve this uncertainty; either by purchasing online, or by inspecting the item in a physical store. We model this heterogeneity by considering

[3] This assumption, which is standard in the literature reviewed in §3.2, ensures that the firm does not receive a net profit from returned items.

[4] For example, J.C. Penney claims that about one-third of shoppers spend an additional \$60 when they visit a store to make a return (WSJ 2016).

two customer types:

- The "online-savvy" type who prefers to purchase the item online, and can return a misfit item – either online or to the store – deciding based on the return policy and the "hassle cost" h she experiences when utilizing in-store returns. Returning to the store can be inconvenient to a customer because of reasons such as traveling to the store, waiting in checkout lines, or needing to show a receipt.[5]
- The "store-savvy" type that visits the store to inspect the item before making a purchase. For simplicity, we assume that they do not make a return. This type also accommodates customers who indulge in showrooming which constitutes 17% of online shoppers (KPMG 2016).

We assume that a customer is of the latter (store-savvy) type with some exogenous probability θ. The consumer model in Su (2009) is a special case of ours when only the online-savvy customer segment is present, i.e., $\theta = 0$. One can interpret θ as capturing customers' inherent preference for one mode of shopping (over the other) with a particular omnichannel firm: Since different firms enjoy different levels of online traffic, the value of θ may be firm-specific. For example, Macy's (2015) finds that 53% of its customers prefer to shop at their brick-and-mortar stores. A recent consumer survey by Market Force (2017) found that 64% of respondents visited Nordstrom's website in the past 90 days as opposed to 39% for H&M; therefore, concluding that Nordstrom's value of θ is relatively low compared to H&M is valid. Alternatively, one can think of a firm serving two types of customers who differ in their costs of visiting a physical store before purchase: For a proportion θ of the customers this cost is negligible (thus, normalized to zero), whereas for the remaining customers this cost is prohibitively high.

Online customers are affected by two primary competing aspects of returns: (1) paying for returns (50% of customers claim this to be an issue according to UPS 2016a); and (2) the inconvenience of returning to a store (21% of customers, UPS 2016a). Thus, we consider online returns to be (relatively) hassle-free since most retailers provide a return label (and a package); however, they may levy return fees. We model any such fees by allowing the refund amount r to be less than the price p. For analytical convenience, our model ignores the hassle cost of online purchases (including any time spent for searching, payment etc.) as well as any hassle costs of store-savvy customers in visiting the store; these costs would result in some customers not purchasing the product but would not affect any of our main results. We also assume that $c \leq EV = \frac{1}{2}$ as in Su (2009).

Given the firm's price and return policy choices, online-savvy and store-savvy customers will make their purchase (and return, if necessary) decisions as follows.

Online-savvy type: An online-savvy customer makes two sequential decisions: (i) buy the product or not, and then, if she buys, (ii) keep the product or return (after observing her valuation for the item). This combined with the two options for returning the product – return online for

[5] The optimal hassle cost to induce has been the subject of research as well (Davis et al. 1998, Hsiao and Chen 2014), but we consider this cost to be exogenous.

refund r or in-store for a full refund – results in three alternatives.

1. Buy online, and (possibly) return online (BORO): In this case, she will return the product in stage (ii) if her valuation V is less than the refund r. Therefore, if she buys the product in stage (i), she will receive expected utility $u_{\text{BORO}} = E\max\{V, r\} - p$.
2. Buy online, and (possibly) return in-store (BORS): In this case, she will return the product in stage (ii) if her valuation V is less than the refund p for returning the product less the hassle cost h of doing so. Therefore, if she buys the product in stage (i), she will receive expected utility $u_{\text{BORS}} = E\max\{V - p, -h\}$.
3. Do nothing: In this case, she will receive a utility of zero (outside option).

The online-savvy customer chooses the alternative that maximizes her utility. We can show that BORO is dominated by BORS if and only if $p - r > h$, and the customer purchases the product if and only if $u_{\text{BORO}} \geq 0$ ($u_{\text{BORS}} \geq 0$) under BORO (BORS). The firm's expected profit $\pi_{OS}(p, r)$ from online-savvy customers depends on their return behavior:

$$\pi_{OS}(p, r) = \begin{cases} p - c - (r - s_e)r, & \text{if } p - r \leq h, u_{\text{BORO}} \geq 0 \\ p - c - (p - s_p)(p - h), & \text{if } p - r > h, u_{\text{BORS}} \geq 0 \end{cases} \tag{3.1}$$

where $(p-c)$ represents the margin the firm makes if the item is kept by the customer, r and $(p-h)$ represent the probability that a customer makes a return under BORO and BORS respectively, and r and p represent the refund for a returned item under BORO and BORS respectively.

Store-savvy type: A store-savvy customer may choose to visit the store, and, if so, she resolves her valuation uncertainty by inspecting the product in-store. She then buys the product if her valuation V is larger than p; thus her expected utility from visiting the store is $u_{\text{Store}} = E\max\{V - p, 0\}$. She receives a utility of zero if she does not visit the store. Since u_{Store} is non-negative, the store-savvy customers *always* choose to visit the store regardless of the price.[6] (The store-savvy customer may visit the retailer's website to check the product online, but she will still have to visit the store in order to know her V.) The firm's expected profit $\pi_{SS}(p)$ from store-savvy customers is

$$\pi_{SS}(p) = (p - c)(1 - p), \tag{3.2}$$

where $(1 - p)$ represents the probability that a store-savvy customer realizes a valuation $V > p$, and therefore purchases the product.

Combining the profit from the two types of customers, we can express the firm's total expected profit as $\theta\pi_{SS}(p) + (1 - \theta)\pi_{OS}(p, r)$. We are interested in the setting where the firm chooses to sell to both segments. We focus on this setting for two reasons: 1) Most omnichannel firms have non-zero sales from both types of customers – otherwise, the cost of operations may not justify its being "omnichannel" in the first place; and 2) The constraints in our model (on u_{BORO} and u_{BORS}) only exist due to online-savvy customers; ignoring this segment removes the question of

[6] Recall that her hassle cost is zero.

returns, which is the key decision we seek to study in this work. We do analyze the setting where the firm serves only the store-savvy customers, as an extension in §3.6.1. We also consider other extensions, such as incorporating customer heterogeneity in the hassle cost for in-store returns in §3.6.2, evaluating the potential benefits of providing fit information to customers in §3.6.3, and allowing product returns from store-savvy customers in §3.6.4.

3.4 Analysis of special return policies

We begin our analysis with two special return policies commonly studied in the literature: 1) the full-refund policy; and 2) the no-returns policy. In the full-refund case, the firm offers a "money-back guarantee" to ensure customer satisfaction. In our model, this implies that the refund must be equal to the price; i.e., $r = p$. Such generous full-refund policies are offered by many retailers, including Nordstrom and Macy's. In the no-returns case, the firm offers no returns whatsoever or, in other words, "all sales are final." In our model, this implies that the refund is equal to zero (i.e., $r = 0$) and the BORS option does not exist. Several retailers (e.g., Gap and J.Crew) offer such an extreme policy on their "final sale" or "clearance" items, but a few retailers (e.g., American Apparel) also offer such a policy on regular items. We analyze these two special policies and examine when the seller prefers one to the other.

3.4.1 Full-refund policy

Under the full-refund policy, the online-savvy customer always chooses the BORO option to receive utility $u_{\text{BORO}} = E\max\{V - p, 0\} \geq 0$. Thus, regardless of the price p, all online-savvy customers will purchase the product; however, those with realized valuation $V < p$ will return the product (online) and receive a full refund. In this case the firm sets price p in order to maximize its expected profit,

$$\Pi_F(p) = \theta(p - c)(1 - p) + (1 - \theta)(p - c - (p - s_e)p), \tag{3.3}$$

which is obtained from (3.1) and (3.2) by fixing $r = p$. The following lemma characterizes the optimal full-refund policy that maximizes (3.3). All proofs appear in Appendix B.1.

Lemma 3.1. *The optimal price under full-refund is* $p_F = \frac{1}{2}\left(1 + s_e + (c - s_e)\theta\right)$ *and the optimal profit is* $\pi_F = \frac{1}{4}\left(1 + s_e + (c - s_e)\theta\right)^2 - c$.

We can conclude from Lemma 3.1 that the price and profit under the full refund policy are non-decreasing in both salvage value (s_e) and the proportion of store-savvy customers (θ). Under full refunds, a proportion $(1 - p)$ of customers, including both online-savvy and store-savvy customers, pay p and keep the product, and a proportion $(1 - \theta)p$ of the customers (i.e., a proportion p of online-savvy customers) return the product, which then has to be salvaged at (possibly) a net loss. Therefore, at higher θ, there are fewer customers returning the product, and at higher s_e, when they do return the product, the firm is able to salvage the product at a reasonable value. Even though increasing prices makes (online) returns more likely, increasing θ and s_e blunts the cost of these returns, so that together they enable the firm to benefit from increasing prices.

3.4.2 No-returns policy

Under the no-returns policy, an online-savvy customer who purchases online cannot return the product if she is unsatisfied with it. These customers are willing to pay a maximum price of $EV = \frac{1}{2}$, their expected valuation for the product, and will keep it regardless of their realized valuation. In this case the firm sets price p in order to maximize its expected profit,

$$\Pi_N(p) = \theta(p - c)(1 - p) + (1 - \theta)(p - c), \tag{3.4}$$

obtained from (3.1) and (3.2) by setting $r = 0$ (and ignoring BORS), subject to the constraint $p \leq \frac{1}{2}$. The following lemma characterizes the optimal no-returns policy.

Lemma 3.2. *The optimal price under no-returns is $p_N = \frac{1}{2}$ and the optimal profit is $\pi_N = \frac{1}{4}(1 - 2c)(2 - \theta)$.*

According to Lemma 3.2, the firm fully extracts the surplus from online-savvy customers. As more customers are online-savvy, its profits increase – i.e., π_N is decreasing in θ – when $c < \frac{1}{2}$. (Profits are zero when $c = \frac{1}{2}$.) This is because the firm wishes to charge a higher price in order to extract more profit out of its customers, but is unable to do so since the price is constrained by the online-savvy customers' low willingness to pay. (We explore the case when the firm may exit the online-savvy market in order to raise the price i.e., serve only the store-savvy customers, in §3.6.1.)

3.4.3 Preference between full-refund and no-returns policies.

A key question posed in the prior literature is: When is it preferable for a firm to offer full refund policies relative to no returns? The following proposition compares the performance of these two policies.

Proposition 3.1. *The firm prefers no returns when both the salvage value s_e and proportion θ of store-savvy customers are sufficiently low. Moreover, the firm offers full refunds when θ or s_e are sufficiently high.*

This proposition provides sufficient conditions for the firm's preference and cautions us that no returns – as recommended in the related literature – may be worse than offering full refunds. In particular, an omnichannel firm prefers to offer no returns when both the salvage value s_e *and* the proportion θ of store-savvy customers are low; in contrast, if θ were high enough, the firm would offer full refunds even at low salvage values.

The full refund policy affects the firm's profit in two ways: First, beneficially, it allows the firm to charge a higher price since an online-savvy customer can return the product for free, and thus is willing to pay more for it upfront before she realizes her valuation. Second, detrimentally, generous refunds result in returned items that need to be salvaged at a net loss (if $s_e < c$). The relative magnitude of these two effects drives our result: When θ is low, the firm suffers considerably from returned products at low s_e, since a significant segment of its customers choose to shop online, so that the second (adverse) effect is more dominant and the no-returns policy is preferred. On the other hand, when θ is high, the firm is able to extract higher profits from the store-savvy customers (who now form the majority of its customer base) by promising full refunds to its fewer online-savvy customers to support raising prices.

When only online-savvy customers are present (i.e., $\theta = 0$)[7] our result that full refunds are preferred at high salvage values coincides with Su (2009). Note though that the salvage value above which full refunds are preferred to no returns is quite high: more than 83% of the cost c.[8] Thus, the argument that full refunds are offered at high salvage values explains why firms like Nordstrom or Neiman Marcus offer full refunds: Both enjoy higher salvage value s_e owing to their own outlet stores – Nordstrom Rack and Last Call respectively – which allow them to sell returned items at mild discounts (Altug and Aydinliyim 2016). Furthermore, as they operate both channels (i.e., $\theta \geq 0$), we are thus able to extend results from prior work to the omnichannel setting.

In practice, not all firms experience high salvage values: It is estimated that a third-party outlet store buys returned goods for only 10-20% of their value (Stock et al. 2006). Previous literature – specifically Su (2009) – cannot explain why firms such as Macy's and Gap, that predominantly use off-price retailers (e.g., T.J. Maxx), also offer full refunds for regular product categories. In fact, using the logic of Su (2009) would lead to the conclusion that no returns will be preferred by these firms because their salvage values tend to be low. Now we can provide an explanation for this: The proportion of store-savvy customers (θ) at these firms is high. Therefore, even though they face lower salvage values, they choose to offer full refunds in order to benefit from higher margins, while maintaining a presence in the online market.

We have demonstrated that how "omnichannel" the firm is (captured by θ) plays a key role in comparing full refunds against no returns. Specifically we see that policy conclusions based on prior research that considers only a single (online-savvy) segment – that full refunds are only preferred at high enough salvage values – do not in general hold in the omnichannel setting. In order to identify if more general return policies we see in practice – such as full refunds, partial refunds, and in-store returns – arise naturally, we next examine the *optimal* return policy.

[7] $\theta = 0$ implies only one customer type; we consider this case in order to compare our model with prior work.

[8] This is because full refunds are optimal if $\sqrt{2} - 1 < s_e, \forall\theta$: This is a sufficient condition when $\theta > 0$, but it is necessary and sufficient when $\theta = 0$. Since $s_e \leq c \leq 1/2$, $s_e/c > 2(\sqrt{2} - 1)$, or the salvage value must be at least 0.83 times the cost, when $\theta = 0$.

3.5 Optimal return policy

Under the optimal policy, the firm can set both the refund amount r and the price p. A positive $p - r$ can be interpreted in one of two ways: (i) as the fee levied on online returns (e.g., \$5.99 fee charged by H&M and Victoria's Secret); or (ii) as a restocking fee (e.g., 15% fee charged by Macy's for bulky items such as furniture). We assume that the firm allows free in-store returns as before.

Through its policy choice, the firm can induce one of two outcomes depending on online-savvy customers' return behavior – either the BORO option or the BORS option – choosing whichever will yield a higher profit. We first analyze the two induced outcomes – BORO and BORS – separately in §3.5.1 and §3.5.2, respectively. We then identify the better of the two – the optimal policy – in §3.5.3.

3.5.1 BORO outcome

First, we need $u_{\text{BORO}} \geq 0$ so that online-savvy customers purchase the product. Equivalently, from the definition of $u_{\text{BORO}} = E\max\{V, r\} - p$, the highest price that online-savvy customers are willing to pay is $\frac{1}{2}(1 + r^2)$ (Lemma B.5(i) in Appendix B.2). Those customers with $V < r$ will return the product (online). Second, the optimal policy must result in the BORO outcome. From (3.1) this means that $p \leq r + h$. Thus, there are two constraints that must be satisfied: 1) $p \leq \frac{1}{2}(1 + r^2)$ capturing the maximum willingness to pay; and 2) $p \leq r + h$ inducing the BORO outcome. Subject to these constraints, the firm sets p and r in order to maximize its expected profit,

$$\Pi_O(p, r) = \theta(p - c)(1 - p) + (1 - \theta)(p - c - (r - s_e)r), \tag{3.5}$$

which is obtained from (3.1) by considering the BORO option, and (3.2).

The following proposition characterizes the BORO policy.

Proposition 3.2. *The optimal refund (r_O) and price (p_O) under the BORO outcome are*

$$r_O = \begin{cases} r_\beta, & \text{if } 0 \leq h \leq h_\beta \\ r_\gamma, & \text{if } h_\beta < h < h_\alpha \\ r_\alpha, & \text{if } h_\alpha \leq h \end{cases} \qquad p_O = \begin{cases} r_\beta + h, & \text{if } 0 \leq h \leq h_\beta \\ r_\gamma + h, & \text{if } h_\beta < h < h_\alpha \\ \frac{1}{2}(1 + r_\alpha^2), & \text{if } h_\alpha \leq h \end{cases} \tag{3.6}$$

where $h_\beta = \frac{1}{2\theta^2}\left(1 - \sqrt{(c - s_e)\theta^2 + (s_e - 1)\theta + 1}\right)^2, r_\beta = \frac{1}{2}\left(1 + c\theta - 2h\theta + s_e(1 - \theta)\right), r_\gamma = 1 - \sqrt{2h}$, $h_\alpha = \frac{1}{2}(1 - r_\alpha)^2$, *and* $r_\alpha \in [0, 1)$ *is defined as the largest non-negative real root of the cubic equation:* $-\theta r^3 + r(c\theta + \theta - 1) + s_e(1 - \theta) = 0$.

The optimal policy under the BORO outcome satisfies:

(i) The return fee ($p_O - r_O$) is non-decreasing in h.

(ii) The refund r_O is non-increasing in h.

(iii) The refund r_O is at least as large as the salvage value s_e.

(iv) r_α is increasing in θ when $s_e > 0$.

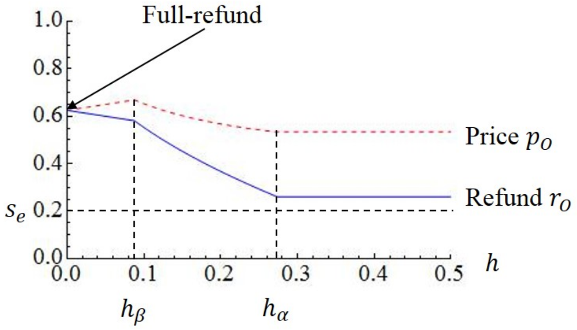
The firm considers two critical forces while deciding the optimal policy – the highest price that customers are willing to pay i.e., $\frac{1}{2}(1+r^2)$, and online-savvy customers' incentive to patronize online returns i.e., the magnitude of the return "fee" $(p-r)$ relative to the return hassle h. Not surprisingly, as store returns become more convenient, the return fee often becomes smaller (Proposition 3.2(i)): The low fee serves to deter online-savvy customers from returning items to the store – i.e., ensuring BORO is optimal – since they tend to compare the fee imposed by the firm against the hassle of returning to the store for a full refund. But, Proposition 3.2(ii) reveals that even the *absolute* refund amount may become more liberal: As h decreases, the incentive constraint $p \leq r+h$ becomes tighter, thereby constraining the firm's pricing power. Responding by decreasing r further would constrain the firm's price – both through the incentive constraint and by reducing online-savvy customers' willingness to pay. So the firm never decreases the (online) refund amount r as in-store returns become more convenient.

Thus the optimal decision depends on the hassle cost, which may be high $(h \geq h_\alpha)$, medium $(h_\beta < h < h_\alpha)$, or low $(h \leq h_\beta)$ as indicated in Proposition 3.2. First, at high h, customers' incentive to consider store-returns is weak, so that only the willingness to pay constraint is binding. Even so, the optimal refund can be *larger* than the salvage value s_e (Proposition 3.2(iii)) – the optimal refund prescribed in Su (2009). The firm benefits from offering higher prices to store-savvy customers[9], and to do so, it must promise large refunds to online-savvy customers. As the proportion θ increases, the firm offers *higher* refunds (Proposition 3.2(iv)). Thus, taking into consideration the omnichannel nature of a firm offers an explanation for refunds that are more liberal than previously concluded.

At low and medium h the firm chooses to steadily increase refunds (when $\theta \neq 0$) as in-store returns become more convenient, indicating that the cross-channel return option offers another explanation for refunds that are very generous. At medium h if the firm decreased prices it would be leaving some surplus to online-savvy customers; by increasing the refund offered, it benefits from charging higher prices (as online-savvy customers' willingness to pay is increasing in the refund). It is easy to see that the optimal price p_O is decreasing in h when $h_\beta < h < h_\alpha$. However, as h grows very small (i.e., low h) this strategy dictates very high prices and refunds, which are undesirable due to the resulting low "real" sales (i.e., unreturned purchases from online-savvy customers and purchases from store-savvy customers) and high volume of returns. Thus, the firm sets its price to be lower than what online-savvy customers are willing to pay. Furthermore, the price is non-decreasing in h with a slope $(= (1-\theta))$ that is decreasing in θ, implying that as customers become more online-savvy, this effect becomes more pronounced: The price decreases more sharply as

[9] We have $r_\alpha = \sqrt{c}$ when $\theta = 1$ (3.6), and $r_\alpha \leq \sqrt{c}, \forall\theta$ (Proposition 3.2(iv)). Thus, $p_O \leq \frac{1}{2}(1+c)$ when $h \geq h_\alpha$. Since $\pi_{SS}(p)$ is uniquely maximized at $\frac{1}{2}(1+c)$, the firm's profit from store-savvy customers $\pi_{SS}(p_O)$ increases with p_O.

Figure 3.1: Firm's optimal price p_O and refund r_O under the BORO outcome.



Note: The following parameter values are used: $c = 0.3, s_e = 0.2$, and $\theta = 0.5$. These values are motivated by a typical omnichannel firm; see Appendix B.3.

in-store returns become more convenient at lower values of θ.

Figure 3.1 shows the optimal policy as a function of h. The values of h defining the three regions are indicated on the h-axis. The refund set equal to salvage value s_e – the optimal policy in prior work (Su 2009) considering only online-savvy customers – is indicated on the vertical axis in order to numerically compare our results against theirs.

Proposition 3.2 also reveals that the optimal policy prescribes full refunds whenever $h = 0$ (indicated in Figure 3.1). In this case online-savvy customers will return items to the store unless they are offered a full refund online, thereby forcing the firm to price as per the full refund policy as in Lemma 3.1. By modeling the hassle of in-store returns, we are able to demonstrate that full refunds may be *optimal* when the firm induces the BORO option. (Later, in §3.5.3, we will show that this can be optimally chosen – considering both BORO and BORS – by the firm.)

3.5.2 BORS outcome

First, the optimal policy must result in the BORS outcome, i.e., $p > r + h$ from (3.1). Second, we need $u_{\text{BORS}} \geq 0$ so that online-savvy customers purchase the product. The constraint $u_{\text{BORS}} \geq 0$ can be equivalently expressed as $p \leq h + 1 - \sqrt{2h}$ (Lemma B.5(ii) in Appendix B.2). These customers will return the product if their realized valuation $V < p - h$, i.e., the volume of returns from online-savvy customers is $(p - h)$.

Because the firm induces the online-savvy customers to choose BORS, the optimal online refund r is zero, or some (possibly negligible) value, i.e., $r_S \in [0, p_S - h)$, where p_S is the optimal price under this outcome. This corresponds to a scenario wherein a firm charges a significant return fee (i.e., low r_S), but allows (free) returns to its physical store. We will consider $r_S = 0$ and derive the optimal price p_S.[10] There are two constraints that must be satisfied: 1) $p \leq h + 1 - \sqrt{2h}$ capturing

[10] The optimal solution is independent of the exact value of $r_S \in [0, p_S - h)$; we choose $r_S = 0$ for simplicity.

the maximum willingness to pay; and 2) $p > h$ inducing the BORS outcome. Subject to these constraints, the firm sets p to maximize its expected profit,

$$\Pi_S(p) \quad = \quad \theta(p-c)(1-p) + (1-\theta)(p-c-(p-s_p)(p-h)), \tag{3.7}$$

which is obtained from (3.1) considering the BORS option, and (3.2). Note that returning an unsatisfactory product to the store is feasible only when $h < \frac{1}{2}$ because otherwise no price p satisfies the constraint that $h < p \leq h + 1 - \sqrt{2h}$. Thus, we assume that $h < \frac{1}{2}$.

The following proposition characterizes the BORS policy.

Proposition 3.3. *The optimal price (p_S) and profit (π_S) under the BORS outcome are given by*

$$p_S = \begin{cases} \frac{1}{2}(1+h+s_p+(c-s_p-h)\theta), & \textit{if } h < \bar{h} \\ h+1-\sqrt{2h}, & \textit{if } \bar{h} \leq h < \frac{1}{2} \end{cases} \tag{3.8}$$

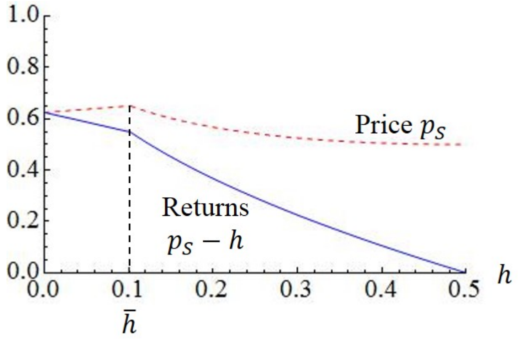
$$\pi_S = \begin{cases} \frac{1}{4}(1+h+s_p+(c-s_p-h)\theta)^2 - c - hs_p(1-\theta), & \textit{if } h < \bar{h} \\ K(s_p), & \textit{if } \bar{h} \leq h < \frac{1}{2} \end{cases} \tag{3.9}$$

where $\bar{h}$ is defined as the value of $h \in \left(0, \frac{1}{2}\right]$ that satisfies $\frac{1}{2}(1+s_p+(c-s_p)\theta) = h\left(\frac{1+\theta}{2}\right)+1-\sqrt{2h}$ and $K(s_p) \equiv \sqrt{2}h^{3/2}(1+\theta) - h^2\theta + \sqrt{2h}(1-s_p(1-\theta)-c\theta) - h(2+\theta(1-c)) - (c-s_p)(1-\theta)$. Moreover, when $\theta \neq 1$, π_S is increasing in s_p $\forall h$, and increasing in h for $h < \bar{h}$.

The firm's optimal policy is determined by two factors – the online-savvy customers' willingness to pay, i.e., $h + 1 - \sqrt{2h}$, and the volume $(p - h)$ of returns as a result of price p (from online-savvy customers). The former is decreasing in h: Some customers, i.e., those with valuation $V \in (p - h, p)$, will keep the item even though they will receive negative utility (ex-post) from doing so, because returning it is not worth the "hassle." As h increases, online-savvy customers increasingly incorporate this effect into their purchasing decision, and thus, they are not willing to pay as much for the product. Correspondingly, the latter is decreasing in h for constant p: The volume of returns decreases as it becomes more inconvenient to return in-store.

Figure 3.2 shows the firm's optimal price p_S and the resulting volume of returns $p_S - h$. Consider an increase in h within the low h region (i.e., $h < \bar{h}$). In this case fewer customers return the product at a given price, so the firm can raise the price (at a rate slower than the growth of h) and still benefit from a lower return volume. Since online-savvy customers willingness to pay is high in this region the firm can afford to do so, and this is evident in the increasing nature of price p_S (Figure 3.2) and profit π_S (Proposition 3.3). At high h (i.e., $h \geq \bar{h}$), this strategy of increasing price with h raises the price above the online-savvy customers' (lower) willingness to pay, so that the firm now extracts maximum consumer surplus by charging $p_S = h + 1 - \sqrt{2h}$, and the price is now decreasing in h.

Figure 3.2: Firm's optimal price p_S and volume of returns $p_S - h$ under the BORS outcome.



Note: The same parameter values are used as in Figure 3.1.

3.5.3 Optimal policy

We focus on the optimal policy when $h < \frac{1}{2}$; otherwise, the BORS outcome is infeasible and the BORO outcome is trivially optimal. The following proposition compares the firm's profit under the two outcomes – BORO and BORS – in order to identify the optimal policy.

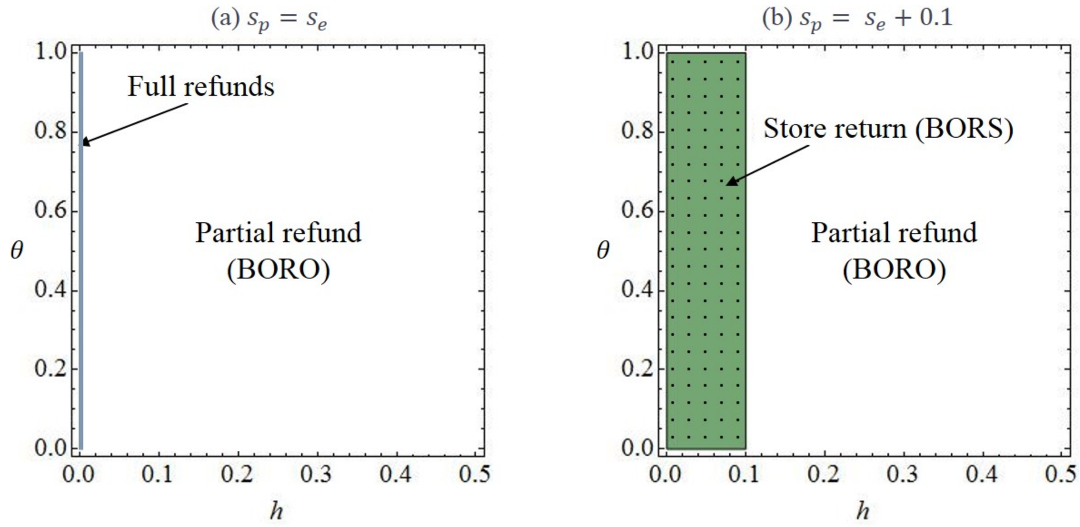
Proposition 3.4. *(i) When $s_p = s_e$, the optimal profit under the BORO outcome, $\pi_O \equiv \Pi_O(p_O, r_O)$, is at least as high as that under the BORS outcome π_S.*

(ii) When $s_p > s_e$, the optimal profit under the BORO outcome π_O is lower than that under the BORS outcome π_S if $\theta \neq 1$ and h is sufficiently small.

Proposition 3.4 highlights the important role of the firm's salvage capability: If the salvage values in-store and online are identical, the optimal policy *always* induces the BORO outcome, and will be given by Proposition 3.2. However, if in-store salvage value is better i.e., $s_p > s_e$ (possibly from cross-selling in-store), Proposition 3.4(ii) presents sufficient conditions for the optimal policy to induce the BORS outcome: There are situations when store returns are *preferred* by an omnichannel firm. In the latter case, the firm deliberately sets low refunds online to induce the BORS outcome, and the optimal policy is given by Proposition 3.3.

Figure 3.3 depicts the firm's preference as stated in Proposition 3.4. We consider identical salvage values across the two channels in Figure 3.3(a) and allow the salvage value to be higher in-store in Figure 3.3(b). When the salvage values are identical, i.e., $s_e = s_p$ (Figure 3.3a), combining Proposition 3.4(i) and Proposition 3.2, full refunds are optimal when $h = 0$, but when $h > 0$, a partial refund is optimal. To see the intuition behind this result, consider low h (very close to zero). When $h = 0$ and $s_e = s_p$, we have $p_O = r_O = p_S = p_F$. For small h around $h = 0$, the refund amount and the proportion of online-savvy customers returning items are both $p - h$ under the BORO outcome (Proposition 3.2), but they are p and $p - h$ respectively, under the BORS outcome (Proposition 3.3). Thus, when $s_e = s_p$, the firm is better off under the former scenario since it gives out a smaller refund.

Figure 3.3: Firm strategies: When to offer full refunds, partial refunds with online returns, or store returns.



Note: The same parameter values are used as in Figure 3.1, except that θ is now variable. Additionally, we use $s_p = s_e + 0.1$; see Appendix B.3 for details.

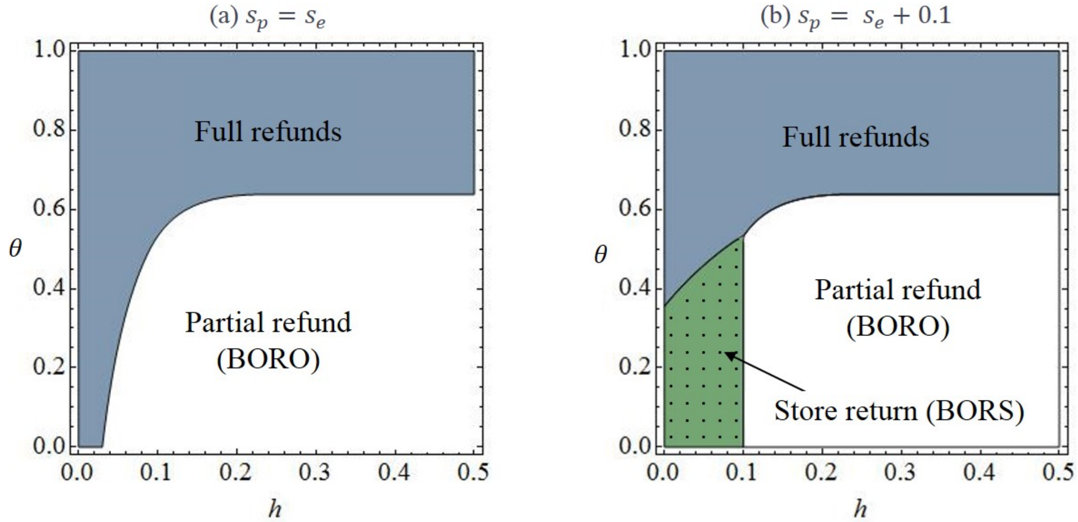
When the salvage value in-store is higher, i.e., $s_p > s_e$, the policy switches from partial refunds to store returns at low h (Figure 3.3b): At low h, customers are able to easily return items to the store. From the firm's perspective, the added benefit $s_p - s_e$ from salvaging in-store negates the higher payout for store returns i.e., p compared to $p - h$ under online returns. Indeed, in many cases e.g., as in Figure 3.3(b), the policy switches exactly when the difference (h) in the payout coincides with the additional benefit ($s_p - s_e$).[11] At high h, many online-savvy customers will keep a misfit item if they are forced to return it to the store, implying that the firm would not enjoy significant benefits from the higher salvage value in-store since return volumes are low. Moreover, the firm would be constrained to charge low prices due to online-savvy customers' low willingness to pay. Thus the firm is better off offering partial online refunds.

Demand stimulating effects of full refunds

It is important to note that our model thus far disregards an additional, favorable effect of full refund policies – the ability to influence customers' long term loyalty with the firm and stimulate future demand. Bower and Maxham III (2012) note that free return policies result in future customer spending that is 158% to 457% more by the end of two years. Moreover, money back guarantees could signal higher quality (Moorthy and Srinivasan 1995), which may result in increased demand

[11] This occurs when the switching value $\tilde{h}$ satisfies $\tilde{h} \geq \bar{h}$ and $h_\beta < \tilde{h} < h_\alpha$ so that the price is the same under the two policies ($= h + 1 - \sqrt{2h}$). In this case setting $\pi_O = \pi_S$ gives the switching value $\tilde{h} = s_p - s_e$.

Figure 3.4: Firm strategies in the presence of demand stimulation.



Note: The same parameter values are used as in Figure 3.3. Additionally, we use $\beta = 1.5$; see Appendix B.3 for details.

as well.

We now incorporate the effect of additional demand stimulated by offering full refunds into our model, through an exogenous multiplicative factor $\beta \geq 1$. Thus, the firm's profit under full refund is $\beta\pi_F$.[12] Figure 3.4 recreates the firm's strategies in Figure 3.3 considering demand stimulation i.e., when $\beta = 1.5$. Not surprisingly, we find that full refunds are preferred more often than before.

The impact of customers' channel preference (θ) on the firm's optimal policy is however more interesting. On the one hand, when the firm enjoys high foot traffic (high θ), it offers full refunds. The intuition for this is similar to that in Proposition 3.1: The firm benefits from promising full refunds to online-savvy customers so that it can charge a sufficiently high price to store-savvy customers. On the other hand, when the firm enjoys high online traffic (low θ), the difference in salvage capabilities plays a key role in determining the return policy. The optimal policy induces in-store returns (BORS) when customers can effortlessly visit the store to return items (low h) if $s_p > s_e$ so that the firm reaps the benefit of the higher in-store salvage value. But, when in-store returns are inconvenient i.e., high h, the firm offers partial online refunds which are generous enough to discourage store returns.

[12] A similar technique is considered by Altug and Aydinliyim (2016) to capture the "demand boost" created by a lenient policy.

Profit increasing effects of improving convenience and accessibility

As the hassle of in-store returns plays an important role in our results, a firm might consider adopting measures to reduce it: It could make it more *convenient* to return a product to the physical store (e.g., either by not requiring a receipt or by allocating a designated area for handling returns) and/or increase the number of locations they operate, thereby improving customers' *accessibility* to their stores. In addition to decreasing h, these measures may also result in increased θ since some online-savvy customers may now choose to visit the store prior to purchase. We classify these measures into two scenarios: (Scenario 1) that results in lower h only; and (Scenario 2) that results in lower h and higher θ. The question is whether such measures necessarily translate into higher profit. We find that the effect depends on the return policy in place.

First consider Scenario 1. While the optimal profit under full refund is constant in h (Lemma 3.1), the optimal profit under partial refund is increasing in h for low h.[13] The driving force here is that, although customers receive full refunds when they return the product to the store, not all misfit items are returned under the store returns policy; a decrease in h causes more customers to return misfit products, hurting the firm's profit. Thus, it may be unprofitable for an omnichannel firm with a partial refund policy to implement measures that only decrease h.

Next, consider Scenario 2. In this case, while the net effect under full refund is higher profit since π_F is increasing in θ when $s_e < c$ (Lemma 3.1), the effect under partial refund is not straightforward since both π_S and π_O can be either increasing or decreasing in θ (Lemma B.6 in Appendix B.2). Therefore, such measures that decrease h and increase θ could *lower* profits under partial refunds, depending on the relative strengths of the two effects (decreasing h and increasing θ) in general.

It is of note that Kumar et al. (2014a) find empirically that an omnichannel retailer's opening of a new store results in increased total revenue. Since their firm's return policy is free returns both online and to the store, our insights – assuming that more customers become store-savvy (Scenario 2) – corroborates this result. In addition, they suggest that the increased ease of in-store returns also drives this effect: While the hassle of in-store returns does not impact the firm's profit under full refunds in our model, this mechanism does arise in our extension where customers are heterogeneous in their hassle costs in §3.6.2. In this case we show that the firm's profits under full refunds do improve when more customers find in-store returns convenient. In contrast, we caution that an omnichannel firm that charges a return fee, or one that experiences returns predominantly in-store, may need to carefully evaluate the impact on its profit, and further empirical research is needed to fully understand this phenomenon in practice.

[13] Under BORS, this occurs when $h < \bar{h}$ and $\theta \neq 1$ (Proposition 3.3). Under BORO, when $h < h_\alpha$ (Proposition 3.2), this occurs because the constraint $p \leq r + h$ becomes tighter at smaller h.

3.6 Extensions

3.6.1 Serving only store-savvy customers

Our analysis has focused on serving both customers types; however, in some cases, it may be profitable to sell only to one segment i.e., the store-savvy type. (This is because the constraints in our model on u_{BORO} and u_{BORS}, and thus, on price p and refund amount r, exist only to ensure serving the online-savvy type.) In this subsection we identify those situations where serving only the store-savvy customers could be more profitable. In this case the profit is $\Pi_{\text{SS}}(p) = \theta(p-c)(1-p)$, so that the unconstrained price and profit are $p_{\text{SS}} = (1+c)/2$ and $\pi_{\text{SS}} = \theta(1-c)^2/4$ respectively.

First, consider the analysis in §3.4. Comparing π_{SS} against the profit π_F under full refunds, we find that the preference depends on the salvage value s_e. If $s_e = c$, then serving only store-savvy customers is *never* preferred (for any $\theta < 1$). When $s_e < c$, this strategy may be preferred at high enough θ, over both full refunds and no returns. Thus, if the online segment is small enough the firm may choose to shut out its pure online shoppers entirely in exchange for more freedom in setting price; in this case, the firm receives no returns. Next, consider the analysis in §3.5. Serving only store-savvy customers is not feasible when $h = 0$; but, it may be optimal when h is high, salvage value s_e is low, and θ is high enough, again for reasons similar to those in §3.4.

Therefore, exiting the online-savvy market may be preferred in the top right region in Figures 3.3 and 3.4. (With sufficient demand stimulation, there may be no difference in Figure 3.4.) Thus, even when we relax our constraint that both segments must be served, our main findings and practical implications for omnichannel firms continue to hold.

3.6.2 Heterogeneous hassle costs

In our model we assume that all online-savvy customers experience the same hassle cost h for returning in-store. However, in reality, customers may face different hassle costs; for instance, due to distance from stores. Suppose a fraction $\delta \in (0,1)$ of online-savvy customers experience negligible hassle cost ($h = 0$), whereas others have a prohibitive cost ($h = 0.5$) so that they will not consider in-store returns at any refund amount or price. In this case the firm sets p and r to maximize its profit $\theta(p-c)(1-p) + (1-\theta)(p-c-\delta(p-s_p)p - (1-\delta)(r-s_e)r)$, subject to $r \leq p \leq (1+r^2)/2$. Here the term $\delta(p-s_p)p$ corresponds to those customers with zero hassle cost who choose to return in-store for a full refund, and the term $(1-\delta)(r-s_e)r$ corresponds to those who have high hassle cost and return online to receive refund r. Thus, we can capture the more realistic scenario wherein customers return both in-store and online.

In this setting we can show that the firm sets its price to extract the maximum surplus i.e., the optimal price p_H and refund r_H satisfy $p_H = (1+r_H^2)/2$, and r_H is the solution to a cubic expression similar to that in Proposition 3.2. Our key findings qualitatively continue to hold in this setting, with the difference that the policies now depend on δ: In particular, we can show that the optimal refund r_H is at least as large as the salvage value s_e and non-decreasing in δ. Thus, the optimal refund continues to be more generous than previously concluded, and as more customers use the free in-store return option, this optimal refund could become even more generous.

As in §3.4, we can derive the optimal policy under the special case of full refunds i.e., when $p = r$. It is easy to show that the optimal profit is non-decreasing in both θ and δ (Lemma B.7 in Appendix B.2), showing that the firm benefits from improving accessibility to its stores when it offers full refunds. This benefit may be either due to more customers visiting the physical store before purchasing (as in §3.5), or due to more customers finding in-store returns convenient.

3.6.3 Complete fit information

Some of the prior work (Bell et al. 2017, Gao and Su 2016b) consider the benefit of providing fit information to customers in an attempt to reduce undesirable returns; for example, firms could invest in virtual showroom technology. Similar to Shulman et al. (2009), we consider the extreme scenario where customers are able to resolve their uncertainty completely using the fit information provided by the firm. (In reality, such perfect information is unlikely, but this abstraction helps us to compare the effects more clearly.) When customers have complete fit information, returns are completely eliminated. Thus, customers are homogeneous and purchase only if their valuation (now revealed exactly) exceeds the price p. The firm sets price $(1+c)/2$ and enjoys profit $(1-c)^2/4$.

We find that the profit under the BORO outcome (with incomplete fit information) may be larger than that under complete information at high enough salvage values and high hassle costs. Thus, even if it were *costless* to provide fit information, it may not be in the firm's best interests to do so. Intuitively, when the firm does not provide this information, it is able to take advantage of customers who choose to keep the item – even though their valuation is lower than the price – by setting the (partial) refund amount optimally. Additionally, at high salvage values, any returns that may result will have a minimal adverse impact on the firm. Our result extends those in Shulman et al. (2009) for a single-channel firm to an omnichannel setting, and cautions omnichannel firms against naively providing information to its customers without accurately evaluating the consequences of doing so.[14]

3.6.4 Returns from store-savvy customers

In our model we assume that store-savvy customers – who know their valuation exactly upon inspecting an item in-store – do not make returns. But, in reality, they may realize that the product is a misfit after they have made the purchase and want to return items. Such returns on in-store purchases occur less often (about 8%, Transport Topics 2017) relative to returns on online

[14] Although, of course, there may be other benefits to providing additional fit information – along with the additional costs – that we do not capture.

purchases (about 33%, WSJ 2013). To model this, we assume that store-savvy customers still realize their valuation upon inspecting in-store but with some probability δ those who purchase an item will return it back to the store for several reasons such as buying impulsively, not spending adequate time inspecting the item's fit, etc. In this case $\pi_{SS}(p) = (1-p)(p-c-(p-s_p)\delta)$. We can show that the optimal refund r_{SS} at high h is the solution to a (similar) cubic expression as in Proposition 3.2, and the optimal refund is once again more generous than previously concluded even when we consider returns from store-savvy customers.

3.7 Conclusion

An omnichannel firm's choice of a return policy, particularly charging fees for returns via mail and allowing free returns to the physical store, has a significant impact on its customers' shopping (and return) behavior. In this chapter we model an omnichannel firm that sells a product through two channels – physical store and online – to customers who face uncertainty around their valuation for this product. We allow customers to be heterogeneous in how they resolve their uncertainty: a customer of the online-savvy type purchases the item online, and after learning her valuation, she may return a misfit item, whereas a customer of the store-savvy type visits the store to inspect the item (and thus, realize her valuation) before making her purchasing decision. The firm optimally decides both the price – which is identical across channels – and the refund amount for returns by mail, while allowing free returns to their physical stores. Thus our model captures several features of omnichannel retailing not considered by previous research.

We identify two drivers for the generous refunds observed in practice when customers return items online: 1) A heterogeneous customer base that uses two different channels to resolve their valuation uncertainty; and 2) The convenience of returning misfit items in-store for a full refund. In particular, we show that full refunds can be optimal, and are likely to be offered by firms with a large store-savvy customer base: This not only enables the firm to enjoy higher margins – by taking advantage of its online-savvy customers' willingness to pay that increases in the refund amount – but also may increase *future* purchases (by signaling higher quality and increasing customer satisfaction). Moreover, we extend prior work suggesting that full refunds are offered by firms that can salvage returned items at mild discounts to the omnichannel setting. Thus, we can explain why firms that own their own outlet stores (e.g., Nordstrom) and those with more store-based customers (e.g. Macy's) both offer full refunds.

We also show that firms with a significant store presence and higher salvage value in-store than online often charge a return fee to nudge their customers to return in-store. For example, both J.C. Penney and Victoria's Secret have a significant store footprint with over 1000 stores, which suggests a lower hassle cost for returning items to their stores, and J.C. Penney claims to have strong cross-selling capabilities in-store. In line with our results, both firms offer partial (online) refunds but free in-store returns; J.C. Penney claims that over 90% of their (online-purchased) returns take place in-store.

Thus, using three key levers – the extent of the omnichannel nature of a firm, salvage capabilities in both channels, and how large its store presence is – we are able to provide an explanation for the different return policies adopted by real omnichannel firms. Our model emphasizes that return policies can be a valuable tool for an omnichannel firm: In addition to serving the classical purpose of avoiding costly returns, the firm could also use return policies to sway customers' choice of channel for returns to a more profitable one, and to extract higher margins from its heterogeneous customer base by exploiting one segment's willingness to pay.

There are several interesting avenues to extend this research further. First, our model abstracts away the leniency of return windows: Some firms such as Macy's allow up to one year for returning items, whereas others such as Forever 21 allow only 30 days. While this directly impacts the salvage value, which we consider, it also impacts customers' return decisions due to both the forgetful nature of customers (e.g., boundedly rational customers as in Su 2009) and any endowment effect they may experience (Janakiraman et al. 2016). Second, it would be interesting to empirically investigate the role of the firm's return policy on the impact of improving accessibility, for example, by opening more physical locations. Finally, competition between omnichannel firms, inventory decisions at the two channels considering returns, and order fulfillment strategies could be valuable additions to our model. Our work is a first attempt at capturing the complex dynamics that motivate the different omnichannel return policy choices we see in practice, and we hope that it presents an analytical foundation to investigate these important questions.

Chapter 4

Modeling Quality of Care in Hospice Operations

4.1 Introduction

Hospice is a philosophy of care that focuses on providing palliative care to terminally ill patients. Eligible patients have a prognosis of 6 months or less to live if their disease were to run its natural course. These patients choose to forgo curative care in lieu of receiving relief from the symptoms of the disease. The goal of the hospice is to improve the quality of life for both the patient and their family: The American Cancer Society defines the purpose of the hospice as providing “humane and compassionate care for people in the last phases of an incurable disease so that they may live as fully and comfortably as possible.”[1]

Hospice care is divided into four levels of which Routine Home Care (RHC) – wherein the patient receives hospice care at their residence – is the most common level of care, accounting for 98% of care days in 2015 (NHPCO 2016). The service is covered under Medicare, which reimburses a hospice on a per patient, per day basis depending on the level of care[2]. More recently, the reimbursement structure for RHC has been modified to reflect the higher level of effort typically required at the start and the end of patients’ length of stay (LoS) (Huskamp et al. 2001): The new structure is two-tiered with a higher payment rate for the first 60 days of care along with an add-on payment for the last 7 days of a hospice patient’s life (NHPCO 2015).

Patients associate a quality level with the service provided by the hospice; however, quality is difficult to quantify in the hospice setting due to inherent differences from other healthcare settings where one may measure the wait time, number of beds available, mortality rate, etc. Indeed, hospice providers note that traditional quality improvement measures may not be a good fit for hospice (Durham et al. 2011). Instead, adherence to the patient’s plan of care, satisfaction ratings, and symptom and pain management are widely considered to be good measures of quality (CMS 2012).

More recently, the Centers for Medicare and Medicaid Services (CMS) has required hospices to report specific quality measures as part of the Hospice Quality Reporting Program (HQRP). Some

[1] https://www.cancer.org/treatment/finding-and-paying-for-treatment/choosing-your-treatment-team/hospice-care.html

[2] The payment rates may differ by geographical regions according to the wage index. Also, patients may have other insurance.

of the measures currently monitored are pain asessment and dyspnea screening[3]. These measures – computed for each hospice provider as the percentage of patient stays where the particular activity was documented – are publicly available (CMS 2017). But there is concern as to whether such programs actually improve quality (Durham et al. 2011). Moreover, HQRP being of the "pay-for-reporting" type, i.e., failure to submit required quality data results in a reduction in the reimbursement rates (CMS 2015), one may question the usefulness of the program as hospices may choose, strategically, not to report quality. We investigate these concerns analytically and study a hospice manager's incentives under the HQRP.

A key modeling feature captured in our work is the impact of quality on the patient census – the number of patients cared for by a hospice – and thus on the hospice's revenue. On the one hand, quality impacts the rate at which patients leave the hospice; for example, Kelley et al. (2013) consider thirty-day re-admission rates and in-hospital death rates as measures of quality of care. On the other hand, the goal of the publicly available quality measures is to "help consumers compare hospice providers on their performance and assist consumers in making decisions that are right for them"[4] i.e., these measures are meant to directly impact whether new patients choose a given hospice or not (National Association for Home Care & Hospice 2015).

The main contribution of this chapter is to model the quality choices of a profit-maximizing hospice and analyze the impact of various regulatory changes on this choice. We specifically ask the following questions:

- How should a profit-maximizing hospice choose the quality of care it provides? Are the quality choices of a for-profit hospice significantly different from that of a nonprofit hospice?
- Does reporting necessarily improve quality of care? What types of hospices may choose to receive the penalty rather than enroll in such reporting programs?
- How does the sensitivity of patient census to quality affect the hospice's choice?

To answer these questions we model a hospice's choice of quality of care by incorporating two key features: 1) a mechanism for incorporating the effect of quality of care on patient census; and 2) a controllable proxy for quality of care. Specifically, quality affects the rate at which new patients join (National Association for Home Care & Hospice 2015), and low quality of care can elevate departures, due to increased hospital re-admission rates and in-hospital deaths (Kelley et al. 2013). And, crucially, we assume that the hospice can control the level of quality it chooses to provide. A high quality of care – to ensure that patients' symptoms are effectively managed through regular monitoring and hospital re-admissions are low – can be achieved through staffing levels, overtime working hours, or active monitoring of patient health, all of which result in higher costs. Thus, the hospice faces a tradeoff between revenue and quality of care costs.

The rest of the chapter is organized as follows. In §4.2, we review the relevant literature. We

[3] https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Hospice-Quality-Reporting/Current-Measures.html

[4] https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Hospice-Quality-Reporting/Hospice-Quality-Public-Reporting-Background-and-Announcements.html

describe our model for quality of care decisions in §4.3, and conduct numerical experiments using parameter values that represent a typical hospice setting in §4.4. Finally, in §4.5, we conclude and discuss future research directions.

4.2 Literature review

There is an extensive literature studying quality of care within healthcare. Lu and Rui (2017) explore online ratings of cardiac surgeons, and they find that those with a low rating may have significantly higher in-hospital mortality rates, a measure of their performance. Using a national data set, Lu and Lu (2016) show that quality at nursing homes can be negatively affected by laws that restrict the use of mandatory overtime work hours for nurses. Previously, Castle and Engberg (2007) found that the quality of care at nursing homes could be improved by addressing staffing issues such as turnover and levels of temporary nurses.

There are several papers on issues pertaining specifically to hospice operations. Ata et al. (2013) investigate a hospice's incentives for patient management under the Medicare reimbursement structure; they find that the hospice may seek out patients with shorter expected LoS. Cherlin et al. (2010) find that there is significant variation in the staffing mix of nurses depending on a hospice's ownership type – for-profit or nonprofit. Related work on quality of care at hospices mainly focuses on investigating which quality measures may be appropriate in this setting (Carlson et al. 2011, CMS 2012). Notably, even though it has been established that low quality of care can elevate hospital re-admission rates in a hospice setting (Kelley et al. 2013), and therefore effect the complex dynamics of patient census, the fundamental question of *how* a hospice should choose quality of care remains an open question. Our work seeks to answer this question by modeling the relationship between quality and patient census, while taking into account the recent regulatory changes that mandate reporting of quality measures.

We formulate the problem of setting quality each week considering the uncertainty in patient census dynamics as a Markov decision process (MDP). Thus, there is relevant work in the literature on optimal control problems. Lee and Kulkarni (2014) investigate a multi-server queueing system wherein both the arrival and service rate can be controlled subject to convex holding and service costs, while earning revenue that is concave in the arrival rate; they find that the optimal arrival and service rates are monotonic in the number of customers. Similarly, Chan and Yom-Tov (2014) study an admission control and speedup problem within the healthcare setting; they assume concave service costs and show that the optimal control policy only chooses the maximum or the minimum of the available actions. In contrast, we consider very general transition probabilities owing to the discrete-time nature (i.e., weekly quality update) of our problem. Furthermore, we impose a different structure on the cost and revenue terms motivated by the application to hospices: Our model is characterized by a quality of care cost that is convex in the chosen quality and linear in the patient census, and revenue that is linear in the patient census.

4.3 Markov decision process formulation

We use a discrete-time infinite-horizon MDP with stationary rewards and general transition probabilities to formulate the quality control problem. This model can be interpreted as a guide for decisions of the hospice; for example, how to adjust staffing levels after realizing the patient census for the week.

We consider a single class of patients, all of whom fall under RHC as this most common level of care (NHPCO 2016). Let $S \in \{0, 1, \dots\}$ be the state space of patient census, and let $\mathcal{Q} = [0, \bar{q}]$ be the set of permissible values for quality q (action)[5]. Our time increment is one week, since staffing choices (e.g., number of per diems to hire, home visits etc.) are typically made for the upcoming week. The hospice earns revenue $r > 0$ per patient, per week and incurs cost $c(q)$ (increasing and convex in q with $c(0) = 0$) per patient, per week for providing quality q.

If we provide quality q in state i, the next state j is given by the random variable $X_i(q) \sim P_{ij}(q)$. Furthermore, let $p(q)$ denote the probability with which any individual patient may depart, and let $\lambda(q)$ denote the arrival rate of new patients. We assume that $\lambda(q)$ is finitely valued and non-decreasing and concave in q, and $p(q)$ is decreasing and convex in q. The non-decreasing nature of the arrival rate in q captures the effect of quality on whether new patients choose a given hospice or not (National Association for Home Care & Hospice 2015), whereas the decreasing nature of the departure probability in q serves as a proxy for either the adverse affects of poor quality such as increased thirty-day re-admission rates and in-hospital death rates (Kelley et al. 2013), or dissatisfied patients choosing alternate hospice providers[6].

Let $v(i)$ denote the maximum expected total β-discounted profits over the infinite horizon when starting in state i. The Bellman optimality equation is given by

$$v(i) = \max_{q \in \mathcal{Q}} \left\{ ri - c(q)i + \Sigma_{j \in S} \beta v(j) P_{ij}(q) \right\}, \forall i \in S. \tag{4.1}$$

We are interested in the optimal stationary policy $\mathbf{q}^* = \{q^*(0), q^*(1), \dots\}$ i.e., the hospice provides quality $q^*(i)$ whenever the number of patients is $i \in S$. The following proposition shows the existence of an optimal policy. All proofs appear in Appendix C.1.

Proposition 4.1. *If there exists a policy* $\boldsymbol{q}^*$ *satisfying* $q^*(i) = \arg\max_{q \in \mathcal{Q}} \{ri - c(q)i + \Sigma_{j \in S} \beta v(j) P_{ij}(q)\}$, *then* $\boldsymbol{q}^*$ *is an optimal policy.*

Even though the rewards are unbounded, Proposition 4.1 shows that a stationary optimal policy exists and suggests an approach to compute this policy.

Before turning to finding an optimal policy, we will first consider an alternate model where the arrival rate is independent of the quality provided, i.e., $\lambda(q) = \lambda_0$. This model can be interpreted as representing a scenario in the absence of quality reporting, where the hospice has a steady market

[5] For simplicity we assume that all patients are provided the same quality. The hospice could create a personalized plan of care for each patient; in this case we would consider q as an aggregate quality measure.

[6] https://www.kindredhealthcare.com/our-services/hospice/faqs

share. Because patient departures are primarily determined by the actual quality of care provided and not the reported measures, we assume that $p(q)$ is the same in the absence and presence of reporting[7]. The following lemma characterizes the optimal policy for this model.

Lemma 4.1. *The optimal policy when* $\lambda(q) = \lambda_0$ *is* $q_0 = \min\left\{\bar{q}, \arg\max\left\{\frac{r-c(q)}{1-\beta(1-p(q))}\right\}\right\}, \forall i$. *Moreover, the return is given by* $v_0(i) = \alpha i + \gamma$, *where* $\alpha = \frac{r-c(q_0)}{1-\beta(1-p(q_0))}$ *and* $\gamma = \frac{\beta\lambda_0\alpha}{1-\beta}$.

Lemma 4.1 shows that a constant quality policy is optimal in this alternate model. To see the intuition behind this policy, consider the effect of a given action (quality q) on the hospice: Each patient earns the hospice a constant reward, and for a given q, has a constant probability of departing by the next stage. The arrival of new patients is unaffected by quality, and the only trade-off – between the reward (decreasing in q) and the departure probability (also decreasing in q) – is identical for *each* patient. Therefore, the optimal policy chooses quality q_0 regardless of the number of patients.

The intuition behind the linearity of $v_0(i)$ is similar: Each patient earns the hospice a fixed return, which is independent of the state i due to the optimality of the constant quality policy. Thus, we have a linear term in i. In addition, new patients that arrive to the system at rate λ_0 provide additional return; not surprisingly, the return γ from new patients is directly proportional to the per patient return rate α with an appropriate discounting factor – the term β in the numerator denotes that the return is incurred starting from the subsequent stage whereas the term $1-\beta$ in the denominator captures discounting over the infinite horizon. The per patient return α is simply the profit per patient $r - c(q_0)$ discounted by $\beta(1 - p(q_0))$, where $1 - p(q_0)$ reflects the probability that a customer will not depart by the next stage.

In the presence of reporting (when $\lambda(q)$ is increasing in q) the problem is no longer separable by patient. This is because providing a higher quality increases the arrival rate of new patients, and the hospice may choose to utilize this to modify the patient census to more profitable levels, the value of which must be balanced against the *aggregate* change in profit across its patient census, which depends on the current census. Thus, the optimal policy is in general complicated, and we must resort to heuristics, which we do next.

4.3.1 A heuristic policy

By exploiting the linear structure of the return function in Lemma 4.1 we can evaluate the first step of a policy improvement algorithm for the original model in closed form. This is because the summation term in (4.1) simplifies to terms involving only the expected value of next state $X_i(q)$ when $v(j)$ is linear. The resulting policy is our heuristic, which is defined below.

Definition 4.1. *The heuristic policy* $\boldsymbol{q}_H$ *is given by* $q_H(i) = \arg\max_{q\in\mathcal{Q}}\{r - c(q) + \beta\alpha(1-p(q)) + \beta\alpha\frac{\lambda(q)}{i}\}, \forall i > 0$, *and* $q_H(0) = \bar{q}$, *where* $\alpha = \frac{r-c(q_0)}{1-\beta(1-p(q_0))}$.

[7] Certainly published annual quality scores could effect departures, but we view this as a second-order effect, and disregard it.

A brief derivation of $\mathbf{q}_H$ is provided in Appendix C.1. The following lemma presents some structural properties of the heuristic policy and the optimal policy in the absence of reporting.

Lemma 4.2. *For all states $i \in S$,*
(a) $q_H(i)$ is non-increasing in i and $q_H(i) \geq q_0$,
(b) $q_H(i)$ and q_0 are non-decreasing in r,
(c) If $q_H(i) \neq \bar{q}$, then $q_H(i)$ becomes larger as $\lambda'(q)$ becomes larger; otherwise, it stays the same.

Lemma 4.2(a) reveals that the quality level of our heuristic is non-increasing in the patient census. Moreover, it is bounded below by q_0 – the quality chosen by the optimal policy in the absence of reporting – indicating that the quality provided in every state by our heuristic in the presence of reporting is at least as large as that in the absence of reporting. While the hospice extracts the maximum value out of its existing patient census at quality q_0 (Lemma 4.1), providing a higher quality increases the number of new patients that are expected to join. This is particularly beneficial when the patient census i is small (Lemma 4.2(a)) as the resulting costs from providing higher quality will not be too large when i is small.

Lemma 4.2(b) shows that the quality chosen by both the optimal policy in the absence of reporting and the heuristic policy is non-decreasing in the reimbursement rate r: As patients earn more revenue for the hospice it makes sense for the hospice to work harder to attract and retain them. And thus it can provide a higher quality of care for its patients. As the arrival rate becomes more sensitive to quality, the quality chosen under the heuristic may be larger (Lemma 4.2(c)). When quality goes a long way – by attracting more patients – it makes sense for the hospice to provide a higher quality of care for its patients.

In the next section we will run several numerical experiments to evaluate the performance of this heuristic policy. We will also identify, numerically, whether the structural properties of the heuristic extend to the optimal policy; for example, whether the quality under the optimal policy is also non-increasing in $i \in S$.

4.4 Numerical results

In this section we will examine the performance of our proposed heuristic policy using numerical approaches. We will first calibrate the parameters of our model to capture a typical hospice as well as design experiments to evaluate multiple variants in §4.4.1; we then compare the performance of the optimal and heuristic policies in §4.4.2. We numerically explore the strategic choices of a hospice under pay-for-reporting in §4.4.3.

4.4.1 Experimental setup

Since 98% of patient care days fall under RHC (NHPCO 2016), we will consider a single class of RHC patients as in §4.3. We will consider two types of hospices – for-profit and nonprofit – in our experiments owing to significant differences both operationally and in performance; for example, the average LoS was 105 and 65 days for for-profit and nonprofit hospices, respectively (MEDPAC 2017).[8]

We use the Hospice Compare dataset published by CMS (2017) to identify the average quality of hospices. This dataset contains reported measures of quality from hospice providers collected as part of the HQRP as well as the type of hospice ownership i.e., for-profit or nonprofit. We focus on one specific measure to calibrate our model – NQF #1637 Pain Assessment – for two reasons: 1) It shows the highest variation between hospice ownership types in the dataset, and indeed, this measure stands out in its ability to help customers differentiate between hospices (STAT 2017); and 2) It is defined as "Percentage of patient stays during which the patient screened positive for pain and received a comprehensive assessment of pain within 1 day of the screening," which suggests that it is a more in-depth and rigorous proxy for quality of care compared to others which are either focused on the initial assessment period or deal with a specific concern limited to only some patients. The average score for the Pain Assessment measure was found to be 80% and 75% for for-profit and nonprofit hospices respectively. We limit the choice of quality to $q \in [0, 1]$, so that the calibrated values for quality are 0.8 and 0.75 for for-profit and nonprofit hospice types, respectively. We will use these values in calibrating other parameters of our model.

Given the fundamental differences in the staffing mix of hospices by ownership type (Cherlin et al. 2010), we will differentiate between the two types solely based on their cost function. We assume that the cost of quality is given by $c(q) = c_0 q^2$, where c_0 depends on the type of hospice.[9] To compute the cost of quality we use estimated data from Killaly and Mukamel (2007): The median daily cost is \$91.47 for a cancer patient and \$61.11 for a non-cancer patient whereas the fixed cost per admission is \$897. Noting that cancer patients form 27.7% of the patients (NHPCO 2016), we have the median daily cost per patient as \$69.5.[10] We next use the average LoS by hospice type – 105 days and 65 days for for-profit and nonprofit respectively (MEDPAC 2017) – to amortize the fixed cost per admission into the daily costs, thus obtaining a total effective cost per patient per day of \$78.1 and \$83.3 for for-profit and nonprofit hospices respectively. Assuming the quality to be the calibrated value described above, we can estimate c_0 such that the average daily cost matches our estimate. The value of r is given by the Medicare reimbursement rate for RHC, which

[8] The larger LoS for for-profits may be due to the lower share of cancer patients – who have short LoS (NHPCO 2016) – in their patient mix (Wachterman et al. 2011).

[9] We use this functional form for $c(q)$ as it is simple, convex and increasing in q, and widely used in related work: for example, Ata et al. (2013) uses a similar function for the cost of recruiting hospice patients at a given rate.

[10] This is obtained by taking the weighted average of \$91.47 and \$61.11 assuming that the proportion of cancer patients is 27.7%. An alternate approach is to use a different proportion depending on the hospice ownership type to compute this cost separately for each type, but this will only amplify any difference in the cost parameters since the proportion of cancer patients at for-profits may be lower (Wachterman et al. 2011).

Table 4.1: Parameter values for $\lambda(q)$ and $p(q)$ used in experiments.

No.	$\lambda(q)$ parameters	Value
1	(a_0, a_1)	(2, 1)
2	(a_0, a_1)	(4, 1)
3	(a_0, a_1)	(3, 2)
4	(a_2, a_3)	(2, 0.5)
5	(a_2, a_3)	(4, 1)

No.	$p(q)$ parameters	Value
1	(b_0, b_1)	(0.1, 0.04)
2	(b_0, b_1)	(0.15, 0.15)
3	(b_0, b_1)	(0.3, 0.2)
4	b_2	0.5
5	b_2	0.35

was \$159.34 in 2015 (CMS 2014). We normalize $r = 1$ and scale the costs appropriately, which gives $c_0 = 0.76$ and $c_0 = 0.92$ for for-profit and nonprofit hospice types respectively[11].

We will now describe the functional forms for the patient census dynamics. In order to ensure a fair comparison in the strategies of the two types of hospices, these functions are assumed to be independent of the hospice type[12]. Specifically, we will run the following 50 experiments (25 each for the two types of hospices) with the following functions for the arrival rate and departure probability[13]:

- $\lambda(q) \in \{a_0 + a_1 q, a_2 + a_3\sqrt{q}\}$,
- $p(q) \in \{b_0 - b_1 q, b_2(1 - \sqrt{q})\}$,

where the parameters are shown in Table 4.1.

Noting that the average daily census in 2015 was 63 (NHPCO 2016) and the average LoS was 87 days (MEDPAC 2017), the average arrival rate is 0.73 patients per day (or, 5.1 patients per week) by Little's Law. Assuming that the LoS is exponential with a mean of 87 days, we can compute the probability of a patient departure in a week as 0.08.[14] The parameters in Table 4.1 are chosen such that the resulting values of $\lambda(q)$ and $p(q)$ at the calibrated quality vary (smaller or larger) roughly by up to a factor of 2 times that of the above average values.

4.4.2 Comparison of optimal and heuristic policies

To compute the optimal policy we implement an augmented value-iteration algorithm exploiting the structure of the heuristic policy: First, we use the heuristic policy to compute a good starting return function to be used in the value-iteration step. Specifically, using the heuristic policy in the place of the optimal policy in the R.H.S. of (4.1) yields a system of simultaneous linear equations, the solution to which is the starting return function $\mathbf{v}_H$. Second, we implement a value-iteration

[11] The fact that for-profit hospices have a lower cost is in line with the findings of Cherlin et al. (2010).

[12] One can certainly assign different functions for the two hospice types, for example to capture differences in the mix of patients (Wachterman et al. 2011). By ensuring that both types are faced with identical patients, our approach allows us to isolate the effects of their quality choices.

[13] The functional forms are either linear or sublinear in order to explore different potential relationships while still keeping them simple.

[14] The exponential assumption allows us to compute this probability while ignoring historical LoS effects.

algorithm that improves this return function by recalculating the policy that maximizes the R.H.S. of (4.1) until the return function converges. The corresponding policy is our optimal policy, and the corresponding return function is the optimal return $\mathbf{v}^*$. Let $q^*(i)$ be the quality chosen in state $i \in S$ by the optimal policy.

We derive the stationary probability $\pi^a(i)$ of being in state i under policy a by using the transition matrix corresponding to that policy; we obtain the stationary probabilities $\pi^*(i)$ and $\pi^h(i)$ corresponding to the optimal and heuristic policies respectively. We use these to compute the expected quality Q^a for each experiment under policy a as follows: $Q^* = \sum_{i \in S} q^*(i)\pi^*(i)$ and $Q^h = \sum_{i \in S} q_H(i)\pi^h(i)$. Similarly, we can also compute the expected patient census $N^* = \sum_{i \in S} i\pi^*(i)$ and $N^h = \sum_{i \in S} i\pi^h(i)$ under the optimal and heuristic policies respectively. We can use the average quality, and therefore the arrival rate, to compute the expected LoS by using Little's Law.

We assume that the discount factor β is 0.95, and to facilitate computation, we discretize quality to increments of 0.001 and truncate the space space at a large value[15]. New patients arrive according to a Poisson process with rate $\lambda(q)$, and each existing patient either departs before the next stage with probability $p(q)$ or stays with the hospice in the next stage with probability $1 - p(q)$.

Table 4.2: Comparison of optimal and heuristic policies; averages over 25 settings for each type.

Type	Avg. quality		Avg. patient census		Avg. LoS (days)		Avg. profit	
	Heuristic	Optimal	Heuristic	Optimal	Heuristic	Optimal	Heuristic	Optimal
For-profit	0.640	0.637	74.28	73.91	158.6	157.9	408.83	408.86
Nonprofit	0.523	0.521	42.56	42.35	90.2	89.8	365.54	365.57

Note: For average profit we use $v_H(0)$ and $v^*(0)$ assuming that the system starts empty.

We compute several performance metrics – the expected quality, the expected patient census, and the expected LoS – under the heuristic and optimal policies for each of the 25 experiments described in Table 4.1. We can also compute the discounted profits starting from any state, i.e., $\mathbf{v}_H$ and $\mathbf{v}^*$ under the heuristic and optimal policies respectively. We then aggregate these metrics across the experiments for the two types of hospices, and report these averages in Table 4.2.

The heuristic performs extremely well across all metrics for both types of hospices and is a very good approximation for the optimal policy. In all our experiments, we find that the return when starting in the empty state under the heuristic policy i.e., $v_H(0)$ for both hospice types, is always within 0.05% of that under the optimal policy i.e., $v^*(0)$, and the difference between them under the two policies averaged across all experiments is $\approx 0.01\%$, confirming that the heuristic is a very good one. Moreover, the average quality under the heuristic is always within 3% of the optimal policy across our experiments.

The two hospice types differ in their cost structure: One would expect that the lower cost of quality would enable a for-profit hospice to provide a higher quality of care thus attracting more

[15] We determine this large value such that the stationary probability of being in this state is at most 1e^{-10}.

Figure 4.1: Optimal and heuristic policies: For-profit, $(a_0, a_1) = (2, 1)$, and $(b_0, b_1) = (0.15, 0.15)$.

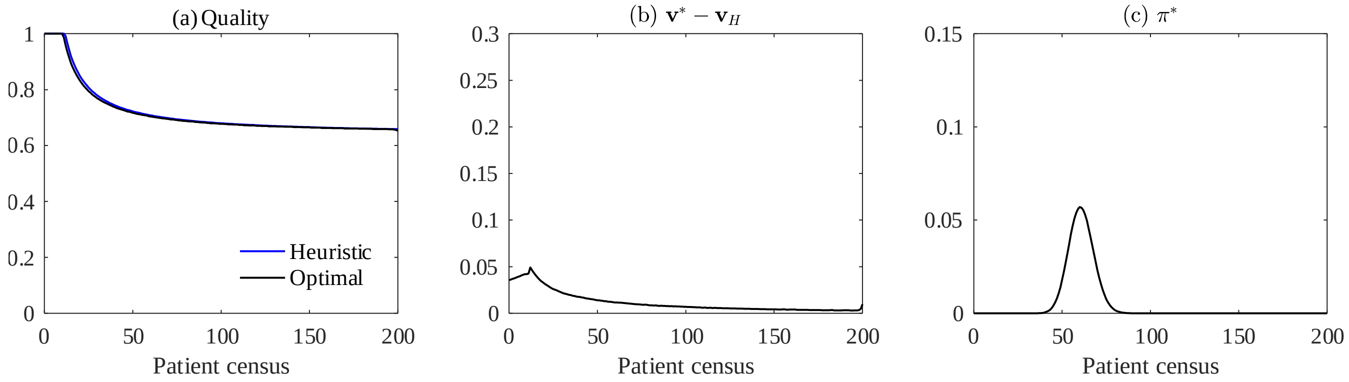


Figure 4.2: Optimal and heuristic policies: For-profit, $(a_0, a_1) = (3, 2)$, and $(b_0, b_1) = (0.15, 0.15)$.

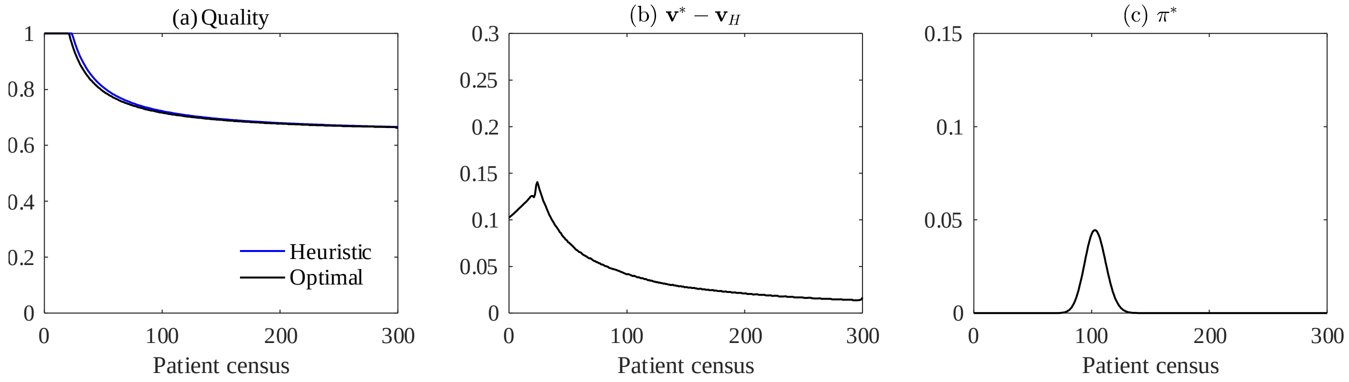


Figure 4.3: Optimal and heuristic policies: For-profit, $(a_0, a_1) = (3, 2)$, and $(b_0, b_1) = (0.1, 0.04)$.

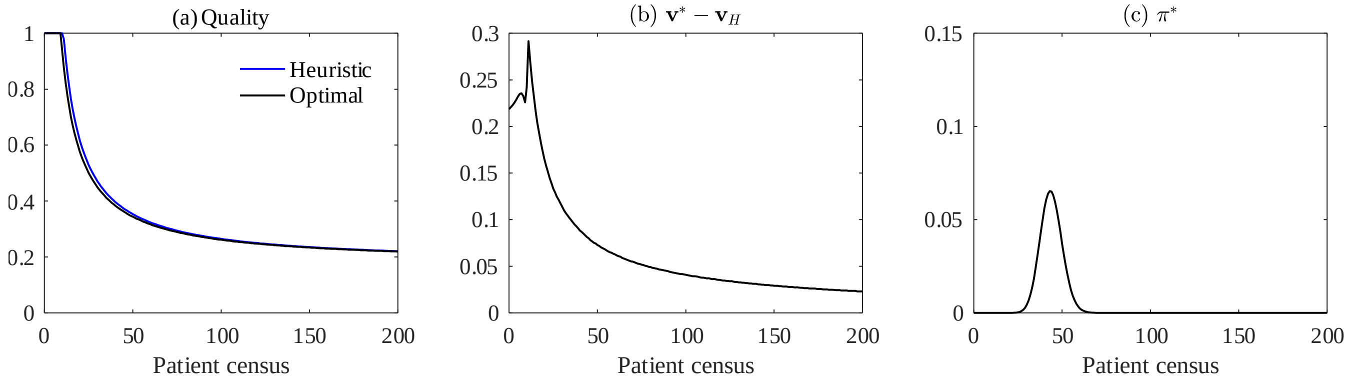


Figure 4.4: Optimal and heuristic policies: Nonprofit, $(a_0, a_1) = (2, 1)$, and $(b_0, b_1) = (0.15, 0.15)$.

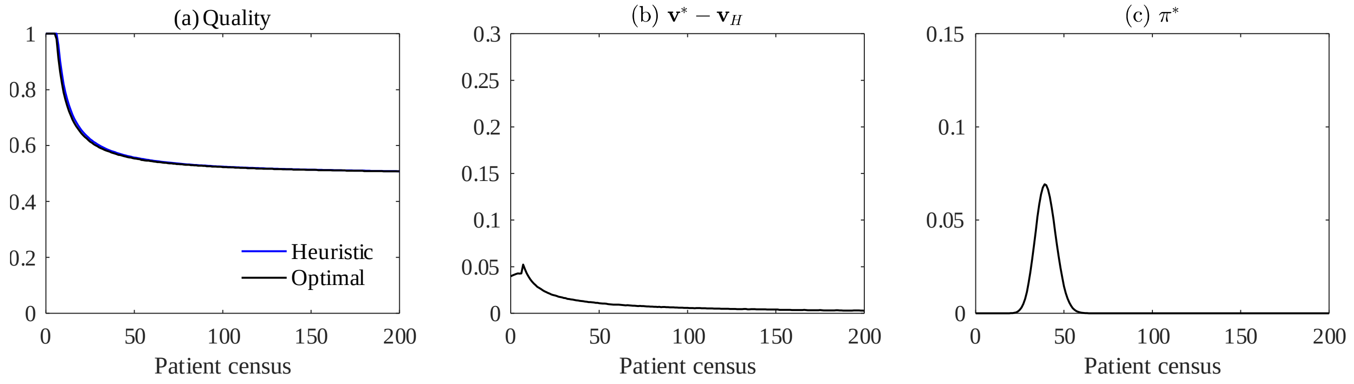


Figure 4.5: Optimal and heuristic policies: Nonprofit, $(a_0, a_1) = (3, 2)$, and $(b_0, b_1) = (0.15, 0.15)$.

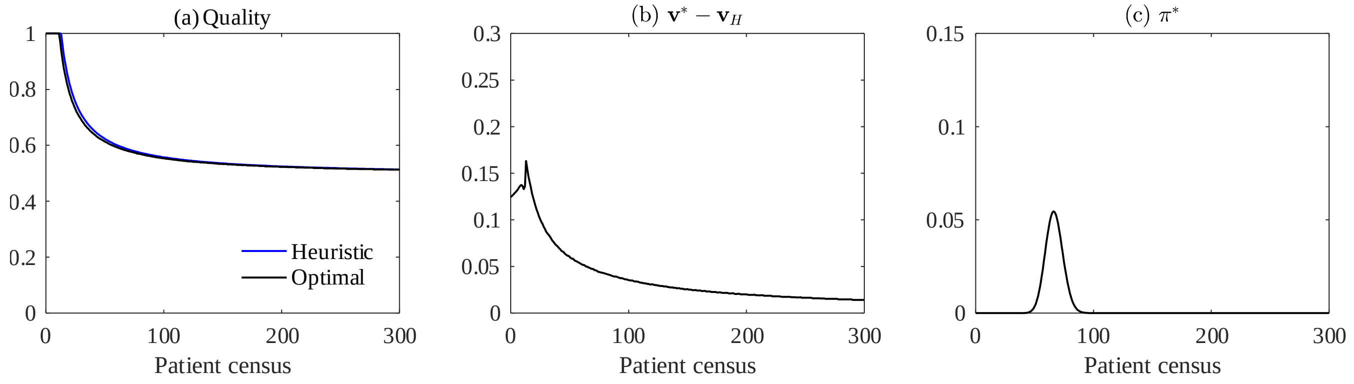
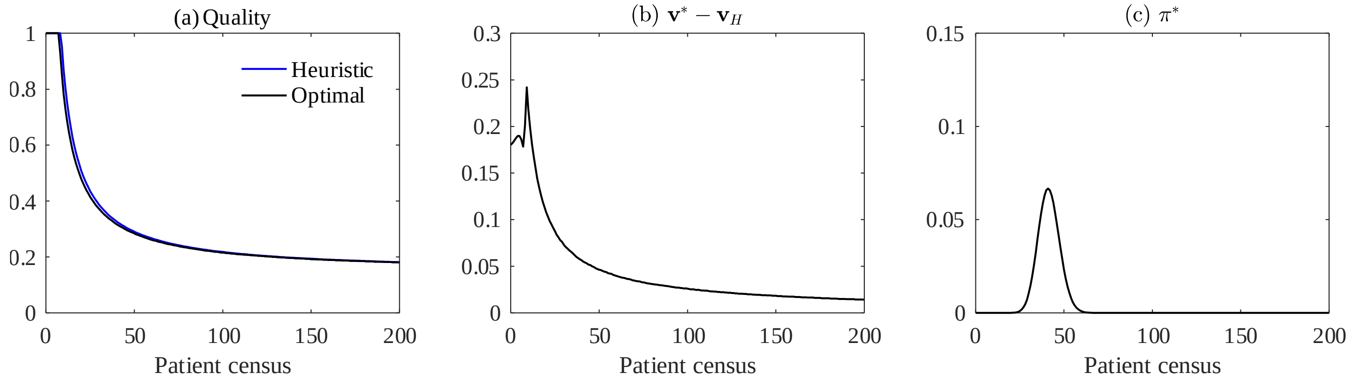


Figure 4.6: Optimal and heuristic policies: Nonprofit, $(a_0, a_1) = (3, 2)$, and $(b_0, b_1) = (0.1, 0.04)$.



patients, who in turn provide more revenue. Table 4.2 shows this to be true; the for-profit type provides a higher quality, resulting in a longer LoS, compared to the nonprofit type, which is in line with MEDPAC (2017).[16]

Figures 4.1-4.3 compare the performance of the two policies as a function of the patient census in three representative experiments for a for-profit type hospice: Panel (a) shows the quality chosen by the optimal and the heuristic polices, panel (b) shows the difference in the profit functions, and panel (c) shows the stationary probability distribution corresponding to the optimal policy[17]. (Figures 4.4-4.6 show the same for a nonprofit type hospice.) The optimal policy has a similar structure to the heuristic; both policies are non-increasing in the number of patients. (Lemma 4.2(a) showed this to be true for the heuristic policy.) While this property does not hold in general for the optimal policy (see Appendix C.2), we find that it is true in all our settings. Moreover, the quality under the heuristic policy is very close to that under the optimal policy across all states in addition to on average (Table 4.2). Interestingly, the heuristic always chooses a higher quality level than the optimal policy. This is likely due to the myopic nature of the update the heuristic makes (by design) to the quality level q_0 chosen in the absence of reporting; the optimal policy considers a more integrated view incorporating the profit across the patient census through the optimality equation (4.1).

We find that the difference between the two policies is more pronounced when the arrival rate is highly sensitive to quality. Figure 4.2 corresponds to an experiment with an arrival rate that is more sensitive to quality relative to that in Figure 4.1, and shows a larger difference in the return function under the two policies (panel b). This is not surprising because the heuristic was developed by first assuming that the arrival rate is independent of quality, and as quality affects arrivals more strongly, the return under the heuristic policy will likely be further away from that under the optimal policy. Furthermore, Figure 4.2 shows a slightly larger quality level compared to Figure 4.1 (panel a)[18], which is in line with Lemma 4.2(c): As the sensitivity of the arrival rate to quality increases, the heuristic chooses a larger quality level.

Figure 4.3 corresponds to an experiment with a departure probability that is less sensitive compared to that in Figure 4.2, and shows a significantly lower quality level and a larger difference in the return function. The former is intuitive since quality does not sway patients' departure choices as much so that it is not worthwhile to increase it. The significant adjustments to the quality level across states in Figure 4.3, i.e., the steeper decline from the maximum level chosen in the empty state in both cases, manifests in a larger difference between the two return functions.

[16] In reality the longer LoS for for-profit types may also be due to the patient mix (Wachterman et al. 2011).

[17] The stationary probability distribution corresponding to the heuristic is very similar and thus omitted.

[18] The average quality under the optimal policy – the weighted average of the optimal policy in panel (a) with the distribution in panel (c) – is 0.704, 0.715 and 0.370 in Figures 4.1, 4.2 and 4.3 respectively.

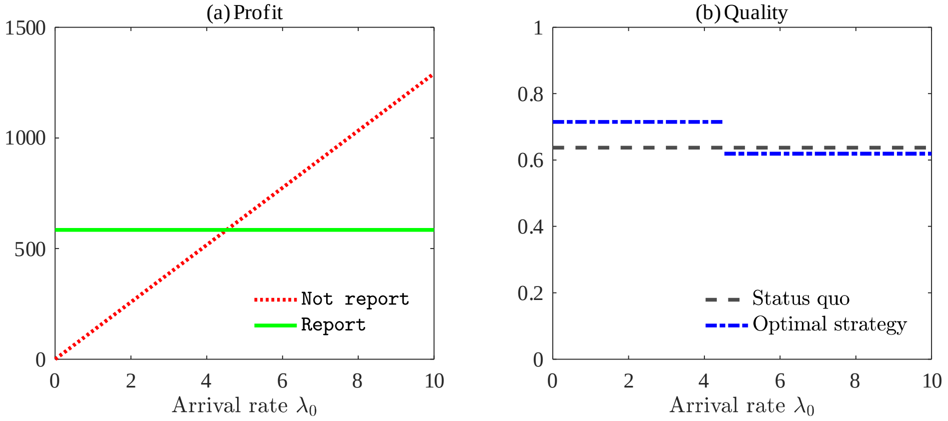
4.4.3 Incentives for quality reporting

One of the goals of the quality reporting program is to encourage providers to improve the quality of their services. Lemma 4.2(a) showed that the quality chosen by the heuristic policy is at least as large as the optimal quality in the absence of reporting. Moreover, §4.4.2 revealed that the heuristic is a very good approximation for the optimal policy, and for most realistic parameter settings, the optimal policy under reporting also chooses quality levels that are at least as large as that in the absence of reporting. Thus, if a hospice *chooses* to report, the quality of care will most likely improve.

But the question remains as to what type of hospices would choose to enroll in such reporting programs? We explore this question in light of the pay-for-reporting incentive structure that is currently in use: hospices that fail to submit quality metrics will be face a penalty of 2 percentage points on their Medicare reimbursement rate (CMS 2015). Let δ denote this penalty. Thus, the hospice faces two choices: (1) `Report` quality and expose itself to reputation effects influencing new patients' joining decisions; or (2) `Not report` quality, receive a lower reimbursement rate, but ensure that customers remain ignorant about its quality of care when they make their joining decision.

We will evaluate the profit under both choices separately; we then study the hospice's strategic choice regarding reporting quality. The optimal strategy is defined as the choice (`Report` or `Not report`) that gives the hospice a larger profit when starting from the empty state. We also consider the status quo scenario in the absence of quality reporting and under the standard reimbursement rate (i.e., without penalty). Lemma 4.1 gives the quality and profit under the status quo scenario and `Not report`: The quality is q_0 and the profit is $v_0(0) = \gamma$, where q_0 and γ depend on the reimbursement rate, i.e., r and $(1-\delta)r$ in status quo and `Not report` respectively. The average quality Q^* and profit $v^*(0)$ under the `Report` choice is computed numerically as in §4.4.2.

Figure 4.7: Comparison of strategies and quality of care under pay-for-reporting.



Note: This figure is for a for-profit with $(a_0, a_1) = (3, 2)$ and $(b_0, b_1) = (0.15, 0.15)$, and $\delta = 0.02$.

Figure 4.7 shows the profit and the quality level under the different outcomes as a function of λ_0 – the arrival rate in the absence of reporting. The profit $v^*(0)$ under `Report` is independent of λ_0, whereas the profit γ under `Not report` is linear in λ_0 (Lemma 4.1). As a result, the profit curves under the two choices cross each other at exactly one point, say λ_T, as shown in Figure 4.7(a), and the optimal strategy is to report whenever λ_0 is smaller than λ_T. A "fair" value for this arrival rate – one that corresponds exactly to the quality q_0 the hospice provides in the status quo scenario – is given by $\lambda_0 = \lambda(q_0)$. For the example in Figure 4.7 this value is $\lambda_0 = 4.3$ while the threshold $\lambda_\text{T} = 4.5$ suggesting that a hospice that enjoys 5% larger market share than what is considered fair would choose not to report.

In other words it need not be optimal for a hospice to enroll in reporting. Thus, the incentive to report quality depends on the extent of the market captured by the hospice in the absence of reporting: If the hospice enjoys a disproportionate market share (high λ_0), then it may be better off not revealing its quality: Revealing this may just chase customers away. Moreover, the threshold λ_T is increasing in δ[19], suggesting that imposing such a penalty may serve as a way to encourage reporting to be the strategic choice.

Figure 4.7(b) compares the quality level chosen by the hospice under the optimal strategy (identified above) against that under the status quo scenario. Interestingly, the hospice *reduces* its quality level when it chooses not to report. Indeed, Lemma 4.2(b) showed that the quality level in the absence of reporting is non-decreasing in the reimbursement rate, or equivalently non-increasing in the penalty δ. As the penalty increases, the drop in quality for hospices that choose to not report quality will be more pronounced.

Thus, the pay-for-reporting incentive structure is a double-edged sword: On the one hand, it may encourage a hospice to enroll in reporting, and if chooses to do so, the quality may improve. On the other hand, some hospices may strategically choose not to report quality, and may actually reduce their quality level due to the reduced reimbursement rate. An alternate approach is to reward hospices that report quality instead of penalizing those that fail to report. In this case the quality under the optimal strategy would be at least as good as that in the status quo scenario. But, the threshold to choose reporting may be smaller relative to that in Figure 4.7(a)[20], in which case fewer hospices would strategically report quality.

4.5 Conclusion

CMS has recently required hospices to report their performance across a given set of quality measures. These measures are publicly available and are expected to help patients make better choices

[19] The threshold λ_T is given by the solution to $v^*(0) = \frac{\alpha \lambda_\text{T} \beta}{1-\beta}$ (Lemma 4.1). We know that α is increasing in the reimbursement rate i.e., $r(1-\delta)$ (or, decreasing in δ) from the proof of Lemma 4.2(b). Thus, the threshold λ_T is increasing in δ.

[20] This occurs when the impact of the reward on boosting $v^*(0)$ is dominated by that of the penalty on lowering γ.

when it comes to hospice care. In this chapter we examine a hospice's choice of quality of care considering the effects of reporting on the rate at which new patients join the hospice.

We model the quality control problem as a discrete-time Markov decision process: Our results serve as a guide for hospice managers to determine any changes – for example, in staffing levels or schedule of visit services – to be made depending on the patient census realized in a given week. The patient census is governed by general transition probabilities, wherein the departure probability and the arrival rate may depend on the quality chosen. We show that a constant quality policy is optimal in the absence of reporting. On the other hand, when quality is reported it affects the rate of new arrivals, and the optimal policy is, in general, complicated.

We exploit the structure of the problem to derive a closed-form heuristic that performs very well against the optimal policy: The average quality under the heuristic is within 3%, and the profit is within 0.05%, of the optimal policy across our numerical experiments. Not surprisingly, a for-profit hospice – characterized by lower quality costs – provides higher quality, and experiences longer stays, on average compared to a nonprofit one.

We derive several structural properties of our heuristic: The policy chooses quality levels that are non-increasing in the patient census, demonstrating the benefit of providing better quality – thereby attracting more patients – when the cost of doing so is not too large. Moreover the heuristic offers a quality that is at least as large as that in the absence of reporting. Through our numerical experiments we find that these properties translate very well to the optimal policy for realistic parameter settings; however, we also identified situations where it does not.

By comparing two systems – before and after reporting of quality is in place – we find that reporting does incentivize hospices to provide better quality of care. However, not all hospices may be willing to participate in a voluntary quality reporting program: Those that enjoy a disproportionate share of the market in the absence of reporting, and thus risk losing customers upon reporting their actual quality publicly, will be less inclined to do so. While the current "pay-for-reporting" approach implemented by CMS does serve as a way to encourage hospices to report quality, it does have some unintended consequences: Hospices that strategically choose not to report may actually reduce their quality level due to the penalty imposed for not reporting. To the best of our knowledge, this is the first work that models the relationship between quality of care and patient census explicitly to explore incentives for participating in quality reporting in the hospice setting.

There are several directions to extend this work. The patient population may be extended to multiple classes where patients may have different diagnoses and/or require different levels of care. An interesting effect to consider in this case is possible movement of patients between classes – both naturally as well as due to poor monitoring by hospice staff – and whether reimbursement rates should be class-dependent. One may also wish to explore the joint effects of having alternate reimbursement structures – for example, the two-tier rate for RHC – along with quality reporting.

Another variant is to consider the impact of offering specialty hospice programs such as for cardiac or pulmonary care. While patients in these programs typically require higher levels of care, offering such programs may help hospices differentiate themselves from their competition. Finally, it is of interest to identify a metric of quality that is both controllable and amenable to fair comparison across hospices with different mixtures of patients.

Chapter 5

Conclusion

This dissertation develops stochastic models to analyze problems in service operations, particularly dealing with issues in healthcare and retail operations. Within healthcare, we explore when multiple listing for organ transplantation is fair given criticism that this provides disproportionate advantages to wealthier patients (Chapter 2), and the quality of care chosen by a hospice in light of recent regulations around reporting (Chapter 4). Within retail operations, we investigate the different options for return policies of a typical omnichannel firm, taking into account the fact that customers are increasingly using more than one channel for their interactions (Chapter 3). The insights generated through these studies provide guidance for practitioners as well as policy makers.

The work in this dissertation makes several contributions to the theory of operations management. The analysis on fairness under redundancy was facilitated by the development of a novel coupling method and a counting procedure to track customers (presented in Chapter 2). The coupling method allows one to process state transitions out of temporal order to obtain an ordering of the queue lengths, while the counting procedure helps define queue lengths in spite of customers' (dynamically) changing positions in queues. We believe that these methods are of interest to the research community for two reasons. First, the coupling method presents a new way to study systems that are difficult to analyze using traditional sample path approaches. And in conjunction with sample path techniques, the counting procedure may be invoked whenever there is no straightforward way to record customers. Moreover, both of these methods lend themselves to extensions involving customer abandonments – which are notoriously difficult to study analytically – demonstrating their power to aid in the analysis of otherwise complex queueing systems. Second, this work has implications beyond redundancy systems: The smallest workload policy is closely related to certain redundancy policies, and our results extend prior work in this area considering a single (flexible) class to two customer classes, only one of which is flexible.

The work on fairness in systems with redundancy helps identify when, and by how much, joining multiple lines may be unfair as compared to joining the shortest queue, and as such could be a useful tool for policy makers in organ transplantation. In addition, the optimal queue joining analysis indicates when a shopper should join the shortest checkout line at the supermarket depending on whether she is aware of which shoppers are "paired up." Future work could broaden the scope of the model to include aspects of organ transplantation such as the acceptance decisions of patients, endogenous listing when redundancy comes with a cost, and priority-based policies.

The work investigating return policies of omnichannel firms provides a potential explanation for the different policies adopted by real retailers, and helps guide when a firm should choose full refunds versus when it should charge a fee for returns. Future work on this front could explore the inventory aspect of omnichannel operations: How can an omnichannel firm gain from jointly managing its inventory levels and return policy, especially given the high return rates in the online apparel industry. Another direction is to explore how return windows – the time that a customer has to return an item for a refund – impact customers' return decisions (e.g., due to either the forgetful nature of customers or the endowment effect). This in turn affects the firm's policy: A lenient policy (i.e. a large window) increases the chance that an item may need to be salvaged at a loss, but it could also reduce returns. Some firms adopt innovative policies for returns, thereby introducing new directions for future work: Firms (e.g., Jet) may offer customers the option to self-select their return behavior (such as opting out of free returns) at the time of purchase using a menu of prices, and one could investigate when a firm should offer such an option and how to design such a contract.

The findings from the quality choices of hospices suggests that the current approach of reporting metrics publicly is a step in the right direction, but cautions policy makers that the incentive structure may result in unintended consequences such as a reduction in the quality levels. Future work could empirically investigate the impact of quality on patient dynamics, incorporating a study of a hospice's actual quality choices. One can also explore the relationship between patient satisfaction ratings and the various metrics of hospice quality currently reported to identify which key metrics matter the most to patients. Another direction to pursue is customers' response to quality disclosure and its implications for the hospice's quality choices.

An important application on the intersection of queueing theory and omnichannel strategy is *omnichannel service* operations. Specifically, how should a firm (e.g., Starbucks) that receives orders both in person and through a mobile interface serve its customers. This is particularly interesting given that an uptick in mobile ordering is a double-edged sword: While it streamlines the ordering and payment process resulting in improved customer wait times, it may also lead to bottlenecks from ineffective juggling between walk-ins and mobile orders, as well as uncertainty with respect to when customers will actually arrive to claim their orders.

Appendix A

Supplements to Chapter 2

A.1 Proofs

Proof of Theorem 2.1. We will first show that $T^{(\text{Ded,CUS})} \leq T^{(\text{Ded,CUE})}$. To compute $T^{(\text{Ded,CUE})}$ we use the stationary distribution from Visschers et al. (2012); Appendix A.2.1 provides details on the steps involved. From Lemma A.5, when $p = 0.5$, we have

$$T_1^{(\text{Ded,CUE})} = \frac{1}{2\mu} + \frac{1}{2\mu - \lambda} + \frac{1}{2\mu - \lambda(1-r)} - \frac{1}{2\mu + r\lambda}, \tag{A.1}$$

$$T_2^{(\text{Ded,CUE})} = \frac{1}{2\mu - \lambda} + \frac{1}{r(2\mu + r\lambda)} - \frac{1-r}{2r\mu}, \tag{A.2}$$

$$T^{(\text{Ded,CUE})} = \frac{1}{2\mu - \lambda} + \frac{1-r}{2\mu - \lambda(1-r)} + \frac{r}{2\mu + \lambda r}. \tag{A.3}$$

From Gardner et al. (2016) we have that

$$T_1^{(\text{Ded,CUS})} = \frac{p}{\mu - \lambda(1-r)p} + \frac{1-p}{\mu - \lambda(1-r)(1-p)} + \frac{1}{2\mu - \lambda} - \frac{1}{2\mu - \lambda(1-r)}, \tag{A.4}$$

$$T_2^{(\text{Ded,CUS})} = \frac{1}{2\mu - \lambda}. \tag{A.5}$$

When $p = 0.5$ this gives us

$$T_1^{(\text{Ded,CUS})} = \frac{1}{2\mu - \lambda} + \frac{1}{2\mu - \lambda(1-r)}, \tag{A.6}$$

$$T^{(\text{Ded,CUS})} = \frac{1}{2\mu - \lambda} + \frac{1-r}{2\mu - \lambda(1-r)}. \tag{A.7}$$

Together we have $T^{(\text{Ded,CUE})} - T^{(\text{Ded,CUS})} = \frac{r}{2\mu + \lambda r}$ when $p = 0.5$. Since $\frac{r}{2\mu+\lambda r} \geq 0$, we have our result.

Next, we show that $T^{(\text{Ded,CUE})} \leq T^{(\text{Ded,JSQ})}$. We first highlight classical results on majorization. Let vectors $x, y \in \mathbf{N}^n$ be restricted to take non-negative integer values, and let $\chi_x(q)$ be the index of the q-th largest component of x. Let $\mathbf{e}_i \in \mathbb{N}^n$ be the vector whose components are zero except for the i-th component which is one. Define $A_q(x) = x + \mathbf{e}_{\chi_x(q)}$ as the vector that is obtained by adding one to the q-th largest component, and $D_q(x) = x - \mathbf{1}(x_{\chi_x(q)} > 0)\mathbf{e}_{\chi_x(q)}$ as the vector obtained by subtracting one from the q-th largest component of x if this component is strictly positive. Then, from Definition 3.2 and Lemma 3.2 in Liu et al. (1995) we have:

(L1) x is weakly submajorized by y i.e., $x \prec_w y$ if $\sum_{q=1}^{k} x_{[q]} \leq \sum_{q=1}^{k} y_{[q]}, k = 1, 2, \ldots, n$, where $x_{[\cdot]}$ denotes the component of x in decreasing order;

(L2) if $x \prec_w y$, then $A_q(x) \prec_w A_s(y), 1 \leq s \leq q \leq n$;

(L3) if $x \prec_w y$, then $D_q(x) \prec_w D_s(y), 1 \leq q \leq s \leq n$.

Consider a sample path, ω, that contains arrival times and class types for each customer, and service completion times at each server. We consider two systems – R and J – differentiated only by the policy (CUE and JSQ respectively) utilized by class 2 customers on ω. We assume that both systems start empty. The queue length vector in the J system is straightforward; queue length in the R system is defined as follows. For each redundant customer we assign one of the two jobs to be the "real" job: In particular, the redundant job that is closer to the server is considered to be real job (adds one to that queue), and the other one is not counted, breaking ties arbitrarily.

Arrivals do not change which job of a redundant customer is counted in this way, but the assignment may change after a departure event: After a service completion, if the "real" job of a redundant customer faces *more* jobs ahead of it than the other copy, then the latter copy will be counted as the "real" job; otherwise the assignment is unchanged. Conceptually, we think of the systems evolving as we sequentially process events along ω. We also consider two "modified" systems, $\tilde{R}$ and $\tilde{J}$, on ω that differ from the corresponding original systems in the sequence in which the given arrival and service events are processed. The three scenarios that the modified systems could be in, and the associated sequencing rules for advancing states, are:

(A) Any queue (in $\tilde{R}$ or $\tilde{J}$) is idle. In this case advance both systems to the next event (service or arrival).

(B) No queue (in $\tilde{R}$ or $\tilde{J}$) is idle and there is no redundant customer in the *queue* in the $\tilde{R}$ system. In this case advance both systems to the next event (service or arrival).

(C) No queue (in $\tilde{R}$ or $\tilde{J}$) is idle and there is at least one redundant customer in the *queue* in the $\tilde{R}$ system. In this case advance both systems to the next *service* event, deferring arrival events as necessary.

Both systems start empty and have the arrival and service events considered in the order they occur until the arrival of the first redundant customer to face two busy servers in $\tilde{R}$. This moves the system to (C), and only service events will be considered; arrival events, if any, will be tracked but will be added to the queue only when a service completion moves the system either to (A) if any of the queues are empty, or to (B) if the last redundant customer in the queue enters service in $\tilde{R}$. In either case, any customers that arrived when the system was in (C) are added to the system sequentially and instantaneously when the system transitions into (A) or (B). Finally, we move from (A) upon an arrival that leaves no queue idle, either to (B) if there are no redundant customers in the queue or to (C) if there is at least one redundant customer in the queue in $\tilde{R}$, whereas we move from (B) either to (A) if one of the queues idles or to (C) if a redundant customer arrives in $\tilde{R}$.

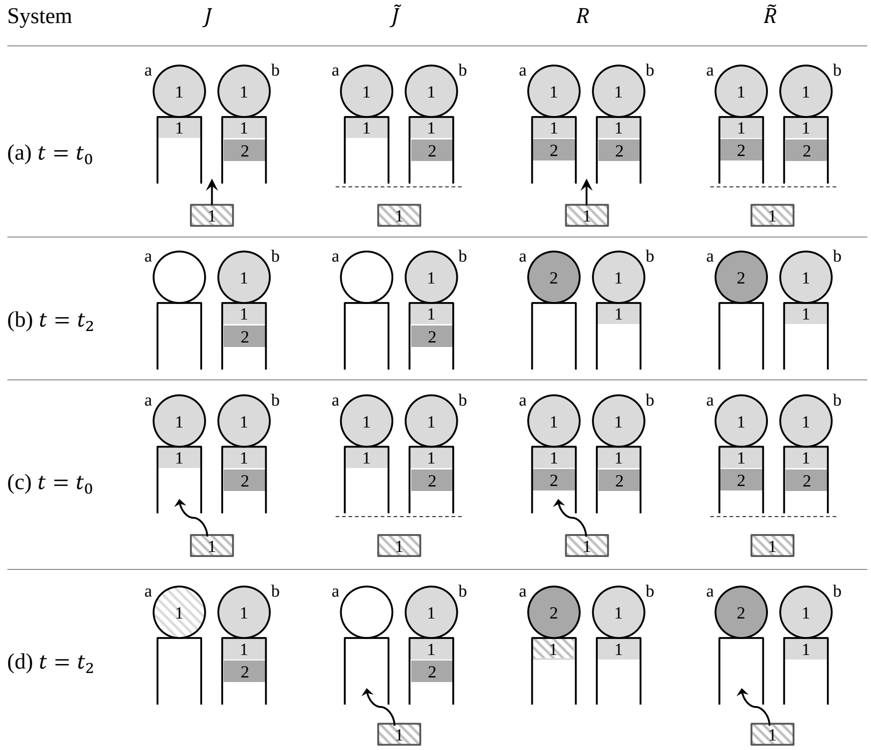
The routing probabilities for customers are coupled between the four systems as follows. If there are no redundant jobs in the queue in $\tilde{R}$ when a class 1 customer arrives (say customer n_0 arriving at time t_0), then all of the arrival and service events prior to time t_0 have been processed at all four queues. In this case customer n_0 joins either the shorter or the longer queue in all four systems at t_0. Ties, if any, may be broken arbitrarily; however, the routing in the modified system should be identical to that in the corresponding original system.

On the other hand, when a class 1 customer arrives and the system is in scenario (C), she is not assigned in the $\tilde{J}$ and $\tilde{R}$ systems but is in J and R. These assignment choices, to J and R, occur as follows: Customer n_0 looks forward to time τ when the $\tilde{(\cdot)}$ systems will transition out of state (C) i.e., τ is the minimum of the time until the last class 2 customer in front of customer n_0 in the R system enters service and the time until one of the servers idles in J. (We prove, in Lemma A.1, that the systems J and $\tilde{J}$, and R and $\tilde{R}$, are coupled on ω.) Customer n_0 evaluates the queue lengths of jobs that arrived prior to time t_0 and remain in the system at time τ. She chooses the shorter of these queues (in systems R and J) with probability 0.5, and the longer of these queues with probability 0.5, breaking ties arbitrarily (but consistently between the original and modified systems). When customer n_0 is eventually added to the $\tilde{J}$ and $\tilde{R}$ systems, she joins the queues the same way – either the shorter or the longer queue measured at time τ, according to what was selected for the J or R systems. If a class 2 customer arrives to the J system but is not assigned a queue upon arrival in the $\tilde{J}$ system immediately, then whenever she needs to be added in the $\tilde{J}$ system, she is assigned whichever queue she is presently at in the J system. There is no explicit routing for class 2 customers joining R and $\tilde{R}$ as they are added to both queues; however, if both the servers are idle, she joins whichever server she joins in J.

We present an illustration in Figure A.1. We start with two class 1 customers at each server (denoted a and b) followed by a class 2 customer (Figure A.1a). The modified systems are currently in scenario (C), and no further arrivals are considered (indicated by the dashed line in $\tilde{J}$ and $\tilde{R}$). Suppose the next three events are: (i) a class 1 customer n_0 arrives to the system at time $t = t_0$; (ii) a service completion occurs at server a at time t_1; and (iii) another service completion occurs at server a at time t_2. To identify which queue customer n_0 must be sent to in J and R at time t_0, we look forward in time to time $\tau = t_2$ (Figure A.1b) when the modified systems are no longer in scenario (C) (they are in scenario (A)). If we choose (according to a coin flip) the shorter queue in J at time t_2, then customer n_0 should be sent to queue a in both J and R at time t_0 (Figure A.1c). Note that she is not assigned a queue in $\tilde{J}$ and $\tilde{R}$ at this time. When the system eventually evolves to time t_2, customer n_0 will be assigned to queue a (i.e., the shorter queue at time t_2) in $\tilde{J}$ and $\tilde{R}$ (Figure A.1d), so that the systems J and $\tilde{J}$, and R and $\tilde{R}$, remain coupled.

Define $w(x)$ and $\tilde{w}(x)$ as the state vector ahead of customer x (*after* she is assigned a queue in the modified system) i.e., comprising all customers who arrived before x and are still in the system, in the original and modified systems respectively. The following lemma states that the two sets of

Figure A.1: An illustration of system state evolution in $J, R, \tilde{J}$ and $\tilde{R}$.



Note: A class $i \in \{1, 2\}$ customer is denoted by i in all four systems; customer n_0, described in Proof of Theorem 2.1, is indicated using diagonal stripes.

systems remain coupled.

Lemma A.1. *The two sets of systems – J and $\tilde{J}$, as well as R and $\tilde{R}$ – are coupled. In particular, $w(i) = \tilde{w}(i)$ for every customer i in both sets of systems.*

Proof. If $w(i) = \tilde{w}(i)$ for a customer i upon her assignment to a queue in the modified system, this remains true until customer i departs the system because the service events are coupled exactly between the original and modified systems. So we prove Lemma A.1 by induction on customers at the time when they are assigned a queue in the $\tilde{(\cdot)}$ system.

Consider the first set of systems: J and $\tilde{J}$. Since the two systems start empty, the statement is true for the first customer (regardless of her class). Let the two systems be coupled for the $(n-1)^{th}$ customer, say x, i.e., $w(x) = \tilde{w}(x)$ at the time when x is assigned a queue in $\tilde{J}$. We need to show that the next arrival to be assigned, say y, satisfies $w(y) = \tilde{w}(y)$ as well. If x is of class 1, she joins either the longer queue or the shorter queue in both (defined according to $w(x)$ and $\tilde{w}(x)$). If x is of class 2, she would have joined the shorter queue in J upon arrival, and joins the same relative position as in J upon being assigned a queue in $\tilde{J}$. Thus, the systems are coupled just after assigning x as well. Now, either y has already arrived and is waiting for an assignment in $\tilde{J}$ or y is yet to arrive and be assigned. In both cases, there may be some (possibly zero) service completions (regardless of the scenario (A), (B) or (C)) that occur before y is assigned. Each of these service completions, if any, are coupled i.e., they occur from the same queue in both systems.

We also need to show that y, if delayed in $\tilde{J}$, does not depart in J before being assigned in $\tilde{J}$. This could happen if the following three events occur together: (i) the queue that y should join in $\tilde{J}$ idles; (ii) y is not assigned immediately in $\tilde{J}$; and (iii) y is the next job in service at J. But if y is about to enter service in J, the corresponding queue will be idle in $\tilde{J}$, because the systems are coupled. Since the modified systems assign any delayed jobs as soon as one of the queues becomes empty (if not sooner), y will be added to the queue (and begins service) in $\tilde{J}$ immediately, so that y can never depart in J before being assigned in $\tilde{J}$. Thus, the two systems are coupled to begin with, and any service completions are coupled, so that the two queues remain coupled when assigning y as well. Therefore, by induction, the two systems remain synchronized, and $w(y) = \tilde{w}(y)$.

For the second set of systems – R and $\tilde{R}$ – we use a similar induction. Under the induction assumption as before, we claim that the two systems remain coupled just after assigning x. To see this, let us consider the class of customer x: If x is of class 2 type, then we add her to *both* queues in the two systems, and if one or more servers are idle, we add her to the same server in both systems. On the other hand, if x is of class 1 type, then we assign her to whichever queue (longer or shorter) we assigned her in $\tilde{J}$. Again, there may be some (possibly zero) service completions before assigning y to a queue, but since the systems are coupled just after x is assigned and any service completions are coupled, the two queues remain coupled when it is time to assign y. As before, y does not depart in R before being assigned in $\tilde{R}$. □

Let $T^{\tilde{J}}$ and $T^{\tilde{R}}$ denote the mean overall response times in $\tilde{J}$ and $\tilde{R}$ respectively. The following lemma provides an ordering of the response times.

Lemma A.2. $T^{\tilde{R}} \leq T^{\tilde{J}}$.

Proof of Lemma A.2. At any instant t, let the queue length vector in $\tilde{R}$ and $\tilde{J}$ be given by $\tilde{q}^R(t) = (r_1, r_2)$ and $\tilde{q}^J(t) = (j_1, j_2)$ respectively, where the first component of each vector represents the longer queue. We first use induction on time t to show that the states satisfy weak submajorization (as in (L1)) i.e., at any instant t, we have

$$r_1 + r_2 \leq j_1 + j_2, \tag{A.8}$$

$$r_1 \leq j_1. \tag{A.9}$$

It is easy to see that (A.8) and (A.9) are trivially true for $t = 0$. Assume that (A.8) and (A.9) are true until time t. Let the queue lengths after the next event (time t') be given by $\tilde{q}^R(t') = (r_1', r_2')$

and $\tilde{q}^J(t') = (j'_1, j'_2)$ in $\tilde{R}$ and $\tilde{J}$ respectively. We will show

$$r'_1 + r'_2 \leq j'_1 + j'_2, \tag{A.10}$$

$$r'_1 \leq j'_1. \tag{A.11}$$

First off, note that if the system is in scenario (C), there is at least one redundant job in the queue in $\tilde{R}$. Then unless the shorter queue in system $\tilde{J}$ is idle i.e., a departure causes a transition to (A), there can be no arrivals. And, even if it is idle and the next event is an arrival to the shorter queue, there can be no further arrivals i.e., $j'_2 = 1$ and the system moves back to (C). In other words, when there is a redundant customer in the *queue*, either (i) there are no jobs behind it and $j_2 \geq 0$; or (ii) there are jobs behind it (including possibly other redundant jobs) and $j_2 \in \{0, 1\}$.

Let z denote the number of redundant customers in the queue, and when $z = 1$, let $I \in \{0, 1\}$ denote whether that redundant customer is the last job ($I = 1$) in the system or not ($I = 0$) in $\tilde{R}$. If the next event is a class 1 arrival, since we couple the routing such that the arrival is sent to the shorter queue in both $\tilde{R}$ and $\tilde{J}$, or to the longer queue in both, by (L2), (A.10) and (A.11) hold. If the next event is a class 2 arrival, the job is "counted" towards the shorter queue in the $\tilde{R}$ system. Even though the class 2 customer uses the JSQ policy in the J system, the queue that she would have joined upon arrival may have become the longer queue when she is assigned in the $\tilde{J}$ system. Thus, the job may either be added to the shorter or the longer queue in $\tilde{J}$ and, by (L2) we have our result. Next, we need to consider service completions in the various possible sets of states:

1. $z = 0$. If the server at the longer (shorter) queue completes, a job departs from the longer (shorter) queue in both systems, so by (L3) we have our result.
2. $z \geq 2, j_2 = 0$ or $z = 1, I = 0, j_2 = 0$. We have $r_2 \geq 1$ since there is at least one redundant customer in the queue, and $r_1 + r_2 \leq j_1$ from (A.8) so that $r_1 \leq j_1 - 1$. The next state is either $(r_1 - 1, r_2)$ or $(r_1, r_2 - 1)$ in $\tilde{R}$ depending on the way the redundant job(s) rearrange in queue, and whether $r_1 = r_2$ or not. If the service is at the longer queue, we have state $(j_1 - 1, 0)$ in $\tilde{J}$. In this case $r'_1 + r'_2 = r_1 + r_2 - 1 \leq j_1 - 1 = j'_1 + j'_2$ from (A.8) so that (A.10) holds true. If the next state in $\tilde{R}$ is $(r_1 - 1, r_2)$, then $r'_1 = r_1 - 1 \leq r_1 \leq j_1 - 1 = j'_1$ so that (A.11) holds true. If the next state in $\tilde{R}$ is $(r_1, r_2 - 1)$, then $r'_1 = r_1 \leq j_1 - 1 = j'_1$ so that (A.11) holds true. If the service is at the shorter queue, we have state $(j_1, 0)$ in $\tilde{J}$. Since $j_1 - 1 < j_1$, we have our result.
3. $z \geq 2, j_2 = 1$ or $z = 1, I = 0, j_2 = 1$. In this case $r'_1 + r'_2 = r_1 + r_2 - 1 \leq j_1 = j'_1 + j'_2$ from (A.8) so that (A.10) holds true. There are at least 3 jobs in $\tilde{R}$, so there must be at least 2 jobs in the longer queue in $\tilde{J}$. Thus, if the service is at the longer queue, we have state $(j_1 - 1, 1)$ in $\tilde{J}$. If the next state in $\tilde{R}$ is $(r_1 - 1, r_2)$, then $r'_1 = r_1 - 1 \leq j_1 - 1 = j'_1$ from (A.9) so that (A.11) holds true. If the next state in $\tilde{R}$ is $(r_1, r_2 - 1)$, as long as there is at least one redundant customer in the queue, the servers in $\tilde{R}$ never become idle upon a service completion, so $r_2 - 1 \geq 1$. Along with $r_1 + r_2 \leq j_1 + 1$ from (A.8), this gives us $r_1 \leq j_1 - 1$

so that $r_1' = r_1 \leq j_1 - 1 = j_1'$ and (A.11) holds true. If the service is at the shorter queue, we have state $(j_1, 0)$ in $\tilde{J}$. Since $j_1 - 1 < j_1$, we have our result.

4. $z = 1$, $I = 1$ and $j_2 \geq 0$. We consider two separate sets of states in $\tilde{R}$: (I) $r_1 \neq r_2 + 1$ so that service at the longer queue results in state $(r_1 - 1, r_2)$ if $r_1 > r_2 + 1$ and $(r_1, r_2 - 1)$ if $r_1 = r_2$, and service at the shorter queue results in state $(r_1, r_2 - 1)$; and (II) $r_1 = r_2 + 1$ so that, depending on the underlying position of the redundant customer, service at the shorter queue may result either in state $(r_1, r_2 - 1)$ or $(r_1 - 1, r_2)$, whereas a service at the longer queue results in state $(r_1 - 1, r_2)$. Regardless of which state we are in, it is easy to show that (A.10) is true: We have $r_1' + r_2' = r_1 + r_2 - 1 \leq j_1 + j_2 - 1 \leq j_1' + j_2'$ from (A.8), where the final inequality is because the number of total jobs in $\tilde{J}$ after the service completion is either that at time t less one if the service completion resulted in a departure, or that at time t if the service completion was at the shorter queue and $j_2 = 0$. If we are in state (I), then the new state upon service completion is exactly the same as if there were no redundant customer, so (A.11) follows from (L3). Similarly, if we are in state (II) and the service completion is from the longer queue, we have (A.11) from (L3). However, if we are in state (II) and the service completion is from the shorter queue we may either be in state $(r_1, r_2 - 1)$ or $(r_1 - 1, r_2)$. The departure of a job from the shorter queue in $\tilde{J}$ results in state $(j_1, \max\{j_2 - 1, 0\})$. For the former state: $r_1' = r_1 \leq j_1 = j_1'$ from (A.9), and for the latter state $r_1' = r_1 - 1 \leq r_1 \leq j_1 = j_1'$ from (A.9) again, so that (A.11) is true in both cases.

Thus, the modified systems satisfy (A.8) and (A.9) $\forall t \geq 0$. Let u denote the number of unassigned customers (it is equal in both), then the total number of customers at time t is $r_1 + r_2 + u$ and $j_1 + j_2 + u$ in the $\tilde{R}$ and $\tilde{J}$ systems respectively. Then, from (A.8), the time average number of customers N^s in system $s \in \{\tilde{R}, \tilde{J}\}$ satisfies $N^{\tilde{R}} \leq N^{\tilde{J}}$. Finally, since our system is ergodic and $N^s = \lambda T^s$ by Little's Law, we have our result. □

According to Lemma A.1 the two systems J and $\tilde{J}$ (similarly, $\tilde{R}$ and R) are coupled exactly. In particular, each customer arrives and departs at the same time in the corresponding original and modified systems (although they may not be assigned a queue immediately in the modified system). As a result, the mean overall response time in J (R) is exactly equal to that in $\tilde{J}$ $(\tilde{R})$, and, by Lemma A.2, we have $T^{\tilde{R}} \leq T^{\tilde{J}}$. Thus, $T^{(\text{Ded,CUE})} \leq T^{(\text{Ded,JSQ})}$. □

Proof of Theorem 2.2. To show that class 1 customers prefer CUS over CUE for $p = 0.5$, we compute the difference (using ((A.6)-(A.1))) $T_1^{(\text{Ded,CUE})} - T_1^{(\text{Ded,CUS})} = \frac{1}{2\mu} - \frac{1}{2\mu + r\lambda}$, which is increasing in r and is zero for $r = 0$, so that $T_1^{(\text{Ded,CUE})} > T_1^{(\text{Ded,CUS})}$ for $r \in (0, 1)$.

Next, we show that their response time is decreasing in r: It is easy to see that this is true for CUS since r is in the denominator (with a positive coefficient) of a positive term in the expression of $T_1^{(\text{Ded,CUS})}$ as in (A.6). For CUE, rearranging the terms in (A.1): $T_1^{(\text{Ded,CUE})} = \frac{1}{2\mu} + \frac{1}{2\mu - \lambda} + \frac{\lambda}{(2\mu - \lambda(1-r))(2\mu + r\lambda)}$. Since r is in the denominator (with positive coefficients) of a positive term we have that $T_1^{(\text{Ded,CUE})}$ is also decreasing in r.

Finally, we show the preference for CUE over JSQ. Let $n_i, i \in \{1, 2\}$ denote the number of jobs at server i when $\pi_2 = \text{JSQ}$. Consider a tagged class 1 customer x arriving to an arbitrary state (n_1, n_2). Her wait time in either queue will be n_1/μ or n_2/μ with probability 0.5 each, and her expected wait time will be $(n_1 + n_2)/2\mu$. By PASTA the mean wait time for a class 1 customer is given by $N^{(\text{Ded, JSQ})}/2\mu$ where $N^{(\text{Ded, JSQ})}$ is the expected number of customers in this system. The total number in system when $\pi_2 = \text{CUE}$ is given by $N^{(\text{Ded, CUE})} = \lambda T^{(\text{Ded, CUE})}$ by Little's Law, and it can be easily verified that $T_1^{(\text{Ded, CUE})} = \frac{1}{\mu} + \frac{N^{(\text{Ded, CUE})}}{2\mu}$, where $T_1^{(\text{Ded, CUE})}$ and $T^{(\text{Ded, CUE})}$ are given by (A.6)-(A.1) and (A.7)-(A.3) respectively. Thus class 1 wait time under $\pi_2 \in \{\text{CUE, JSQ}\}$ is given by $N^{(\text{Ded},\pi_2)}/2\mu$, and they prefer the system with the smaller $N^{(\text{Ded},\pi_2)}$. From Little's Law, this is the one with the smaller mean overall response time i.e., when $\pi_2 = \text{CUE}$ by Theorem 2.1. □

Proof of Theorem 2.3. We are interested in the mean wait time (i.e., time in queue) $W_1^{\pi_2}$ for class 1 customers when class 2 customers follow policy π_2 and $r \to 1$. Suppose a tagged class 1 customer arrives to the system to find w_j amount of work at server j. (For every class 2 customer, only "real" work is counted i.e., one of the two copies' work will be counted at the corresponding server in CUE, and in CUS provided both copies are not in service at the same time; if both copies end up in service under CUS, then the minimum of the service times is counted in the work at both servers for this customer.) This tagged customer, upon joining queue j, has to wait until w_j amount of work is completed, so that class 1 wait time at server j is w_j, and the overall (or weighted average) class 1 wait time is $W(p, r) = pw_1 + (1 - p)w_2, \forall p, r$.

Suppose the limit of the wait time at each server, i.e., $\lim_{r \to 1^-} w_j$, exists. Then, by the sum law of limits, the limit of $W(p, r)$ also exists as $r \to 1^-$. The mean wait time $W_1^{\pi_2}$ is simply equal to $pw_1 + (1 - p)w_2$ where w_j represents the work at server j when $r = 1$. In this case the stationary probability of any state is given by simple systems such as an M/M/1 working at rate 2μ when $\pi_2 = \text{CUS}$ and an M/M/2 when $\pi_2 = \text{CUE}$, and the work w_j is independent of j since the queues are symmetric. Thus, the value of p does not matter. We will compute $W_1^{\pi_2}$ for $\pi_2 \in \{\text{JSQ, CUE, CUS}\}$ separately.

(i) $\pi_2 = \text{JSQ}$. In this case the total work is $\frac{n_1+n_2}{\mu}$ in state $n = (n_1, n_2)$. Since the expectation of n_1+n_2 with respect to the stationary JSQ distribution is the expected number N^{JSQ} in a JSQ system, and w_j equals half the total expected work in the system, we have $W_1^{\text{JSQ}} = \frac{N^{\text{JSQ}}}{2\mu}$.

(ii) $\pi_2 = \text{CUE}$. Let n denote the number of class 2 customers in the system (effectively an M/M/2 when $r = 1$). For $n = 1$, the work at a server (in expectation) is $w(1) = \frac{1}{2\mu}$. For $n > 1$, the work in either queue is the sum of the time required for $(n - 2)$ service completions at any server (so that there are no jobs in queue) and one service completion at a particular server (since there is exactly one job at each server). The former is given by $\frac{n-2}{2\mu}$ and the latter is given by $\frac{1}{\mu}$. Thus, the work at either server$w(n) = \frac{n}{2\mu}, \forall n \geq 0$. Since the expectation of n with respect to the stationary M/M/2 distribution is the expected number N^{CUE} in an M/M/2 system, $W_1^{\text{CUE}} = \frac{N^{\text{CUE}}}{2\mu}$.

(iii) $\pi_2 =$ CUS. Let n denote the number of class 2 customers in the system (effectively an M/M/1 with rate 2μ when $r = 1$). For $n \geq 0$, the work at either server $w(n)$ is the sum of the time required for n service completions at rate 2μ so that $w(n) = \frac{n}{2\mu}$. As before, $W_1^{\text{CUS}} = \frac{N^{\text{CUS}}}{2\mu}$.

Thus, $W_1^{\pi_2} = \frac{N^{\pi_2}}{2\mu}, \forall p, \pi_2 \in \{\text{CUS, CUE, JSQ}\}$. Finally, $T_1^{(\text{Ded},\pi_2)} = \frac{1}{\mu} + W_1^{\pi_2}$, and we have our result.

Since the total number in an M/M/1 serving at rate 2μ is at most that in an M/M/2 serving at rate μ per server (Wolff 1989, p. 258), we have $N^{\text{CUS}} \leq N^{\text{CUE}}$. From Foss (1989) we know that the policy that routes arriving customers to the queue with the smallest workload is optimal. Since the CUE (or M/M/2) system is equivalent to the smallest workload policy, the number of jobs in a system of two parallel queues where jobs are routed using the JSQ policy is at least as large as when they use CUE so that $N^{\text{CUE}} \leq N^{\text{JSQ}}$. Thus, we have our preference ordering. □

Proof of Theorem 2.4. For this proof (and the proofs of Theorems 2.8, 2.11 and 2.12) we will adopt the "tagged job" approach of Akgun et al. (2012) summarized as follows. First, consider a separate class of "extra" jobs which constitute a fraction ϵ of the total arrival rate.

Next, tag the first extra job in any busy period. Let K and K_e be random variables denoting the total number of new customers (of any class) and new extra customers that arrive to the system during this busy period respectively. It is easy to see that both $E[K]$ and $E[K_e]$ are finite if the length of the busy period is finite. In fact, since arriving customers belong to a class in a probabilistic fashion, the conditional probability of K_e given $K = k$ is distributed as a Binomial with parameters ϵ and k. Therefore, we have $P(K_e = 0) = \sum_{k\geq 0}(1-\epsilon)^k P(K = k)$ or $\lim_{\epsilon\to 0} P(K_e = 0) = 1$. Similarly, $P(K_e = 1) = \sum_{k\geq 1} \epsilon k(1-\epsilon)^{k-1} P(K = k)$ or $\lim_{\epsilon\to 0} \frac{P(K_e=1)}{\epsilon} = E[K] < \infty$. Finally, $E[K_e] \geq P(K_e = 1) + 2P(K_e = 2)$ by definition of expectation, $E[K_e|K = k] = \epsilon k$, and $E[K_e] = \epsilon E[K]$. Together, this gives us $0 \leq \frac{P(K_e\geq 2)}{\epsilon} \leq \frac{1}{2}\left(E[K] - \frac{P(K_e=1)}{\epsilon}\right)$, or $\lim_{\epsilon\to 0} \frac{P(K_e\geq 2)}{\epsilon} = 0$. These results can be summarized as follows:

$$P(K_e = 0) = 1 - O(\epsilon), \qquad P(K_e = 1) = O(\epsilon), \qquad P(K_e \geq 2) = o(\epsilon). \tag{A.12}$$

Furthermore, suppose we are interested in the response time T_e for extra customers, then we can condition on the number of extra customers to obtain $T_e = T_e^0(1 - O(\epsilon)) + T_e^1 O(\epsilon) + \sum_{j\geq 2} T_e^j o(\epsilon)$, where T_e^j is the response time given that j more of them arrive after the tagged (extra) job in the current busy period. This enables us to compute $\lim_{\epsilon\to 0} T_e = T_e^0$ and $\lim_{\epsilon\to 0} \frac{T_e}{\epsilon}$ etc.

For the proof of Theorem 2.4, our parameters must satisfy $\lambda < \mu$ for stability since $p = 1$ and $r \to 0^+$. We have finite busy periods even if there is only one server that serves all incoming jobs; therefore, we have finite busy periods for our present system with two servers as well $\forall \pi_2 \in \{\text{CUS, CUE, JSQ}\}$. We can use the tagged job approach as in Akgun et al. (2012). Consider the *extra class* to be the class 2 customers using policy π_2. They form a proportion $\epsilon \equiv r$ of all arrivals.

We need $T_1^{(\text{Ded, CUS})} \geq T_1^{(\text{Ded, CUE})} \geq T_1^{(\text{Ded, JSQ})}$ in a neighborhood close to $r = 0$. (For $r = 0$ we have $T_1^{(\text{Ded, CUS})} = T_1^{(\text{Ded, CUE})} = T_1^{(\text{Ded, JSQ})}$ since all three systems are identical.) In other words, the difference in the mean class 1 response times for any two adjacent policies i.e., $\Delta T_1^{\text{(CUS-CUE)}} \equiv T_1^{\text{(Ded, CUS)}} - T_1^{\text{(Ded, CUE)}}$ and $\Delta T_1^{\text{(CUE-JSQ)}} \equiv T_1^{\text{(Ded, CUE)}} - T_1^{\text{(Ded, JSQ)}}$, must be non-negative.

First, we will show that $\Delta T_1^{\text{(CUS-CUE)}} \geq 0$. We consider two systems: (i) CUS where extra customers follow the CUS policy; and (ii) CUE where extra customers follow CUE. We fix the sample path of arrivals to both systems and couple the service times at each of the servers. We consider some arbitrary state (n_1, n_2) where the sample paths for the two systems differ i.e., the next event is the arrival of the tagged class 2 customer, the first in the busy period. Using the tagged job approach (Akgun et al. 2012) and (A.12), we know that the limit of $\Delta T_1^{\text{(CUS-CUE)}}$ as $\epsilon \to 0$ is given by this value conditional upon no other extra customers arriving during this busy period. Suppose there are $n_1 \geq 0$ customers (all class 1) at server 1 when the tagged job arrives; the other server is idle i.e., $n_2 = 0$.

Under $\pi_2 =$ CUS the tagged job joins both queues and enters service at server 2 while the other copy of this job joins the queue at server 1; however, under $\pi_2 =$ CUE the job enters service at server 2 either because it is idle (if $n_1 > 0$) or since $\eta = 1$ (if $n_1 = 0$). If the two copies of the tagged job both eventually enter service, the class 1 customers (if any) joining behind the tagged job during this busy period i.e., of server 1 in CUS, will experience an increased delay in the CUS system compared to those joining in the CUE system. The expected delay is equal to $\frac{1}{2\mu}$ and occurs with non-zero probability – the probability that n_1 jobs finish at one server before 1 job at the other server – equal to $\left(\frac{1}{2}\right)^{n_1}$. Therefore, in this case $\Delta T_1^{\text{(CUS-CUE)}} \geq 0$.

Let the derivative of the difference at $r = 0$ be denoted by g, i.e., $g \equiv \left.\frac{d\Delta T_1^{\text{(CUE-JSQ)}}}{dr}\right|_{r=0}$. We will show that $g \geq 0$. We consider two systems as before: (i) CUE where extra customers follow the CUE policy; and (ii) JSQ where extra customers follow JSQ. From the definition of the derivative and using the fact that the difference is zero when $r = 0$, we have

$$g = \lim_{\epsilon \to 0^+} \frac{\Delta T_1^{\text{(CUE-JSQ)}}}{\epsilon}. \tag{A.13}$$

This can be simplified using (A.12) so that

$$g = \lim_{\epsilon \to 0^+} \frac{\Delta T_1^{\text{(CUE-JSQ,0)}} (1 - O(\epsilon)) + \Delta T_1^{\text{(CUE-JSQ,1)}} O(\epsilon) + \sum_{j \geq 2} \Delta T_1^{\text{(CUE-JSQ},j)} o(\epsilon)}{\epsilon}, \tag{A.14}$$

where $\Delta T_1^{\text{(CUE-JSQ},j)}$ is the value of the difference $\Delta T_1^{\text{(CUE-JSQ)}}$ conditioned on the number j of additional extra customers arriving during the busy period.

Again, suppose $n_1 \geq 0$ customers (all class 1) and $n_2 = 0$ when the tagged job arrives. It is easy to see that the single tagged job is indistinguishable between the two systems: They join the same server when both servers are idle and they join the idle server whenever $n_1 \geq 1$. Therefore, $\Delta T_1^{\text{(CUE-JSQ,0)}} = 0$. To find the value of $y \equiv \Delta T_1^{\text{(CUE-JSQ,1)}}$, we assume another extra job, say D, arrives to the system, which may be in one of the following states:

(a) $n_2 = 0, n_1 \geq 0$. In this case job D joins the same queue in either system and $y = 0$.

(b) $n_2 = 1, n_1 = 0$. In this case job D joins queue 2 in either system and $y = 0$.

(c) $n_2 = 1, n_1 \geq 1$. In this case job D joins both queues in the CUE system, but joins queue 2 in the JSQ system. With probability $\left(\frac{1}{2}\right)^{n_1}$ D enters service at server 1, creating a delay for any class 1 customers joining after her in the CUE system – compared to those joining in the JSQ system – within this busy period i.e., of server 1 in CUE. Therefore, $y \geq 0$.

Thus, $\Delta T_1^{(\text{CUE-JSQ},0)} = 0$ and $\Delta T_1^{(\text{CUE-JSQ},1)} \geq 0$. Additionally, $O(\epsilon) \geq 0$ since it represents the probability of having one new "extra" customer in a busy period. Together with (A.14) and $\lim_{\epsilon\to 0^+} \frac{o(\epsilon)}{\epsilon} = 0$,

$$g = \lim_{\epsilon\to 0^+} \frac{\Delta T_1^{(\text{CUE-JSQ},1)} O(\epsilon) + \sum_{j\geq 2} \Delta T_1^{(\text{CUE-JSQ},j)} o(\epsilon)}{\epsilon} \geq 0. \tag{A.15}$$

Finally, since the derivative is non-negative, the difference $\Delta T_1^{(\text{CUE-JSQ})}$ is non-decreasing in r when $r = 0$. Combined with the fact that the difference is exactly zero when $r = 0$, this implies that the value is non-negative for low enough r, or equivalently r, and JSQ is preferred to CUE. □

Proof of Theorem 2.5. As in the proof of Theorem 2.3, we will take the limit of class 1 wait time in any state as $r \to 1^-$. A key difference is that the class 1 wait time is now $W(p, r) = w$ where w is the total work at the *shorter* queue. Whenever the limit of class 1 wait time, i.e., $\lim_{r\to 1^-} w$, exists, its value will be equal to that when $r = 1$, and the distribution of the number of customers is the same as the stationary distribution of the system when $r = 1$ i.e., an M/M/1 with rate 2μ under $\pi_2 = \text{CUS}$ and an M/M/2 under $\pi_2 = \text{CUE}$. Under CUS the queue lengths are equal, and the work due to class 2 customers when there are n customers in the system is given by $w(n) = \frac{n}{2\mu}$. Thus, the expected wait time is $W_1^{\text{CUS}} = \frac{N^{\text{M/M/1}}}{2\mu} = \frac{\rho}{2\mu(1-\rho)}$ where $N^{\text{M/M/1}}$ denotes the mean number in system in an M/M/1 working at rate 2μ.

Under CUE the work due to class 2 customers depends on n: When $n \geq 2$, the queue lengths are equal, and the work is given by $w(n) = \frac{n}{2\mu}$ as in the proof of Theorem 2.3. When $n < 2$, the work is zero since at least one of the servers is idle. Thus, the expected queueing time W_1^{CUE} is given by $W_1^{\text{CUE}} = \frac{1}{2\mu}\left(N^{\text{M/M/2}} - q^{\text{M/M/2}}(1)\right)$ where $N^{\text{M/M/2}}$ and $q^{\text{M/M/2}}(n)$ denote the mean number in system and the stationary probability of n customers in system in an M/M/2. Note that $N^{\text{M/M/2}}$ is summed for $n \geq 1$ in an M/M/2; we need to subtract $q^{\text{M/M/2}}(1)$ to correct for the fact that we sum for $n \geq 2$. Since $N^{\text{M/M/2}} = \frac{2\rho}{1-\rho^2}$ and $q^{\text{M/M/2}}(1) = \frac{2\rho(1-\rho)}{1+\rho}$ we have $W_1^{\text{CUE}} = \frac{1}{2\mu}\left(\frac{2\rho}{1-\rho^2} - \frac{2\rho(1-\rho)}{1+\rho}\right) = \frac{\rho^2(2-\rho)}{\mu(1-\rho^2)}$. Since the mean system time is the sum of the mean wait time $W_1^{\pi_2}$ and mean service time $(= \frac{1}{\mu})$ we have the required expressions for the limiting value of class 1 customers' mean system time $T_1^{(\text{JSQ},\pi_2)}$.

To compare the two expressions W_1^{CUS} and W_1^{CUE} we look for sign changes in $\Delta \equiv W_1^{\text{CUE}} - W_1^{\text{CUS}}$ i.e., $\Delta = \frac{\rho^2(2-\rho)}{\mu(1-\rho^2)} - \frac{\rho}{2\mu(1-\rho)} = \frac{\rho}{2\mu(1-\rho^2)}\left(-2\rho^2 + 3\rho - 1\right)$. Since the sign of the quadratic $-2\rho^2 + 3\rho - 1$ changes at $\rho = 0.5$ we have $\Delta \geq 0$ for $\rho \geq 0.5$ and $\Delta < 0$ otherwise. Thus we have that $W_1^{\text{CUE}} > W_1^{\text{CUS}}$ for $\rho > 0.5$ so that class 1 customers prefer CUS instead of CUE, and the reverse is true for $\rho < 0.5$. Finally, $W_1^{\text{CUE}} = W_1^{\text{CUS}}$ for $\rho = 0.5$ so they are indifferent between CUS and CUE. $\square$

Proof of Proposition 2.1. We present a counterexample to show that JSQ is not always optimal. Consider the CUS state described by $\mathbf{S} = \{\mathbf{S}_1; \mathbf{S}_2\}$ where $\mathbf{S}_1 = (A, B, C_1, D_1)$ and $\mathbf{S}_2 = (C_2, D_2, E)$ i.e., C and D are redundant. Define $w_1^{(j)}$ as the wait time for an arriving job if she joins server $j \in \{1, 2\}$, and let $W_1^{(j)}$ be its mean. If the random variable $\{d_j(\mathbf{S})\}$ denotes queueing delay for an arriving class 1 job if it joins queue j when the state is $\mathbf{S}$, then

$$w_1^{(j)} \sim d_j\left(\begin{array}{c} A, B, C_1, D_1 \\ C_2, D_2, E \end{array}\right).$$

Under the independent exponential assumption, the first service event is distributed as $\exp(2\mu)$. With probability 0.5 the service completion could be at either the first or the second server, so

$$w_1^{(j)} \sim \exp(2\mu) + \frac{1}{2}d_j\left(\begin{array}{c} B, C_1, D_1 \\ C_2, D_2, E \end{array}\right) + \frac{1}{2}d_j\left(\begin{array}{c} A, B, D_1 \\ D_2, E \end{array}\right).$$

Similarly, conditioning on the remaining service events one at a time,

$$\begin{aligned} w_1^{(j)} &\sim \exp(2\mu) + \frac{1}{2}\left[\exp(2\mu) + \frac{1}{2}d_j\left(\begin{array}{c} C_1, D_1 \\ C_2, D_2, E \end{array}\right) + \frac{1}{2}d_j\left(\begin{array}{c} B, D_1 \\ D_2, E \end{array}\right)\right] \\ &\quad + \frac{1}{2}\left[\exp(2\mu) + \frac{1}{2}d_j\left(\begin{array}{c} B, D_1 \\ D_2, E \end{array}\right) + \frac{1}{2}d_j\left(\begin{array}{c} A, B \\ E \end{array}\right)\right] \\ &= 2\exp(2\mu) + \frac{1}{4}d_j\left(\begin{array}{c} C_1, D_1 \\ C_2, D_2, E \end{array}\right) + \frac{1}{2}d_j\left(\begin{array}{c} B, D_1 \\ D_2, E \end{array}\right) + \frac{1}{4}d_j\left(\begin{array}{c} A, B \\ E \end{array}\right) \\ &= \frac{11}{4}\exp(2\mu) + \frac{1}{2}d_j\left(\begin{array}{c} D_1 \\ D_2, E \end{array}\right) + \frac{1}{4}d_j\left(\begin{array}{c} B \\ E \end{array}\right) + \frac{1}{4}d_j\left(\begin{array}{c} A, B \\ E \end{array}\right). \end{aligned}$$

A class 2 customer is served at rate 2μ if both her jobs are in service at the same time, so

$$\begin{aligned} w_1^{(1)} &\sim \frac{11}{4}\exp(2\mu) + \frac{1}{2}\exp(2\mu) + \frac{1}{4}\exp(\mu) + \frac{2}{4}\exp(\mu) = \frac{13}{4}\exp(2\mu) + \frac{3}{4}\exp(\mu), \\ w_1^{(2)} &\sim \frac{11}{4}\exp(2\mu) + \frac{1}{2}\left(\exp(2\mu) + \exp(\mu)\right) + \frac{1}{4}\exp(\mu) + \frac{1}{4}\exp(\mu) = \frac{13}{4}\exp(2\mu) + \exp(\mu). \end{aligned}$$

Taking the expectation over $w_1^{(j)}$ gives $W_1^{(1)} = \frac{19}{8\mu}$ and $W_1^{(2)} = \frac{21}{8\mu}$: joining the longer queue 1 results in lower mean system time, implying that JSQ is not optimal. Similarly, $\mathbf{S}_1 = (A, B, C, E_1, F_1)$ and $\mathbf{S}_2 = (D, E_2, F_2, G)$ results in $W_1^{(1)} = \frac{27}{8\mu}$ and $W_1^{(2)} = \frac{29}{8\mu}$: a counterexample for CUE. $\square$

Define the random variable $\bar{w}_1^{(j)}(s) \triangleq \{w_1^{(j)} | \mathbf{S} = s\}$ as queueing delay given that an arriving class 1 customer joins queue j when the *FO* state realized is s. Then the mean queueing delay for a given

PO state if an arriving job joins queue j can be defined as the expectation of the *FO* queueing delay over all possible realizations of the *FO* state: $\tilde{W}^{(j)} = E_{\mathbf{S}}\left[E[\bar{w}_1^{(j)}(S)]\right]$.

Proof of Theorem 2.6. We prove Theorem 2.6 with the aid of the following lemma.

Lemma A.3. *Joining queue 2 is at least as good as joining queue 1 for all "boundary" states of the form* $(i, 0)$ *and* (i, i)*,* $\forall i \geq 0$*.*

Proof of Lemma A.3. For states of the form $s = (i, 0)$ the only possible *FO* state realization is when all i customers belong to class 1, so the queueing delay $\bar{w}_1^{(j)}(s)$ for an arriving class 1 customer at queue j is distributed as: $\bar{w}_1^{(1)}(s) \sim i \cdot \exp(\mu), \bar{w}_1^{(2)}(s) = 0$. Therefore, JSQ is optimal for $(i, 0)$, $\forall i \geq 0$. For states of the form (i, i), by symmetry, the mean queueing delay is the same in either queue. □

Regardless of which queue the arriving customer joins, future arrivals do not impact her response time. We can therefore shut off arrivals and focus on the system in state $n = (n_1, n_2), n_1 > n_2$; when $n_2 = n_1$, JSQ is optimal from Lemma A.3. For our proof, we will consider a single departure event and use a two-step induction, on n_1 and $0 \leq n_2 < n_1$, to show that JSQ is optimal in all resulting states. Clearly the theorem is true for $(1, 0)$ from Lemma A.3. Assume the theorem is true for $n_1 = 1, 2, \ldots, i$ and $n_2 < n_1$ (induction hypothesis I). We need to show that it is optimal for $n_1 = i + 1$ and $n_2 < i + 1$. First, the theorem is true for $n_1 = i + 1$ and $n_2 = 0$ by Lemma A.3. Now, assume it is true for $n_2 \leq k$, $0 \leq k < i$ (induction hypothesis II): we need to prove that JSQ is optimal for $n_2 = k + 1$. Thus, our state is $(i + 1, k + 1)$. Consider the departure events that could occur at state $(i + 1, k + 1)$ for the CUS (CUE) model:

1. A class 1 job departs from queue 1 (A job departs from queue 1 and the next job is not class 2), resulting in state $(i, k + 1)$: JSQ is optimal for $(i, k + 1)$ by induction hypothesis I if $k + 1 < i$, and by Lemma A.3 if $k + 1 = i$.
2. A class 1 job departs from queue 2 (A job departs from queue 2 and the next job is not class 2), resulting in state $(i + 1, k)$: JSQ is optimal for $(i + 1, k)$ by induction hypothesis II.
3. A class 2 job departs (A job departs from either queue and the next job is class 2), resulting in state (i, k): JSQ is optimal for (i, k) by induction hypothesis I since $k < i$.

So joining queue 2 was optimal upon the job's arrival in all three cases. Hence, JSQ is optimal for the state $(i + 1, k + 1)$ under CUS and CUE, and, by induction, the proof is complete. □

A.2 Additional results

A.2.1 Derivation of performance measures under CUE

Consider the system where $\pi_1 = \text{Ded}$. We define two separate classes of class 1 customers – class A with arrival rate λ_A is dedicated to server 1 and class B with arrival rate λ_B is dedicated to server 2.

Let class R denote the redundant customers with $\pi_2 = \text{CUE}$ and λ_R denote their arrival rate. Let the service rates at the two servers be μ_1 and μ_2. Define η as the probability that an incoming class R customer who sees both servers idle chooses server 1. We use W_i to denote the mean wait time for class i customers. Then the following lemma provides the performance measures of interest.

Lemma A.4. *When $\pi_1 = Ded, \pi_2 = CUE$ and $\eta = \frac{\lambda_B+\lambda_R}{\lambda_A+\lambda_B+2\lambda_R}$, the wait time W_i for class i customers is given by:*

$$\begin{aligned} W_A &= \frac{1}{\mu_1-\lambda_A}(g_1+g_{21}) + \frac{1}{\mu_1+\mu_2-\lambda}(g_{12}+g_{21}), && \text{(A.16)}\\ W_B &= \frac{1}{\mu_2-\lambda_B}(g_2+g_{12}) + \frac{1}{\mu_1+\mu_2-\lambda}(g_{12}+g_{21}), && \text{(A.17)}\\ W_R &= \frac{1}{\mu_1+\mu_2-\lambda}(g_{12}+g_{21}), && \text{(A.18)} \end{aligned}$$

where

$$\begin{aligned} g_1 &= \frac{\lambda_A+\eta\lambda_R}{\mu_1-\lambda_A}g_0, && \text{(A.19)}\\ g_2 &= \frac{\lambda_B+(1-\eta)\lambda_R}{\mu_2-\lambda_B}g_0, && \text{(A.20)}\\ g_{12} &= \left[\frac{\lambda_A+\lambda_R}{\mu_1+\mu_2-\lambda}\cdot\frac{\lambda_B+(1-\eta)\lambda_R}{\mu_2-\lambda_B}\right]g_0, && \text{(A.21)}\\ g_{21} &= \left[\frac{\lambda_B+\lambda_R}{\mu_1+\mu_2-\lambda}\cdot\frac{\lambda_A+\eta\lambda_R}{\mu_1-\lambda_A}\right]g_0, && \text{(A.22)}\\ g_1+g_2+g_{12}+g_{21}+g_0 &= 1. && \text{(A.23)} \end{aligned}$$

Proof. We use Visschers et al. (2012) to derive the required performance measures. We will use the same notation as in Visschers et al. (2012). Let m_i denote server i. First, we derive the value of η that satisfies the "assignment condition" (as in Visschers et al. 2012): η solves the equation $(\lambda_A+\eta\lambda_R)(\lambda_B+\lambda_R) = (\lambda_B+(1-\eta)\lambda_R)(\lambda_A+\lambda_R)$. The solution is unique and is stated in Lemma A.4. Our system may belong to one of five types of states: 1) $\pi(\cdot, m_2, \cdot, m_1)$; 2) $\pi(\cdot, m_1, \cdot, m_2)$; 3) $\pi(\cdot, m_1)$; 4) $\pi(\cdot, m_2)$; and 5) the empty state. The probability of the first four states can be computed in terms of the probability, say g_0, of being in the empty state using equation (32) in Visschers et al. (2012), and they are given by (A.19)-(A.22).

Let D_i denote the wait time for class i. Then, we can use Theorem 3 in Visschers et al. (2012) to derive the Laplace-Stieltjes transform (LST) of D_i i.e., $L^{(i)}(s) = E[e^{-sD_i}]$:

$$\begin{aligned} E[e^{-sD_R}] &= g_0+g_1+g_2+(g_{12}+g_{21})\frac{\mu_1+\mu_2-\lambda}{\mu_1+\mu_2-\lambda+s}, && \text{(A.24)}\\ E[e^{-sD_A}] &= g_0+g_2+g_1\frac{\mu_1-\lambda_A}{\mu_1-\lambda_A+s} + \frac{\mu_1+\mu_2-\lambda}{\mu_1+\mu_2-\lambda+s}\left(g_{12}+g_{21}\frac{\mu_1-\lambda_A}{\mu_1-\lambda_A+s}\right), && \text{(A.25)}\\ E[e^{-sD_B}] &= g_0+g_1+g_2\frac{\mu_2-\lambda_B}{\mu_2-\lambda_B+s} + \frac{\mu_1+\mu_2-\lambda}{\mu_1+\mu_2-\lambda+s}\left(g_{21}+g_{12}\frac{\mu_2-\lambda_B}{\mu_2-\lambda_B+s}\right). && \text{(A.26)} \end{aligned}$$

The expected wait time W_i for class i can be computed from the LST $L^{(i)}(s)$ using the relationship: $W_i = -\frac{dL^{(i)}(s)}{ds}\Big|_{s=0}$. Thus, we have the expressions (A.16)-(A.18).

□

Lemma A.5. *When $p = 0.5$ and $\mu_1 = \mu_2 = \mu$, the mean system time T_i for class i customers is given by:*

$$\begin{aligned} T_A = T_B &= \frac{1}{2\mu} + \frac{1}{2\mu - \lambda} + \frac{1}{2\mu - \lambda(1-r)} - \frac{1}{2\mu + r\lambda}, \\ T_R &= \frac{1}{2\mu - \lambda} + \frac{1}{r(2\mu + r\lambda)} - \frac{1-r}{2r\mu}, \end{aligned}$$

where r denotes the fraction of redundant customers.

Proof. When $p = 0.5$ we have $\eta = 0.5$ since $\lambda_A = \lambda_B$. This gives us:

$$g_1 = g_2 = \frac{\lambda}{2\mu - \lambda(1-r)} g_0, \tag{A.27}$$

$$g_{12} = g_{21} = \left[\frac{\lambda(1-r)/2 + \lambda r}{2\mu - \lambda} \cdot \frac{\lambda}{2\mu - \lambda(1-r)}\right] g_0. \tag{A.28}$$

Using (A.23) and Mathematica, we can obtain the values of g_1, g_0 and g_{12} in terms of our parameters. Then, we can use (A.16)-(A.18) we get the required wait time expressions. Finally, the system time T_i for class i is simply $\frac{1}{\mu} + W_i$. □

A.2.2 Modifications to the CTMC in §2.5

Given a state $\mathbf{S} = \{\mathbf{S}_1; \mathbf{S}_2\}$, let $n_i(\mathbf{S})$ denote the length of the vector $\mathbf{S}_i, i \in \{1, 2\}$, and let $\mathbf{J} \circ \{k\}^x$ be the operator that adds x copies of job type k to state $\mathbf{J}$ according to action a. The modified transitions $\tilde{p}\{\cdot|a\}$ relative to those of the original CTMC $p\{\cdot|a\}$ given policy $a \in \{1, 2\}$ are:

1. $\tilde{p}\left\{(\mathbf{S}, m = 0), (\mathbf{S}_a^j, m = 0)|a\right\} = p\left\{\mathbf{S}, \mathbf{S}_a^j|a\right\}, \forall j, a \in \{1, 2\}, n_1(\mathbf{S}) < k$, or $a = 2, n_1(\mathbf{S}) = k, n_2(\mathbf{S}) < k$. Arrivals when the *FO* state is not full are the same as before.
2. $\tilde{p}\{(\mathbf{S}, m = 0), (\mathbf{S}, m = 1)|a = 1\} = \lambda$ if $n_1(\mathbf{S}) = k$. Arrivals when the longer *FO* state is full are added to the central queue if $a = 1$.
3. $\tilde{p}\{(\mathbf{S}, m = 0), (\mathbf{S}, m = 1)|a = 2\} = \lambda \mathbf{I}_{n_2(\mathbf{S})=k} + \lambda r \mathbf{I}_{n_2(\mathbf{S})<k}$. Class 1 arrivals when the shorter *FO* queue is full are added to the central queue if $a = 2$; class 2 is added if the longer *FO* queue is full.
4. $\tilde{p}\{(\mathbf{S}, m), (\mathbf{S}, m + 1)|a\} = \lambda$ if $0 < m < M$. Arrivals when the central queue is not empty are added to the central queue.
5. $\tilde{p}\{(\mathbf{S}, m = 0), (\mathbf{S}_d^j, m = 0)|a\} = p\{\mathbf{S}, \mathbf{S}_d^j|a\}, \forall j, a \in \{1, 2\}$. Transitions due to service completions when the central queue is empty are the same as before.

6. $\tilde{p}\{(\mathbf{S}, m), (\mathbf{S}_d^j, m)|a\} = p\{\mathbf{S}, \mathbf{S}_d^j|a\}, \forall a \in \{1, 2\}, m > 0$ if $n_1(\mathbf{S}_d^j) = k$ and $n_2(\mathbf{S}_d^j) > 0$. Transitions when a service completion does not lead to an idle server and the longer queue remains full are the same as before.
7. $\tilde{p}\{(\mathbf{S}, m), (\mathbf{S}_d^j \circ \{2\}, m-1)|a\} = p\{\mathbf{S}, \mathbf{S}_d^j|a\}r, \forall a \in \{1, 2\}, m > 0$ if $n_1(\mathbf{S}_d^j) < k$. When a service completion leads to an empty slot in the longer *FO* queue, a class 2 customer (probability r) from the central queue enters the *FO* state.
8. $\tilde{p}\{(\mathbf{S}, m), (\mathbf{S}_d^j \circ \{1\}, m-1)|a = 1\} = p\{\mathbf{S}, \mathbf{S}_d^j|a = 1\}(1-r), m > 0$ if $n_1(\mathbf{S}_d^j) < k$. When a service completion leads to an empty slot in the longer *FO* queue, a class 1 customer (probability $1 - r$) enters the empty slot from the central queue if $a = 1$.
9. $\tilde{p}\{(\mathbf{S}, m), ((\mathbf{S}_1^{'} \backslash \{1\}; \mathbf{S}_2^{'}) \circ \{2\}, m - 1)|a\} = p\left\{\mathbf{S}, \mathbf{S}_d^j|a\right\} r, \forall a \in \{1, 2\}, m > 0$ if $n_1(\mathbf{S}_d^j) = k$ and $n_2(\mathbf{S}_d^j) = 0$ where $\mathbf{S}^{'} = \mathbf{S}_d^j$. When a service completion leads to an idle server and the longer queue is full, a class 2 customer (probability r) from the central queue enters the *FO* queue by replacing a class 1 customer in the longer queue. This is to ensure that the response time is a lower bound.
10. $\tilde{p}\{(\mathbf{S}, m), (\mathbf{S}_d^j, m)|a = 1\} = p\{\mathbf{S}, \mathbf{S}_d^j|a = 1\}(1 - r), m > 0$ if $n_1(\mathbf{S}_d^j) = k$ and $n_2(\mathbf{S}_d^j) = 0$. When a service completion leads to an idle server and the longer queue is full, a class 1 customer (probability $1 - r$) does not enter the *FO* state if $a = 1$.
11. $\tilde{p}\{(\mathbf{S}, m), (\mathbf{S}_d^j \circ \{1\}^x \circ \{2\}, m - x - 1)|a = 2\} = p\{\mathbf{S}, \mathbf{S}_d^j|a = 2\}(1 - r)^x r, m > 0$ if $n_1(\mathbf{S}_d^j) < k$ and $1 \leq x \leq min\{m - 1, -n_2(\mathbf{S}_d^j) + k - 1\}$. When a service completion leads to an empty slot in the longer *FO* queue, several class 1 customers (probability $(1 - r)^x$) may enter the *FO* queue from the central queue followed by a class 2 customer (probability r) (either until the central queue is empty or until no slots are left in the *FO* queue) if $a = 2$.
12. $\tilde{p}\{(\mathbf{S}, m), (\mathbf{S}_d^j \circ \{1\}^x, m - x)|a = 2\} = p\{\mathbf{S}, \mathbf{S}_d^j|a = 2\}(1 - r)^x, m > 0$ if $n_1(\mathbf{S}_d^j) < k$ and $x = min\{m, -n_2(\mathbf{S}_d^j) + k\}$. When a service completion leads to an empty slot in the longer *FO* queue, as many class 1 customers (probability $(1-r)^x$) as possible (either until the central queue is empty or no slots are left in the *FO* queue) enter the *FO* queue (according to policy $a = 2$) from the central queue if $a = 2$.
13. $\tilde{p}\{(\mathbf{S}, m), (\mathbf{S}_d^j \circ \{1\}, m - 1)|a = 2\} = p\left\{\mathbf{S}, \mathbf{S}_d^j|a = 2\right\}(1 - r), m > 0$ if $n_1(\mathbf{S}_d^j) = k$ and $n_2(\mathbf{S}_d^j) = 0$. When a service completion leads to an idle server, a class 1 customer enters service if $a = 2$.

A.2.3 Bounds on value of information

Consider the state described by $\mathbf{S} = \{\mathbf{S}_1; \mathbf{S}_2\}$ where $\mathbf{S}_1 = (2, \ldots, 2, 1, \ldots, 1)$ and $\mathbf{S}_2 = (1, \ldots, 1, 2, \ldots, 2)$ where the number of class 2 customers is y, and the number of class 1 customers in queue 1 (queue 2) is z (x).

(i) $\boldsymbol{y \rightarrow \infty}$ **and** $\boldsymbol{z < x}$. Let $W_1(x, z)$ and $W_2(x, z)$ denote the expected waiting time for an arriving class 1 customer upon joining queue 1 and 2 respectively (subscript y is omitted since

$y \to \infty$). By conditioning on a single service completion, we have $W_j(x,z) = \frac{1}{2\mu} + 0.5W_j(x-1,z) + 0.5W_j(x,z), j \in \{1,2\}$ (as $y \to \infty$), which simplifies to yield a recursive equation in x: $W_j(x,z) = \frac{1}{\mu} + W_j(x-1,z), x \geq 1$. Therefore, $W_j(x,z) = \frac{x}{\mu} + W_j(0,z), \forall x \geq 1, j \in \{1,2\}$. Since $W_1(0,z) = \frac{y}{2\mu} + \frac{z}{\mu}$ and $W_2(0,z) = \frac{y}{2\mu}$, we have $W_1(x,z) = \frac{x}{\mu} + \frac{y}{2\mu} + \frac{z}{\mu}$ and $W_2(x,z) = \frac{x}{\mu} + \frac{y}{2\mu}$. The ratio of the queueing time (from joining the shorter queue 1 relative to the optimal policy i.e., joining queue 2) $\delta \equiv \frac{W_1(x,z)}{W_2(x,z)} = \frac{x+y/2+z}{x+y/2}$ represents a measure of the value of information. If $z = x-1$ and x scales with y then the ratio approaches $\frac{5}{3}$, however, the absolute difference, $W_1(x,z) - W_2(x,z) = \frac{z}{\mu}$, could be arbitrarily unbounded.

(ii) $\boldsymbol{x = z = 1}$. Then, the queueing times (as defined in (i) above) are: $W_1 = T + \frac{\hat{Y}}{2\mu} + \frac{z}{\mu}$ and $W_2 = T + \frac{\hat{Y}}{2\mu} + \frac{\hat{X}}{\mu}$ where T denotes the minimum of the time until x jobs are served at server 2 or until y jobs are served at server 1, and Y (X) denotes the number of redundant customers (non-redundant customers at server 2) left behind at time T. Let the expected values of T, X and Y be denoted by $\hat{T}, \hat{X}$ and $\hat{Y}$ respectively. Since $x = z = 1$, by conditioning on the different service completions possible, we have the following distribution $T \sim \frac{n}{2\mu}$ with probability (w.p.) $2^{-n}, \forall n \in \{1, y\}$ and $T \sim \frac{y}{2\mu}$ w.p. 2^{-y} so that $\hat{T} = \frac{1}{\mu}(1-2^{-y})$. Similarly, the number of redundant customers left at time T is $Y \sim (y-n+1)$ w.p. $2^{-n}, \forall n \in \{1, y\}$ and $Y \sim 0$ w.p. 2^{-y}, and the number of non-redundant customers at server 2 is $X \sim 1$ w.p. 2^{-y} and $X \sim 0$ w.p. $(1-2^{-y})$, so that $\hat{Y} = y - 1 + 2^{-y}$ and $\hat{X} = 2^{-y}$. Thus, $W_1 = \frac{y+3-2^{-y}}{2\mu}$ and $W_2 = \frac{y+1+2^{-y}}{2\mu}$. The ratio $\delta = \frac{W_1}{W_2} = \frac{y+3-2^{-y}}{y+1+2^{-y}}$ achieves its maximum value $\delta = \frac{19}{13}$ at $y = 2$, and the absolute difference $\frac{1-2^{-y}}{\mu}$ achieves its maximum value $\frac{1}{\mu}$ as $y \to \infty$.

A.3 Proofs of extensions

Proof of Theorem 2.7. We are interested in class 1 wait time $W_1^{\pi_2}$ when class 2 customers follow policy π_2 as $r \to 1^-$. Similar to the proof of Theorem 2.3 in Appendix A.1, the limiting value exists under the condition stated in the theorem and the limiting value will equal that when $r = 1$ and independent of p_j. Moreover, the stationary probability of being in a particular state when $r = 1$ is given by simple systems. We will compute $W_1^{\pi_2}$ – considering the work at a server – for each case separately (similar to the proof of Theorem 2.3):

(i) $\pi_2 =$ JSQ. The total work $w(n)$ at server $j \in \{1, \ldots, m\}$ in state $n = (n_1, \ldots, n_m)$ is n_j/μ, so that the expected wait time $W_1^{\text{JSQ}} = N^{\text{JSQ}}/m\mu$.

(ii) $\pi_2 =$ CUE. If the number i of class 2 customers in the system is such that $0 < i \leq m$, then with probability i/m the work at a server is $1/\mu$ and 0 otherwise, so that the expected work at a server is $w(i) = i/m\mu$. For states with $i > m$ the work is the sum of the time for $(i-m)$ service completions at any server $(= (i-m)/m\mu)$ and one service completion at a particular server $(= 1/\mu)$. Thus the wait time is $W_1^{\text{CUE}} = N^{\text{CUE}}/m\mu$.

(iii) $\pi_2 =$ CUS. The work at any server when there are n customers is the time for n service completions at any of the m servers so that $w(n) = n/m\mu, \forall n \geq 0$ and the expected wait time is $W_1^{\text{CUS}} = N^{\text{CUS}}/m\mu$.

Thus, $W_1^{\pi_2} = \frac{N^{\pi_2}}{m\mu}, \forall p, \pi_2 \in \{\text{CUS, CUE, JSQ}\}$. Since $N^{\text{CUS}} \leq N^{\text{CUE}} \leq N^{\text{JSQ}}$ (Foss 1989, Wolff 1989) we have the required result. □

Proof of Theorem 2.8. Again, since $\lambda < \mu$ we have finite busy periods. We will show that $\Delta T_1^{(\text{CUS-JSQ})} \geq 0$. We consider two systems as in the proof of Theorem 2.4 in Appendix A.1 – the CUS and the JSQ systems – distinguished by the policy used by class 2, and focus on this customers' impact provided no other extra customers arrive during this busy period. Let $n_1 \geq 0$ denote the number of customers at server 1 when this tagged job arrives; all other servers are idle. The tagged job joins a queue 2-s in the JSQ system but joins all queues in the CUS system. Since there may be a delay to any class 1 customers joining after this job in the current busy period i.e., of server 1 in CUS, and this delay occurs with non-zero probability (if the tagged job reaches service at server 1 as well), class 1 customers prefer JSQ to CUS. The policy of the tagged customer under CUE is identical to that under JSQ so that they are indifferent between JSQ and CUE. □

Proof of Theorem 2.9. First, we show that JSQ is not always optimal under the *FO* regime. Consider $\pi_2 = \text{CUS}$, $s = m$, and state $\mathbf{S} = \{\mathbf{S}_1; \mathbf{S}_2 \ \ldots; \mathbf{S}_m\}$, where $\mathbf{S}_1 = (C_1, D_1, E)$ and $\mathbf{S}_j = (A^{(j)}, B^{(j)}, C_j, D_j), j \in \{2, \ldots, m\}$ such that $A^{(j)}$ and C_1 are currently in service, $A^{(j)}$, $B^{(j)}$ and E belong to class 1, and C and D are redundant. Define $W_1^{(j)}$ as class 1 wait time upon joining queue j. Then, for $m = 5, \mu = 1, W_1^{(1)} = 2.2234$ and $W_1^{(j)} = 2.1941, \forall j \in \{2, \ldots, m\}$, showing that JSQ is not optimal in this state.

Next, we show that JSQ is optimal under the *PO* regime if $s = m$. This proof is on the same lines as in the case of two servers in the proof of Theorem 2.6. We first modify Lemma A.3 to show the optimality of JSQ in "boundary" states of the form $(i_0, i_0, \ldots, i_0), i_0 \geq 0$ and $(i_1, i_2, \ldots, 0, 0), \forall i_j > 0, 1 \leq j \leq k, \forall k < m$. (We assert this without proof, as it is straightforward.) Then we will condition on a single departure event at any state n and use an m-step induction on $n_j, j \in \{1, \ldots, m\}$ to show that JSQ is optimal in all resulting states (reordering states as necessary so that the queue lengths are non-increasing in j). As before, our final induction step is to show that JSQ is optimal for state $i = (i_1 + 1, i_2 + 1, \ldots, i_m + 1)$ given that the induction hypothesis at stage n_j is true, i.e., JSQ is optimal for state $i = (i_1 + 1, i_2 + 1, \ldots, i_{j-1} + 1, n_j, n_{j+1}, \ldots, n_m)$ $\forall n_j \leq i_j, \forall n_k \leq n_{k-1}, j < k \leq m$. The base case at stage n_j is that JSQ is optimal for state $i = (i_1 + 1, i_2 + 1, \ldots, i_{j-1} + 1, 0, \ldots, 0, 0)$ which is true from the Lemma. The proof follows by conditioning for CUS (CUE):

1. A class 1 job departs from queue $j < m$ (a job departs from queue $j < m$ such that the next job is not of class 2), resulting in state $i = (i_1 + 1, i_2 + 1, \ldots, i_j, \ldots, i_m + 1)$ i.e., the k-th index of state i is i_j where $k = j + \mathbf{I}_{i_j = i_{j+1}}$: JSQ is optimal by inductive hypothesis at stage n_k if $i_j \neq 0$, and by the Lemma if $i_j = 0$.
2. A class 1 job departs from queue m (a job departs from queue m such that the next job is

not of class 2), resulting in state $i = (i_1 + 1, i_2 + 1, \ldots, i_m)$: JSQ is optimal by inductive hypothesis at stage n_m if $i_m \neq 0$, and by the Lemma if $i_m = 0$.

3. A class 2 job departs (a job departs from any queue such that the next job is of class 2), resulting in state $i = (i_1, i_2, \ldots, i_m)$: JSQ is optimal by inductive hypothesis at stage n_1.

   Finally, to see why JSQ is not optimal when $s < m$ under the *PO* regime, consider $m = 3, s = 2$ and state $n = (1, 1, 1)$ under CUS. There could be two underlying *FO* states, namely $\mathbf{S}^{(1)} = \{A_1; A_2; B\}$ and $\mathbf{S}^{(2)} = \{C; D; E\}$ where A is a class 2 customer who is served at rate 2μ while B, C, D and E are class 1 customers. The arriving class 1 customer prefers to join queue 1 or 2 over queue 3 because the queueing time at queue 1 or 2 lies in $[\frac{1}{2\mu}, \frac{1}{\mu}]$ depending on the relative probability of states $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$, whereas at queue 3 it is $\frac{1}{\mu}$. If JSQ were the optimal policy the customer would be indifferent between joining any of the queues.

□

Proof of Theorem 2.10. The proof is similar to the case without abandonments (proof of Theorem 2.1 in Appendix A.1) with three main differences: (1) We replace a service event with service events or abandonments, both of which are coupled between the modified and original systems; (2) Abandonments in the modified systems may occur before a job is assigned a queue; and (3) Lemma A.2 will be replaced by Lemma A.6. To illustrate (1), the scenarios and sequencing rules that constitute the modified systems are updated as follows:

(A) Any queue (in $\tilde{R}$ or $\tilde{J}$) is idle. In this case advance both systems to the next event (service, abandonment or arrival).

(B) No queue (in $\tilde{R}$ or $\tilde{J}$) is idle and there is no redundant customer in the *queue* in the $\tilde{R}$ system. In this case advance both systems to the next event (service, abandonment or arrival).

(C) No queue (in $\tilde{R}$ or $\tilde{J}$) is idle and there is at least one redundant customer in the *queue* in the $\tilde{R}$ system. In this case advance both systems to the next *service or abandonment* event (whichever is first), deferring arrival events as necessary.

Similarly, in the proof of Lemma A.1, we can consider service and abandonment events as coupled as well. The additional difference is that a job may abandon in J, but in this case, the job abandons in $\tilde{J}$ as well (either before or after being assigned a queue).

Let $A^{\tilde{J}}$ and $A^{\tilde{R}}$ denote the mean number of abandonments in $\tilde{J}$ and $\tilde{R}$ respectively. The following lemma provides an ordering of the response times – $T^{\tilde{R}}$ and $T^{\tilde{J}}$ – and the number of abandonments.

Lemma A.6. *$T^{\tilde{R}} \leq T^{\tilde{J}}$ and $A^{\tilde{R}} \leq A^{\tilde{J}}$.*

Proof. We use the same notation, and adopt a similar approach, as in the proof of Lemma A.2. Let a_r and a_j represent the total number of abandonments upto time t, and let a'_r and a'_j represent the same upto time t'. Assuming that $a_r \leq a_j$, we will show that

$$a'_r \leq a'_j, \tag{A.29}$$

in addition to (A.10) and (A.11).

Let us consider the different events that occur at time t'. We couple the arrival and service completions as in the proof of Lemma A.2. This ensures that if the next event is either an arrival or a service completion, (A.10) and (A.11) are true as before, and (A.29) is true by our induction assumption.

Next we consider the case where the next event is an abandonment. If any of the unassigned customers in either system abandon, then the two modified systems are coupled exactly and the same customer leaves from both systems. This increases the number of abandonments in $\tilde{J}$ and $\tilde{R}$ by one, so that (A.29) is true. This event does not affect the queue lengths, and the unassigned customers remain coupled, so (A.10) and (A.11) are true.

If the next event was an abandonment from one of the two queues, we couple them as follows. (Given that $\tilde{J}$ has at least as many total customers as $\tilde{R}$ due to (A.8), we consider $\tilde{J}$ to be the one that "drives" this coupling.) We couple each abandonment from $\tilde{J}$ with an abandonment in $\tilde{R}$ (as much as possible) in such a way that the orderings – (A.10), (A.11) and (A.29) – are preserved.

More rigorously, we start by numbering the customers in both systems. First, label those in queue 1 in both systems from 1 to r_1. Next, label from r_1+1 to $s_2 \equiv \min\{j_2, r_2\} + r_1$. Define $M_{LL} = \{1, r_1\}$ and $M_{SS} = \{r_1+1, s_2\}$. There may be some unlabeled customers at queue 1 in $\tilde{J}$ (if $r_1 < j_1$) and at queue 2 in either $\tilde{J}$ (if $r_2 < j_2$) or $\tilde{R}$ (if $r_2 > j_2$). Now we consider two cases: (a) $r_2 > j_2$ and (b) $r_2 \le j_2$.

Under (a) there are $r_2 - j_2$ left at queue 2 in $\tilde{R}$ and $j_1 - r_1$ left at queue 1 in $\tilde{J}$. From (A.9), we know that $j_1 - r_1 \ge r_2 - j_2$, so label the unlabeled customers at queue 2 in $\tilde{R}$ and at queue 1 in $\tilde{J}$ from s_2+1 to $s_2+r_2-j_2$. The remaining customers at queue 1 in $\tilde{J}$ will be labeled $s_2+r_2-j_2+1$ to $s_2+j_1-r_1$ if $j_1 > r_1$ so that all customers are labeled. Define $M_{LS} = \{r_1+j_2+1, r_1+r_2\}$ and $M_L = \{r_1+r_2+1, j_1+\min\{j_2, r_2\}\}$.

Under (b) there are $j_1 - r_1$ left at queue 1 and $j_2 - r_2$ left at queue 2 in $\tilde{J}$. In this case all unlabeled customers are at $\tilde{J}$ so we label the ones at queue 1 from s_2+1 to $s_2+j_1-r_1$ if $j_1 > r_1$ (again, define M_L as before) and those at queue 2 from $s_2+j_1-r_1+1$ to $s_2+j_1-r_1+j_2-r_2$, so that all customers are labeled. Define $M_S = \{r_2+j_1+1, j_1+j_2\}$.

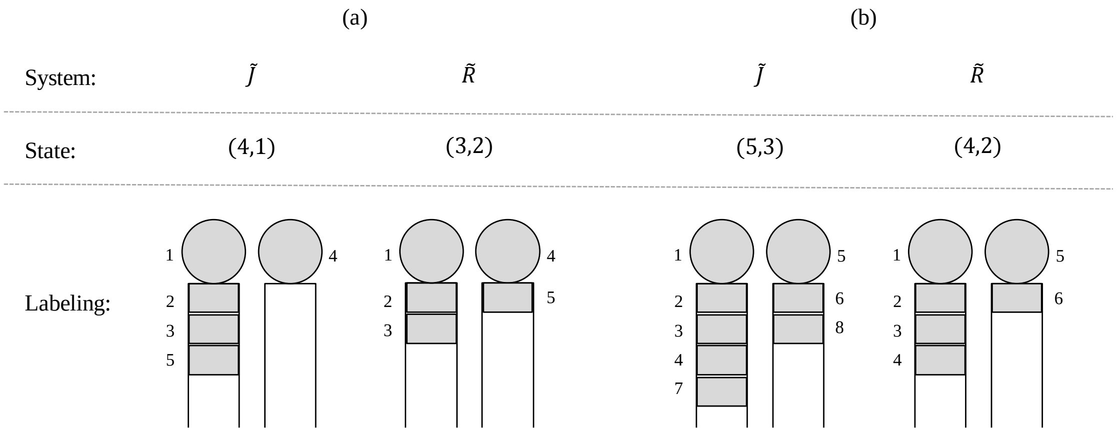
Note that $M_{xy}, x \in \{L,S\}, y \in \{L,S\}$ indicates that the abandonment is from the longer (shorter) if $x = L$ ($x = S$) queue in $\tilde{J}$ and from the longer (shorter) if $y = L$ ($y = S$) queue in $\tilde{R}$. Similarly, $M_z, z \in \{L,S\}$ indicates that it is from the longer (shorter) if $z = L$ ($z = S$) queue in $\tilde{J}$ and there is no abandonment in $\tilde{R}$. Thus, we have the following mutually exclusive sets (and labels) under the two cases:

(a) M_{LL} (from 1 to r_1), M_{SS} (from r_1+1 to r_1+j_2), M_{LS} (from r_1+j_2+1 to r_1+r_2), and M_L (from r_1+r_2+1 to j_1+j_2).

(b) M_{LL} (from 1 to r_1), M_{SS} (from r_1+1 to r_1+r_2), M_L (from r_1+r_2+1 to j_1+r_2), and M_S (from j_1+r_2+1 to j_1+j_2).

Figure A.2 shows an example for each case.

Figure A.2: Sample states and labeling.



Suppose the abandonment is of customer h under the above numbering. If $h \leq r_1 + r_2$ then she departs in both systems, $h \in M_{LL} \cup M_{SS} \cup M_{LS}$ and $a'_r = a_r + 1 \leq a_j + 1 = a'_j$, and if $r_1 + r_2 < h \leq j_1 + j_2$ then she departs only in $\tilde{J}$, $h \in M_L \cup M_S$ and $a'_r = a_r < a_j + 1 = a'_j$. Thus, (A.29) is true. To show (A.10) and (A.11), we consider the following cases:

(i) $h \in M_S$. This only occurs if $r_1 + r_2 < j_1 + j_2$ and the abandonment is from the shorter queue in $\tilde{J}$ which preserves $r_1 \leq j_1$. Furthermore, $r'_1 + r'_2 = r_1 + r_2 \leq j_1 + j_2 - 1 = j'_1 + j'_2$ so that (A.10) is true.

(ii) $h \in M_L$. Again, this only occurs if $r_1 + r_2 < j_1 + j_2$, and by definition of M_L, this case exists only if $r_1 < j_1$. Since the abandonment is from the longer queue in $\tilde{J}$ and $r_1 < j_1$, this implies $j'_1 = \max\{j_1 - 1, j_2\} \geq j_1 - 1 \geq r_1 = r'_1$. Furthermore, $r'_1 + r'_2 = r_1 + r_2 \leq j_1 + j_2 - 1 = j'_1 + j'_2$ so that (A.10) is true.

(iii) $h \in M_{SS} \cup M_{LL}$. In this case the abandonment is from both systems, and the state evolution is identical to if there was a coupled service completion in both systems either at the shorter queue (if $h \in M_{SS}$) or at the longer queue (if $h \in M_{LL}$) so that (A.10) and (A.11) are true.

(iv) $h \in M_{LS}$. This only occurs if $r_2 > j_2$. Combined with (A.8), this gives us $r_1 < j_1$ since $r_1 + j_2 < r_1 + r_2 \leq j_1 + j_2$. Since the abandonment is from the longer queue in $\tilde{J}$, we have $j'_1 = \max\{j_1 - 1, j_2\} \geq j_1 - 1$. Since it is from the shorter queue in $\tilde{R}$, r'_1 takes one of two values – r_1 or $r_1 - 1$, so that $r'_1 \leq r_1$ – depending on the rearrangement of redundant jobs as in the proof of Lemma A.2. Thus, $r'_1 \leq r_1 \leq j_1 - 1 \leq j'_1$ so (A.11) is true. The total number in both systems decreases by one so (A.10) is true.

Thus, the modified systems satisfy (A.10), (A.11) and (A.29) $\forall t \geq 0$. As before, we have $T^{\tilde{R}} \leq T^{\tilde{J}}$, and $A^{\tilde{R}} \leq A^{\tilde{J}}$. □

Finally, as the modified systems are coupled exactly with their original counterparts, we have our result. □

Proof of Theorem 2.11. We can show this result by using a similar approach to the proof of Theorem 2.4 in Appendix A.1. (We have finite response times due to abandonments.) To show class 1's preference between CUE and CUS, we consider the state such that there are $n_1 \geq 0$ customers at server 1 and $n_2 = 0$ customers at server 2, when the tagged job, say J, arrives. As in the proof of Theorem 2.4 we only need to consider one "extra" job in the busy period. Customer J joins server 2 in both systems in addition to joining queue 1 in the CUS system, as before. The individual customer abandonments are coupled exactly: If any class 1 customers abandon, they do so in both systems, so the number of class 1 customers, and the total number of abandonments, in either system is the same. If J abandons or finishes service at server 2 before its copy at queue 1 enters service, the two systems are identical. But there is a non-zero probability – equal to n_1 departures (either due to service or abandonment) occurring at server 1 before one at server 2 – that the copy at queue 1 enters service. In this case we will consider the total number of class 1 customers in the two systems and show that there are at least as many in CUS during the current busy period, and we will have our result by Little's Law. Similarly, to show that $A_1^{(\text{Ded,CUE})} \leq A_1^{(\text{Ded,CUS})}$ we need to consider the number of abandonments within this case as well.

Let x be the number of class 1 customers in the system when the copy of J at queue 1 enters service: If $x \geq 1$, then there is a class 1 job in service in the CUE system. Any abandonments of class 1 are exactly coupled between the two systems, so the number of class 1 customers, and the total number of abandonments, in either system is the same. If a service completion occurs at server 2 or a class 2 abandonment occurs, the two systems are identical from then on. If a service completion occurs at server 1, there is an impact if $x \geq 1$: The total number of class 1 customers will be larger (by one) in the CUS system. Moreover, the abandonments are coupled between the two systems, so an abandonment in CUS need not always correspond to one in CUE. When it does not correspond, there is an abandonment in CUS but not in CUE; thus, the number of class 1 abandonments in CUS is larger, and the two systems are identical from then on. Repeating this argument proves our result.

To show class 1's preference between JSQ and CUE, we will consider another extra job D that arrives to the system and the different possibilities for the system state. (As in the proof of Theorem 2.4 we only need to consider one additional "extra" job that arrives in the busy period.) In cases (a) $n_2 = 0, n_1 \geq 0$ and (b) $n_2 = 1, n_1 = 0$, the two systems are identical since D joins the same server in both systems, as before. In case (c) $n_2 = 1, n_1 \geq 1$, D joins queue 2 in the JSQ system and both queues in the CUE system. If any class 1 customers abandon, they do so in both systems, so the number of class 1 customers, and the total number of abandonments, in either system is the same. If D abandons before entering service in CUE or if D enters service at server 2 in CUE, the two systems are identical. As before, there is a non-zero probability that D enters service at server 1 under CUE. Repeating the previous argument using the number of class 1 customers in the system when the copy of D at queue 1 enters service proves our result. □

Proof of Theorem 2.12. We have finite busy periods since $\lambda < \mu_1$ as in proof of Theorem 2.4 in

Appendix A.1. First, we will show class 1's preference between CUE and CUS by using a similar approach to that in the proof of Theorem 2.4. Suppose the state is $n_1 \geq 0, n_2 = 0$ when the tagged job arrives. As in the proof of Theorem 2.4 we only need to consider one "extra" job in the busy period. This job joins both queues in the CUS system, but joins server 2 in the CUE system. The non-zero probability that the copy of the tagged job will enter service at server 1 under CUS is given by $\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{n_1}$. The delay, when it occurs, is $\frac{1}{\mu_1+\mu_2}$, which is the expected service time for a CUS customer when both her copies are in service. (Under CUE there is no corresponding delay.) Thus, class 1 customers arriving after the tagged job during this busy period i.e., of server 1 under CUS, experience additional delay in the CUS system compared to CUE.

Next, we will show class 1's preference between CUE and JSQ. We will consider another extra job D that arrives to the system and the different possibilities for the system state. (As in the proof of Theorem 2.4 we only need to consider one additional "extra" job that arrives in the busy period.) In cases (a) $n_2 = 0, n_1 \geq 0$ and (b) $n_2 = 1, n_1 = 0$, the two systems are identical since D joins the same server in both systems. In case (c) $n_2 = 1, n_1 \geq 1$, D joins queue 2 in the JSQ system and both queues in the CUE system. There is a non-zero probability that D enters service at server 1 under CUE. In this case class 1 customers will experience additional delay in the CUE system compared to JSQ, and we have our result. □

Appendix B

Supplements to Chapter 3

B.1 Proofs

Proof of Lemma 3.1. The first derivative of $\Pi_F(p)$ with respect to (w.r.t.) p is $\Pi'_F(p) = -2p + 1 + s_e + (c - s_e)\theta$. Since the second derivative is negative, setting $\Pi'_F(p) = 0$ yields the $p_F = \frac{1}{2}(1 + s_e + (c - s_e)\theta)$. Now, $0 \leq s_e + (c - s_e)\theta = s_e(1 - \theta) + c\theta \leq c < 1$ since $0 \leq s_e \leq c$, so that $0 < p_F < 1$ and, thus, a valid optimal price. Substituting p_F in $\Pi_F(p)$ gives the optimal profit π_F. □

Proof of Lemma 3.2. The first derivative of $\Pi_N(p)$ w.r.t. p is $\Pi'_N(p) = 1 + c\theta - 2p\theta$. Proceeding as in the Proof of Lemma 3.1, the unconstrained maximizer is $\frac{1}{2\theta}(1 + c\theta)$. Now, $\frac{1}{2\theta}(1 + c\theta) \geq \frac{1}{2}$ since $c \geq 0$ and $\theta \leq 1$. Thus, the solution violates the constraint $p \leq \frac{1}{2}$ when $\theta \in [0, 1)$ or $c \in (0, \frac{1}{2}]$, and the constraint is binding when $\theta = 1$ and $c = 0$. Since, $\Pi'_N(p) > 0$ for $p < \frac{1}{2\theta}(1 + c\theta)$, the optimal price $p_N = \frac{1}{2}$. Substituting p_N in $\Pi_N(p)$ gives π_N. □

Proof of Proposition 3.1. Consider $s_e = \theta = 0$. In this case, the optimal profit π_F under full refund is $\pi_F = \frac{1}{4} - c$ but that under no returns is $\pi_N = \frac{1}{2}(1 - 2c) = \frac{1}{2} - c > \pi_F$. By continuity, the firm prefers no returns when s_e and θ are sufficiently small.

For any fixed s_e and c, π_F is non-decreasing in θ whereas π_N is non-increasing in θ. Similarly, for any fixed c and θ, π_F and π_N are non-decreasing and constant in s_e respectively. Thus, if we can show that $\Delta_{FN} = \pi_F - \pi_N$ is positive at some value of θ_1 and s_1, then we can conclude that Δ_{FN} continues to stay positive for all $s_e > s_1$, keeping $\theta = \theta_1$ and c fixed, or for all $\theta > \theta_1$ keeping $s_e = s_1$ and c fixed. Consider $s_e = 0$. Then, $\Delta_{FN} = \frac{1}{4}((1+c\theta)^2 - 4c - (1-2c)(2-\theta)) = \frac{1}{4}(c^2\theta^2 + \theta - 1)$. Δ_{FN} is increasing in $\theta \in [0, 1]$ – since $c^2\theta^2 + \theta - 1$ has a positive derivative w.r.t. θ (i.e., $2c^2\theta + 1$) – and takes values $-\frac{1}{4}$ and $\frac{c^2}{4}$ at $\theta = 0$ and $\theta = 1$ respectively. Thus, Δ_{FN} crosses the θ-axis once at $\theta_0(c) \in (0, 1)$ for $c > 0$, and stays positive $\forall \theta > \theta_0(c)$, where $\theta_0(c) = \frac{1}{2c^2}(\sqrt{1 + 4c^2} - 1)$ is the positive solution to the quadratic. Next, consider $\theta = 0$ so that $\Delta_{FN} = \frac{1}{4}((1 + s_e)^2 - 4c - 2(1 - 2c)) = \frac{1}{4}(s_e^2 + 2s_e - 1)$. Again, Δ_{FN} is increasing (slope is $(s_e + 1)/2$), takes negative $(-1/4)$ and positive $(1/16)$ values at $s_e = 0$ and $s_e = 1/2$ respectively, so it crosses once at $s_2 = \sqrt{2} - 1$ and is positive $\forall s_e > s_2$. □

Proof of Proposition 3.2. We first prove two lemmas.

Lemma B.1. *The function* $f_\alpha(r) \equiv \Pi_O\left(\frac{1}{2}(1+r^2), r\right)$ *is uniquely maximized at* r_α*, defined as the largest non-negative real root of:* $-\theta r^3 + r(c\theta + \theta - 1) + s_e(1-\theta) = 0$*, and* $r_\alpha \in [0, 1)$.

Proof. The first derivative of $f_\alpha(r) = \frac{1}{2}c\left(r^2\theta + \theta - 2\right) - \frac{1}{4}\theta\left(1-r^2\right)^2 + rs_e(1-\theta) + \frac{1}{2}(1-r^2)$ w.r.t. r is given by $f'_\alpha(r) = -\theta r^3 + r(c\theta + \theta - 1) + s_e(1-\theta)$. To find the maximizer, we need to find the roots of this cubic equation. First consider $s_e \neq 0, \theta \neq 1$ and $\theta \neq 0$. We apply Descartes' rule of signs to find the maximum number of positive roots. Regardless of the sign of the second term $c\theta + \theta - 1$, we only have one sign change so the number of positive roots is exactly 1. To show that this positive root is between 0 and 1, we look at the value of the derivative at $r = 0$ and $r = 1$: $f'_\alpha(0) = s_e(1-\theta) > 0$ since $s_e \neq 0, \theta \neq 1$ and $f'_\alpha(1) = c\theta - 1 + s_e(1-\theta) \leq c - 1 < 0$ since $s_e \leq c$ and $c \leq 1/2$. Thus, $f'_\alpha(r)$ crosses the r axis at some point $r_\alpha \in (0, 1)$.

Note that the second derivative of $f_\alpha(r)$ w.r.t. r is given by $f''_\alpha(r) = -3\theta r^2 + (c\theta + \theta - 1)$. To show that r_α is the required maximizer, we need $f''_\alpha(r_\alpha) < 0$: From the definition of r_α we have $c\theta + \theta - 1 = \theta r_\alpha^2 - s_e(1-\theta)/r_\alpha$ so that $f''_\alpha(r_\alpha) = -3\theta r_\alpha^2 + (c\theta + \theta - 1) = -2\theta r_\alpha^2 - s_e(1-\theta)/r_\alpha < 0$. Thus, we also know that $f'_\alpha(r)$ is positive for $r \in [0, r_\alpha)$ and negative for $r \in (r_\alpha, 1]$. We will now consider the three special cases: $\theta \in \{0, 1\}$ and $s_e = 0$.

(a) $\theta = 0$. In this case, $r = s_e$ is the solution to $f'_\alpha(r) = 0$ and the second derivative is -1 so the unique maximizer is $r = s_e$.

(b) $\theta = 1$. In this case, there are two solutions to $f'_\alpha(r) = 0$, namely $r = 0$ and $r = \sqrt{c}$, but the value of $f''_\alpha(r)$ is c and $-2c$ at the two solutions respectively. When $c > 0$, the second derivative is negative only at $r = \sqrt{c}$, so this is the unique maximizer. When $c = 0$, the second and third derivatives are zero, but the fourth derivative is negative, so that $r = 0$ is the unique maximizer.

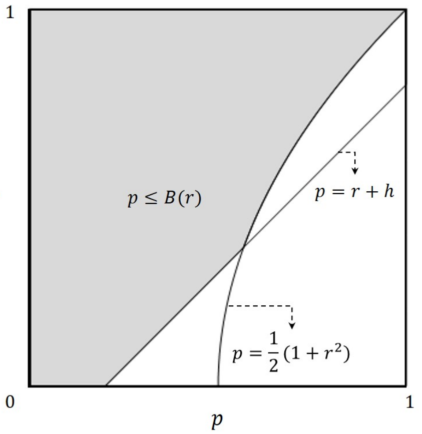
(c) $s_e = 0$. In this case, there are two solutions to $f'_\alpha(r) = 0$, namely $r = 0$ and $r = \sqrt{\frac{c\theta+\theta-1}{\theta}}$, but the value of $f''_\alpha(r)$ is $c\theta + \theta - 1$ and $-2(c\theta + \theta - 1)$ at the two solutions respectively. Thus, the solution depends on the sign of $c\theta + \theta - 1$: if it is positive, the solution is $r = \sqrt{\frac{c\theta+\theta-1}{\theta}}$. Note that $c\theta + \theta - 1 < \theta$ since $\theta \leq 1$ and $c \leq 1/2$ so that $\sqrt{\frac{c\theta+\theta-1}{\theta}}$ lies in $(0, 1)$. On the other hand, if it is negative the solution is $r = 0$. If it is zero, then $r = 0$ is the solution since the fourth derivative is negative (similar to when $c = 0$ in (b) above), but it is a highly flat maximum.

□

Lemma B.2. *The function* $f_\beta(r) \equiv \Pi_O(r + h, r)$ *is uniquely maximized at* $r_\beta \equiv \frac{1}{2}\left(1 + c\theta - 2h\theta + s_e(1-\theta)\right)$*, and* $r_\beta \in [0, 1)$ *if* $h \leq \frac{1}{2}$.

Proof. The function simplifies to a quadratic $-r^2 + r(1 + c\theta - 2h\theta + s_e(1-\theta)) + (c-h)(h\theta - 1)$ which is concave in r and maximized at $r_\beta \equiv \frac{1}{2}(1 + c\theta - 2h\theta + s_e(1-\theta))$. We have $r_\beta \leq \frac{1}{2}(1+c) < 1$ since $h \geq 0, \theta \geq 0, s_e \leq c$ and $c \leq 1/2$, and $r_\beta \geq \frac{1}{2}(1 + s_e - \theta) \geq 0$ since $h \leq 1/2, 0 \leq s_e \leq c$ and $\theta \leq 1$, so that $r_\beta \in [0, 1)$ is a valid maximizer. □

Figure B.1: Feasible region under the BORO outcome.



Note: The feasible region $p \leq B(r)$ is shaded.

Next, we prove Proposition 3.2. First, note that $\Pi_O(p, r)$ is concave and quadratic in p and r. We evaluate the partial derivatives δ_p and δ_r of $\Pi_O(p, r)$ w.r.t. p and r: $\delta_r = -(1 - \theta)(2r - s_e)$, $\delta_p = 1 + c\theta - 2p\theta$. Setting them to zero gives us the unconstrained maximizers: $r_U = \frac{s_e}{2}$, $p_U = \min\{1, \frac{1+c\theta}{2\theta}\}$. When $c \neq 0$ or $\theta \neq 1$, this solution violates the constraint $p \leq \frac{1}{2}(1 + r^2)$ in one of two ways: (1) Case $p_U = \frac{1+c\theta}{2\theta}$: Then, $\frac{1}{2}(1 + r_U^2) - p_U = \frac{1}{2}(1 + \frac{s_e^2}{4} - \frac{1+c\theta}{\theta}) = \frac{1}{8\theta}(-4 + \theta(s_e^2 - 4c + 4)) \leq \frac{1}{8\theta}\left(-4 + \theta(2 - c)^2\right) < 0$ where the first inequality is due to $s_e \leq c$ and the second inequality is due to $\theta < 1$ and $0 < c \leq 1/2$; or (2) Case $p_U = 1$: Then, $\frac{1}{2}(1 + r_U^2) - p_U = \frac{1}{2}(s_e^2/4 - 1) < 0$. When $c = 0$ and $\theta = 1$, the constraint is binding with $p_U = 1/2, r_U = 0$ (since $s_e = 0$). Note that if $p_U = 1$, then $\delta_p > 0$ for all $p < 1$; however, if $p_U = \frac{1+c\theta}{2\theta}$, we have $\delta_p > 0$ for all $p < p_U$ and $\delta_p < 0$ for all $p > p_U$. Similarly, $\delta_r > 0$ for all $r < r_U$ and $\delta_r < 0$ for all $r > r_U$.

Second, we shall prove that the optimal solution will always lie on the boundary of the feasible region. The two-dimensional feasible region is non-convex and is defined by $p \leq B(r)$, where $B(r) \equiv \min\{(1 + r^2)/2, r + h\}$ because of the constraints $p \leq (1 + r^2)/2$ and $p \leq r + h$. Figure B.1 shows the feasible region. Note that $B(r)$ is strictly increasing in r. Thus, the feasible region can be defined as $p \leq B(r)$ or equivalently, $r \geq B^{-1}(p)$. Consider any interior point solution, say (r_1, p_1). By definition, $r_1 > B^{-1}(p_1)$. We will show that (r_1, p_1) cannot be an optimal solution. Let us consider three regions where (p_1, r_1) can lie:

(a) $p_1 \geq B(r_U)$: We have $r_1 > B^{-1}(p_1) \geq r_U$ since (p_1, r_1) is an interior point and by the definition of this case. Since $\delta_r < 0$ for $r > r_U$, we can decrease r to $B^{-1}(p_1)$ from r_1 (keeping p constant at p_1) and increase profits.

(b) $p_1 < B(r_U)$ and $r_1 \leq r_U$: We have $p_1 < B(r_U) \leq p_U$ since (p_U, r_U) either lies on the boundary (when $c = 0$ and $\theta = 0$) or strictly outside it. Since $\delta_p > 0$ for $p < p_U$, we can increase p to $B(r_1)$ from p_1 (keeping r constant at r_1) and increase profits.

(c) $p_1 < B(r_U)$ and $r_1 > r_U$: Since $r_1 > r_U$ we have $\delta_r < 0$ and since $p_1 < p_U$ (from (b) above) we have $\delta_p > 0$. Thus, we can decrease r to r_U (keeping p constant at p_1), and then increase p to $B(r_U)$ (keeping r constant at r_U), thereby reaching the boundary. Both these steps result in an increase in profits.

Thus, (p_1, r_1) cannot be an optimal solution.

Furthermore, $r \leq p$ at the optimal solution: Since the optimal solution lies on the boundary, or $p = B(r)$, the value of $p - r$ is either $(1-r)^2/2$ or h, and thus, non-negative.

Third, we shall characterize the optimum solution for different values of h. It is easy to see that if $h > 1/2$ the constraint $p \leq r + h$ will never bind: This is because $r + h > r + \frac{1}{2} \geq \frac{1}{2}(r^2 + 1) \geq p, \forall r$, where the second inequality is because $r^2 \leq 2r$ when $r \leq 1$. Thus we restrict $h \leq \frac{1}{2}$ for the rest of this proof. Define $r_\gamma = 1 - \sqrt{2h}$ as the unique value of $r \in (0, 1]$ where the two constraints meet. Therefore, the boundary is given by $B(r) = \frac{1}{2}(1 + r^2)$ if $r \geq r_\gamma$, and by $B(r) = r + h$ otherwise. Suppose h is large enough that, by definition, $B(r) = \frac{1}{2}(1 + r^2)$ at the optimal r. Then, the solution can be found by solving the maximization assuming that $p = \frac{1}{2}(1 + r^2)$ holds at the optimum solution, subject to the constraint $r \geq r_\gamma$. Thus, from Lemma B.1, we will have r_α as the optimum solution if $r_\alpha \geq r_\gamma$ or, in other words, $h \geq h_\alpha \equiv \frac{1}{2}(1 - r_\alpha)^2$. This, combined with the fact that $f'_\alpha(r)$ is negative for $r > r_\alpha$, gives us: $r = r_\alpha$, if $h \geq h_\alpha$, and $r = r_\gamma$, otherwise.

Similarly, for small enough h, the boundary is $B(r) = r + h$ at the optimal r, and the solution can be found by solving the maximization assuming that $p = r + h$ holds at the optimum solution, subject to the constraint $r \leq r_\gamma$. Again, from Lemma B.2, we know that the solution will be r_β if $r_\beta \leq r_\gamma$, or in other words, $\Delta(h) \equiv r_\beta - r_\gamma \leq 0$. Now, since $\Delta(h) = \frac{1}{2}\left(1 + c\theta - 2h\theta + s_e(1-\theta)\right) - (1 - \sqrt{2h})$, $\Delta(h) = 0$ has a unique root at $h_\beta \in (0, 1/2]$. This is because of two reasons: (a) the values of $\Delta(h)$ at $h = 0$ and $h = 1/2$ are not of the same sign: $\Delta(0) = \frac{1}{2}\left(1 + c\theta + s_e(1-\theta)\right) - 1 \leq \frac{1}{2}(c - 1) < 0$ since $s_e \leq c < 1$, and $\Delta(1/2) = \frac{1}{2}\left(1 - \theta + c\theta + s_e(1-\theta)\right) \geq 0$ since $(1-\theta) \geq 0, s_e \geq 0$ and $c \geq 0$; and (b) $\Delta(h)$ is increasing in $h < 1/2$: its derivative is $\frac{1}{\sqrt{2h}} - \theta > 0$ since $\theta \leq 1$. Thus, $\Delta(h) < 0$ for $h \leq h_\beta$, and $r_\beta \in [0, 1)$ is a valid maximizer. Note that $\Delta(h)$ is a concave (quadratic) function. Rewriting in terms of $x = \sqrt{h}$, $\Delta(x) = -\theta x^2 + \sqrt{2}x + \frac{1}{2}(c\theta + s_e(1-\theta) - 1)$. Thus, the required root of $\Delta(h)$ is given by $h_\beta = \frac{1}{2\theta^2}\left(1 - \sqrt{(c - s_e)\theta^2 + (s_e - 1)\theta + 1}\right)^2$. This, combined with the concavity of $f_\beta(r)$, gives us: $r = r_\beta$, if $h \leq h_\beta$, and $r = r_\gamma$, otherwise. The following lemma relates h_β and h_α.

Lemma B.3. *$h_\beta \leq h_\alpha$.*

Proof. Define $r_0 \equiv 1 - \sqrt{2h_\beta} = \frac{1}{\theta}\left(\theta - 1 + \sqrt{(c - s_e)\theta^2 + (s_e - 1)\theta + 1}\right)$ and a function $g(s_e) \equiv f'_\alpha(r_0)$. From the proof of Lemma B.1, $g(s_e) = -\theta r_0^3 + r_0(c\theta + \theta - 1) + s_e(1 - \theta)$. The first derivative $g'(s_e)$ w.r.t. s_e is $g'(s_e) = \left(-3\theta r_0^2 + c\theta + \theta - 1\right)\frac{dr_0}{ds_e} + (1 - \theta)$. The goal is to show that $g(s_e) \leq 0, \forall \theta$, so that $r_0 \geq r_\alpha$, and $1 - \sqrt{2h_\alpha} = r_\alpha \leq r_0 = 1 - \sqrt{2h_\beta}$, and we have our result. We can easily show $g(s_e) \leq 0$ to be true using the *Reduce* function in Mathematica. □

Since $h_\beta \leq h_\alpha$, the regions where the solutions for the individual problems in terms of r (under a single constraint) are not equal to r_γ are non-overlapping, and the solution in the intermediate region $h \in (h_\beta, h_\alpha)$ is r_γ. Finally, since the constraint $p \leq r + h$ never binds for $h > \frac{1}{2}$, the optimal solution is the same as when $h \geq h_\alpha$. Thus, the optimal solution is as stated in (3.6).

To prove Proposition 3.2(i)-(iv) we will use the following lemma.

Lemma B.4. *r_α is increasing in s_e when $\theta \neq 1$, and constant otherwise.*

Proof. We first consider $\theta \in (0,1)$. From Lemma B.1 we know that r_α maximizes $f_\alpha(r) = \Pi_O\left(\frac{1}{2}(1+r^2), r\right)$, and the second derivative $f''_\alpha(r_\alpha) < 0$. Now, since $f'_\alpha(r)$ is continuous and differentiable everywhere, we can use the Implicit Function Theorem to evaluate the sign of $\frac{dr_\alpha}{ds_e}$. To do so, we define the implicit function $F(r, s_e) = f'_\alpha(r)$, so that $F(r_\alpha, s_e) = 0$ and $\frac{\partial F(r,s_e)}{dr} = f''_\alpha(r)$. Therefore, $\frac{dr_\alpha}{ds_e} = -\frac{\frac{\partial F(r_\alpha,s_e)}{ds_e}}{f''_\alpha(r_\alpha)}$. Since $f''_\alpha(r_\alpha) < 0$, the sign of $\frac{dr_\alpha}{ds_e}$ is the sign of $\frac{\partial F(r_\alpha,s_e)}{ds_e}$, which is given by $\frac{\partial F(r_\alpha,s_e)}{ds_e} = (1-\theta)$. When $\theta \in (0,1)$ we have $\frac{\partial F(r_\alpha,s_e)}{ds_e} > 0$, and r_α is increasing in s_e. Finally, we consider the special cases: From Lemma B.1 we know that $r_\alpha = s_e$ when $\theta = 0$, and $r_\alpha = \sqrt{c}$ when $\theta = 1$, so we have our result. □

(i) Using the optimal policy, the fee i.e., $p_O - r_O$, takes the value h when $h < h_\alpha$ and $\frac{1}{2}(1-r_\alpha)^2$ otherwise. Thus, the fee is increasing in h when $h < h_\alpha$ and constant in h otherwise.

(ii) By definition, $r_\alpha = 1 - \sqrt{2h_\alpha}$ and $r_\gamma = 1 - \sqrt{2h}$. Since $h < h_\alpha$ in the region that r_γ is optimal we have $r_\gamma = 1 - \sqrt{2h} > 1 - \sqrt{2h_\alpha} = r_\alpha$. Furthermore, r_γ is decreasing in h. To show that $r_\beta \geq r_\gamma$: the optimal solution r_β for any $h < h_\beta$ is at least as high as its value when $h = h_\beta$ since r_β is non-increasing in h. Therefore, it suffices to show that $r_\beta \geq r_\gamma$ for $h = h_\beta$. By the definition of h_β we have $r_\gamma = 1 - \sqrt{2h_\beta} = r_\beta$ at $h = h_\beta$. Thus, we have $r_\beta \geq r_\gamma \geq r_\alpha$, and r_β is non-increasing in h, r_γ is decreasing in h and r_α is constant in h so that r_O is non-increasing in h.

(iii) It is sufficient to show that $r_\alpha \geq s_e$ because of Proposition 3.2(ii). To do so, we will show that $f'_\alpha(r) \geq 0$ at $r = s_e$. We have $f'_\alpha(s_e) = -s_e^3\theta + s_e(c\theta) = s_e\theta(c - s_e^2) \geq 0$ where the inequality is because $0 \leq s_e \leq c, \theta \geq 0$ and $s_e < 1$.

(iv) We will use the Implicit Function Theorem as in the proof of Lemma B.4. Define the implicit function $F(r,\theta) = f'_\alpha(r)$ so that the sign of $\frac{dr_\alpha}{d\theta}$ is the sign of $\frac{\partial F(r_\alpha,\theta)}{d\theta}$ since $f''_\alpha(r_\alpha) < 0$, as before. Now $\frac{\partial F(r_\alpha,\theta)}{d\theta} = r_\alpha(c - r_\alpha^2) + (r_\alpha - s_e)$. Substituting $r = \sqrt{c}$ in $f'_\alpha(r)$ we have $f'_\alpha(\sqrt{c}) = (1-\theta)(s_e - \sqrt{c}) < 0$. The inequality is because when $s_e > 0$ we have $c > 0$ as well, and $s_e \leq c < \sqrt{c}$. This implies that $r_\alpha < \sqrt{c}$ from the proof of Lemma B.1, and thus, $r_\alpha^2 < c$. Combined with $r_\alpha \geq s_e > 0$ (Proposition 3.2(iii)), we have $\frac{\partial F(r_\alpha,\theta)}{d\theta} > 0$ when $\theta \in (0,1)$ and $s_e > 0$.

□

Proof of Proposition 3.3. First, we will show that $0 < \bar{h} \leq 1/2$ exists and is unique. Consider the two expressions $y_L(h) \equiv \frac{1}{2}(1 + h + s_p + (c - s_p - h)\theta)$ and $y_R(h) \equiv h + 1 - \sqrt{2h}$: $y_L(h)$ is non-decreasing in h since its derivative $y'_L(h) = \frac{1}{2}(1-\theta) \geq 0$ whereas $y_R(h)$ is decreasing in h since its derivative $y'_R(h) = 1 - \frac{1}{\sqrt{2h}} < 0, \forall h < 1/2$. Now, $y_L(0) = \frac{1}{2}(1 + c\theta + s_p(1-\theta)) \le \frac{1}{2}(1+c)$ since $s_p \le c$ so that $y_L(0) < 1$. We also have $y_L(h) \ge 1/2$. This is because $y_L(h)$ is non-decreasing in h and $y_L(0) \ge 1/2$ since $s_p, c - s_p$ and θ are all non-negative. On the other hand, $y_R(0) = 1$ and $y_R(1/2) = 1/2$. Thus, $y_R(h)$ is decreasing in h from 1 to 1/2 whereas $y_L(h)$ is increasing from (or constant at) a value at least as large as 1/2. Since both $y_L(h)$ and $y_R(h)$ are continuous, the two curves must cross each other at exactly one point in the interval $(0, 1/2]$ which is precisely the value $\bar{h}$. Furthermore, we also have $y_L(h) \le y_R(h)$ for $h \le \bar{h}$ and $y_L(h) > y_R(h)$ for $\bar{h} < h < 1/2$.

Second, we solve the problem of maximizing $\Pi_S(p)$ subject to $h < p \le y_R(h)$. Note that $y_R(h) > h$ since $h < 1/2$. The first derivative of $\Pi_S(p)$ w.r.t. p is $\Pi'_S(p) = (c-h)\theta + h - 2p + s_p(1-\theta) + 1$. Since the second derivative is negative ($= -2$), setting $\Pi'_S(p) = 0$ yields the unconstrained maximizer $y_L(h)$ (defined above). Now, $y_L(h) > h$ since $h < 1/2$ and $y_L(h) \ge 1/2$. Consider $h \le \bar{h}$: We have $h < y_L(h) \le y_R(h)$, so the unconstrained maximizer $y_L(h)$ satisfies both constraints and is the optimal solution in this region. Next, consider $h > \bar{h}$: We have $y_L(h) > y_R(h) > h$. In this case the unconstrained maximizer $y_L(h)$ violates the constraint $p \le y_R(h)$. Since the first derivative $\Pi'_S(p)$ is positive when $p < y_L(h)$, the optimal price is $y_R(h)$. Thus, the optimal solution is given by (3.8).

Third, we show that π_S is increasing in s_p, or equivalently, $\frac{d\pi_S}{ds_p} > 0$. Consider $h < \bar{h}$ and $\theta \ne 1$: Taking the derivative of π_S w.r.t. s_p we have $\frac{d\pi_S}{ds_p} = \frac{1}{2}(1-\theta)\left(1 + c\theta + s_p(1-\theta) - h(1+\theta)\right)$. Since $h < \bar{h}$ and $\theta \ne 1$, we have $\frac{d\pi_S}{ds_p} > \frac{1}{2}(1-\theta)\left(1 + c\theta + s_p(1-\theta) - \bar{h}(1+\theta)\right)$, and by the definition of $\bar{h}$ (Proposition 3.3), $\frac{d\pi_S}{ds_p} > (1-\theta)(1-\sqrt{2\bar{h}}) \ge 0$. Next, consider $\bar{h} \le h < 1/2$ and $\theta \ne 1$: In this case, $\frac{d\pi_S}{ds_p} = (1-\theta)(1-\sqrt{2h}) > 0$. Also, $\frac{d\pi_S}{ds_p} = 0, \forall h$ if $\theta = 1$.

Finally, we show that π_S is increasing in h for $h < \bar{h}$ when $\theta \ne 1$. Taking the derivative $\frac{d\pi_S}{dh} = (p_S - s_p)(1-\theta)$ when $h < \bar{h}$. Showing that $p_S > s_p$ when $\theta \ne 1$ i.e., $\frac{1}{2}\left[(1 + h + (c-h)\theta) + s_p(1-\theta)\right] > s_p$, is equivalent to showing that $1 + h(1-\theta) > s_p(1+\theta) - c\theta$. Since $s_p \le c < 1$ we have $s_p(1+\theta) - c\theta \le c < 1 \le 1 + h(1-\theta)$. Thus $p_S > s_p$ and $\frac{d\pi_S}{dh} > 0$. □

Proof of Proposition 3.4. (i) First, we show $h_\beta \le \bar{h}$. By its definition (Proposition 3.3), $\bar{h}$ solves the equation: $\frac{1}{2}(1 + c\theta + s_p(1-\theta)) = 1 - \sqrt{2h} + \frac{h}{2}(1+\theta)$. Similarly, h_β solves the equation: $\frac{1}{2}(1 + c\theta + s_e(1-\theta)) = 1 - \sqrt{2h} + h\theta$ from the proof of Proposition 3.2. Define $R_1(h) \equiv 1 - \sqrt{2h} + \frac{h}{2}(1+\theta)$ and $R_2(h) \equiv 1 - \sqrt{2h} + h\theta$. First, since $\theta \le 1$, $R_1(h) \ge R_2(h), \forall h$. Second, $R_1(h)$ and $R_2(h)$ are decreasing in h: their derivatives are $R'_1(h) = -\frac{1}{\sqrt{2h}} + \frac{1+\theta}{2} \le 1 - \frac{1}{\sqrt{2h}} < 0$ and $R'_2(h) = -\frac{1}{\sqrt{2h}} + \theta \le 1 - \frac{1}{\sqrt{2h}} < 0$ respectively, where the inequalities are due to $\theta \le 1$ and $h < 1/2$. Finally, since the LHS of the two equations are identical when $s_e = s_p$ and independent of h, and the RHS values are decreasing and satisfy $R_1(h) \ge R_2(h)$, the unique solutions satisfy $h_\beta \le \bar{h}$.

Next, we show $\bar{h} \le h_\alpha$. Since $\bar{h}$ solves a quadratic in $\sqrt{h}$, we have the solution as $\bar{h} = \left(\frac{\sqrt{2} - \sqrt{Q_1}}{1+\theta}\right)^2$, where $Q_1 = (c - s_p)\theta^2 + (c-1)\theta + 1 + s_p$. Define $\bar{r} \equiv 1 - \sqrt{2\bar{h}} = 1 - \left(\frac{2 - \sqrt{2Q_1}}{1+\theta}\right)$. From the proof of Lemma B.1 we know that $f'_\alpha(r) \le 0$ when $r \ge r_\alpha$. By definition, $h_\alpha = \frac{1}{2}(1 - r_\alpha)^2$ (Proposition 3.2). Thus, showing that $\bar{h} \le h_\alpha$ is equivalent to showing that $\bar{r} \ge r_\alpha$. If we can show that $f'_\alpha(\bar{r}) \le 0$ when $s_e = s_p$, we will have our result. By using the *Reduce* function in *Mathematica*, we find that this is true for all parameter values.

Thus, we can divide h into three regions. We will show $\pi_O \geq \pi_S$ in each region.

(a) $\bar{h} \leq h < 1/2$: In this case the price under the BORS outcome is $p_1 \equiv h + 1 - \sqrt{2h}$. Note that $r_1 \equiv 1 - \sqrt{2h}$ and $p_1 = h + 1 - \sqrt{2h}$ is a feasible solution to the BORO outcome problem. The profit under the BORO outcome $\pi_O \geq \Pi_O(p_1, r_1)$. Thus, $\pi_O \geq \pi_S + (1-\theta)h(p_1 - h)$, where the inequality uses $\pi_S = \Pi_S(p_1) = \theta(p_1 - c)(1 - p_1) + (1-\theta)\left(p_1 - c - (p_1 - s_p)(p_1 - h)\right)$, the definition of $\Pi_O(p_1, r_1)$, and $s_e = s_p$. Since $h < 1/2$ we have $p_1 > h$, or $\pi_O \geq \pi_S$.

(b) $0 \leq h \leq h_\beta \leq \bar{h}$: In this case the optimal refund r_O in the BORO outcome problem is the unconstrained maximizer of $\Pi_O(r+h, r)$ i.e., $r_\beta = \frac{1}{2}\left(1 + c\theta - 2h\theta + s_e(1-\theta)\right)$. From Lemma B.2 we know that $\Pi_O(r+h, r)$ is concave in r. By the definition of this case $h \leq h_\beta$, we have $r_O \leq r_\gamma = 1 - \sqrt{2h}$. The optimal price p_S under the BORS outcome is the unconstrained maximizer of $\Pi_S(p)$ (Proposition 3.3). Define $r_2 \equiv p_S - h = \frac{1}{2}(1 + c\theta + s_p(1-\theta)) - \frac{h}{2}(1+\theta)$. Since $\theta \leq 1$ and $s_p = s_e$ we have $r_2 \leq r_O$. Combined with $r_O \leq 1 - \sqrt{2h}$ and the concavity of $\Pi_O(r+h, r)$, this implies that r_2 is feasible, and results in at most as large a profit as r_O, in the BORO outcome problem. Thus, $\pi_O \geq \Pi_O(r_2 + h, r_2)$ where $\Pi_O(r_2 + h, r_2) = \pi_S + (1-\theta)h(p_S - h)$ by the definition of r_2 and π_S and using $s_p = s_e$. We know that $p_S > h$ from the proof of Proposition 3.3. Therefore, $\pi_O \geq \pi_S + (1-\theta)h(p_S - h) \geq \pi_S$.

(c) $h_\beta < h < \bar{h} \leq h_\alpha$: In this case the optimal policy in the BORO outcome problem is $r_O = 1 - \sqrt{2h}$. From the proof of Proposition 3.2 we know that this is because the unconstrained maximizer r_β of the profit, $\Pi_O(r+h, r)$, violates the constraint $r \leq 1 - \sqrt{2h}$, or $r_\beta > 1 - \sqrt{2h}$. The optimal price p_S under the BORS outcome is the same as in (b). Proceeding as in (b), we set $r_2 \equiv p_S - h$. Since p_S is the unconstrained maximizer and $h < \bar{h}$, we have $p_S < h + 1 - \sqrt{2h}$, or, in other words, $r_2 < 1 - \sqrt{2h}$. Thus, $r_2 < 1 - \sqrt{2h} < r_\beta$, or r_2 is feasible in the BORO outcome problem, and due to the concavity of $\Pi_O(r+h, r)$, results in at most as large a profit as r_O. This implies $\pi_O = \Pi_O(r_O + h, r_O) \geq \Pi_O(r_2 + h, r_2)$. Finally, by the definition of r_2 and π_S and using $s_p = s_e$, we have $\Pi_O(r_2 + h, r_2) = \pi_S + (1-\theta)h(p_S - h)$ as in (b). We know that $p_S > h$ from the proof of Proposition 3.3. Therefore, $\pi_O \geq \Pi_O(r_2 + h, r_2) \geq \pi_S$.

(ii) Consider $h = 0$. Then, $\pi_O = \pi_F$ from Proposition 3.2. From Proposition 3.3 we have π_S is increasing in s_p when $\theta \neq 1$, and $\pi_S = \pi_F$ when $h = 0$ and $s_p = s_e$. Therefore, $\pi_S > \pi_O$ when $s_p > s_e$. By continuity, we have that BORS is preferred at sufficiently small h when $s_p > s_e$ and $\theta \neq 1$. □

B.2 Additional results

Lemma B.5. *(i) $u_{BORO} = E\max\{V, r\} - p \geq 0$ is equivalent to $p \leq \frac{1}{2}(1 + r^2)$.*

(ii) Given $p > r + h$, $u_{BORS} = E\max\{V - p, -h\} \geq 0$ is equivalent to $p \leq h + 1 - \sqrt{2h}$.

Proof. (i) Since $r \in [0,1]$ we have $E\max\{V, r\} - p = \int_0^r r dV + \int_r^1 V dV - p = r^2 + \frac{1}{2}(1 - r^2) - p =$

$\frac{1}{2}(1+r^2)-p$. Thus, $u_{\text{BORO}} \geq 0$ is equivalent to $p \leq \frac{1}{2}(1+r^2)$.

(ii) The condition $p > r+h$ yields $p > r+h \geq h$, and since the price must not exceed the maximum valuation for the item, we also have $p \leq 1$. Taken together, we have $h < p \leq 1$, or $0 < p-h \leq 1$. Using this we can simplify $u_{\text{BORS}} = E\max\{V-p,-h\} = E\max\{V, p-h\}-p = \int_0^{p-h}(p-h)dV + \int_{p-h}^{1} VdV - p = (p-h)^2 + \frac{1}{2}(1-(p-h)^2) - p = \frac{1}{2}\left(p^2 - 2(1+h)p + 1 + h^2\right)$. The quadratic $g(p) \equiv 2u_{\text{BORS}} = p^2 - 2(1+h)p + 1 + h^2$ is convex, takes positive value $1+h^2$ at $p=0$, and takes a non-positive value $h(h-2)$ when $p=1$. Thus, for values of p less than the smaller root of $g(p)$ we have $g(p) \geq 0$, or equivalently $u_{\text{BORS}} \geq 0$ since $u_{\text{BORS}} = \frac{1}{2}g(p)$. The value of this root is easily obtained from the quadratic formula, and is given by $1+h-\sqrt{2h}$. □

Lemma B.6. *Both π_S and π_O may be either increasing or decreasing in θ.*

Proof. Consider BORS and $0 < h < \bar{h}$: When $s_p = c$, $\frac{d\pi_S}{d\theta} = -\frac{h}{2}(1-c+h(1-\theta)) < 0$ so π_S is decreasing in θ, and when $s_p = 0$ and $c = \frac{1}{2}$, $\frac{d\pi_S}{d\theta} = \frac{1}{2}\left(1+h+(\frac{1}{2}-h)\theta\right)\left(\frac{1}{2}-h\right) > 0$ so π_S is increasing in θ.

Consider BORO and $h \geq h_\alpha$: Define $\Pi_O(r,\theta) = \theta((1+r^2)/2-c)(1-(1+r^2)/2) + (1-\theta)((1+r^2)/2-c-(r-s_e)r)$. Then, using the chain rule, we have $\frac{d\Pi_O(r,\theta)}{d\theta} = \frac{\partial\Pi_O(r,\theta)}{\partial\theta} + \frac{\partial\Pi_O(r,\theta)}{\partial r}\frac{dr}{d\theta}$. Since the optimal refund is the unconstrained maximizer r_α we know that $\frac{\partial\Pi_O(r,\theta)}{\partial r} = 0$ when $r = r_\alpha$. Thus, $\frac{d\pi_O}{d\theta} = \left.\frac{\partial\Pi_O(r,\theta)}{\partial\theta}\right|_{r=r_\alpha}$. This simplifies to $\frac{d\pi_O}{d\theta} = (r_\alpha - s_e)r_\alpha + \frac{1}{2}(1+r_\alpha^2)\left(c - \frac{1}{2}(1+{r_\alpha}^2)\right)$. Thus, when $s_e = c > 0$, we have $\frac{d\pi_O}{d\theta} = \frac{1}{4}(1-r_\alpha)^2\left(2c-(1+r_\alpha)^2\right)$. Since $c \leq 1/2$ and $r_\alpha \geq s_e > 0$, we have $\frac{d\pi_O}{d\theta} \leq -\frac{1}{4}(1-r_\alpha)^2 r_\alpha\left(r_\alpha+2\right) < 0$, and thus, $\frac{d\pi_O}{d\theta} < 0$. Next, consider $s_e = 0$ and $c = \frac{1}{2}$, so that $\frac{d\pi_O}{d\theta} = \frac{1}{4}r_\alpha^2(3-r_\alpha^2) \geq 0$. From the proof of Lemma B.1, we know that $r_\alpha > 0$ when $s_e = 0$ provided $c\theta + \theta - 1 > 0$, or $\theta > 2/3$ when $c = 1/2$. Thus, $\frac{d\pi_O}{d\theta} > 0$ when $\theta > 2/3$. □

Lemma B.7. *The profit under full-refund in the case of heterogeneous hassle costs is non-decreasing in both θ and δ.*

Proof. The firm's expected profit is given by $\Pi^H(p) = \theta(p-c)(1-p) + (1-\theta)(p-c-\delta(p-s_p)p-(1-\delta)(p-s_e)p)$. As in §3.4, this problem is also unconstrained, and the optimal price p_F^H is the maximizer of $\Pi^H(p)$. Proceeding as in the proof of Lemma 3.1, $p_F^H = \frac{1}{2}\left(1+c\theta+(1-\theta)(\delta s_p + (1-\delta)s_e)\right)$. Since $s_p \leq c$ and $s_e \leq c$ we have $p_F^H \leq \frac{1}{2}(1+c) < 1$. Now, $p_F^H \geq 0$ since $c \geq 0, 0 \leq \theta \leq 1, 0 \leq \delta \leq 1, s_p \geq 0$, and $s_e \geq 0$.

The derivative of the optimal profit $\Pi^H(p_F^H)$ w.r.t. θ is $y_\theta \equiv \frac{1}{2}(c-\delta s_p-(1-\delta)s_e)[1+c\theta+(1-\theta)(\delta s_p+(1-\delta)s_e)] = (c-\delta s_p-(1-\delta)s_e)p_F^H$ and that w.r.t. δ is $y_\delta \equiv \frac{1}{2}(1-\theta)(s_p-s_e)[1+c\theta+(1-\theta)(\delta s_p+(1-\delta)s_e)] = (1-\theta)(s_p-s_e)p_F^H$. We know $p_F^H \geq 0$. Since $s_p \leq c$ and $s_e \leq c$ we have $c-\delta s_p-(1-\delta)s_e \geq 0$ in the expression for y_θ, so that $y_\theta \geq 0$. Finally, since $s_p \geq s_e$ and $\theta \leq 1$, we have $y_\delta \geq 0$ as well. □

B.3 Parameter values used in figures

The parameter values used in our numerical calculations are chosen with the intention of representing a typical omnichannel setting. In Figures 3.1-3.2 we use $\theta = 0.5, c = 0.3$, and $s_e = 0.2$. We chose $\theta = 0.5$ since Macy's (2015) finds that 53% of its customers prefer to shop at their brick-and-mortar stores. On the one hand, the salvage value at third party outlet stores is about 10-20% of their value (Stock et al. 2006); on the other hand, Su (2009) shows that the salvage value must be more than 83% of the cost (see §3.4.3 for details) for single-channel firms to prefer full refunds. Finally, using our model, the salvage value must be more than 66% of the cost for an *omnichannel* firm with $\theta = 0.5$ to prefer full refunds. This is obtained by comparing π_F against π_N for $\theta = 0.5$. Thus, we set $s_e = 2c/3$. We choose $c = 0.3$ to obtain a full refund profit margin of 30%. The profit margin is computed as $\pi_F/(\pi_F + c_F)$, where c_F is the total cost incurred by the firm. This cost is given by $c_F = \theta(1-p_F)c + c(1-\theta)$ since the firm procures products at cost c for all the online-savvy customers (proportion $1-\theta$) and for $(1-p_F)$ of the store-savvy customers (proportion θ). This yields $s_e = 0.2$. From annual reports we find that omnichannel firms enjoy a gross profit margin in the range of 30% to 40%: The 2016 gross margin was 34.9% and 39.4% for Nordstrom and Macy's respectively, whereas that for J. C. Penney was in the range of 29.4% to 36% from 2013 to 2016.

We chose in-store salvage value $s_p = s_e + 0.1$ in Figures 3.3-3.4. The expected cross-selling revenue is \$20 per sale: J.C. Penney claims that about one-third of shoppers spend an additional \$60 when they visit a store to make a return (WSJ 2016), which yields expected revenue of \$20. Using our model parameters we have $p_F = 0.625$ so that assuming a typical sale to be about \$100 (and normalizing the price to p_F) gives the additional cross-selling revenue $s_p - s_e \approx 20 * 0.625/100 \approx 0.1$. Additionally, in Figure 3.4 we use $\beta = 1.5$: Bower and Maxham III (2012) note that free return policies result in future customer spending that is 158% to 457% more by the end of two years, and in order to be conservative in the effect of demand stimulation, we choose a value closer to the lower end of this range.

Appendix C

Supplements to Chapter 4

C.1 Proofs

Proof of Proposition 4.1. Let $R(i,q)$ denote the reward in state i under action q. Then, $R(i,q) \equiv (r - c(q))i$, which is unbounded. Since the state space S is countable, we will use results in Puterman 1994 (replicated below) to show our result. Note that Puterman (1994) uses λ to denote the discount rate, $r(s,a)$ to denote the reward in state s, and $p(j|s,a)$ to be the transition probability from state s to state j, under action $a \in \mathcal{A}_s$ for $s \in S$. Furthermore, r_d and P_d are the return vector and transition matrix respectively, corresponding to the decision rule d. Let $w(i)$ be as defined in Puterman (1994).

Assumption C.1 (Assumption 6.10.1 in Puterman 1994)**.** *There exists a constant $\mu < \infty$ such that $\sup_{a \in \mathcal{A}_s} |r(s,a)| \leq \mu w(s)$.*

Proposition C.1 (Proposition 6.10.5(a) in Puterman 1994)**.** *Suppose there exists a constant $L > 0$, for which $\sum_{j \in S} p(j|s,a) w(j) \leq w(s) + L$ for $a \in \mathcal{A}_s$ and $s \in S$. Then Assumption 6.10.2 holds.*

Theorem C.1 (Theorem 6.10.4(a) and (c) in Puterman 1994)**.** *Let S be a countable set and suppose that Assumptions 6.10.1 and 6.10.2 hold. Then the following holds true: The optimality equation has a unique solution v^*_λ. And, if there exists a $d^* \in D$ satisfying $d^* \in \arg\max_{d \in D}\{r_d + \lambda P_d v^*_\lambda\}$, then $(d^*)^\infty$ is an optimal policy.*

Lemma C.2. *Our model satisfies Assumption C.1.*

Proof. Let $w(i) \equiv ri, \forall i \in S$. We need to show there exists a constant $\mu < \infty$ such that $\sup_{q \in \mathcal{Q}} |R(i,q)| \leq \mu w(i)$. The highest value of $R(i,q)$ is ri when $q = 0$, and the lowest value of $R(i,q)$ is $(r - c(\bar{q}))i$ when $q = \bar{q}$. Thus, $\sup_{q \in \mathcal{Q}} |R(i,q)| = \max\{ri, |r - c(\bar{q})|i\}$, and Assumption C.1 holds with $\mu = \max\{1, \frac{c(\bar{q})}{r} - 1\}$ since $r > 0$. □

Assumption C.2. *There exists a constant $L > 0$, for which $\sum_{j \in S} p(j|s,a) w(j) \leq w(s) + L$ for $a \in \mathcal{A}_s$ and $s \in S$.*

Lemma C.3. *Our model satisfies Assumption C.2.*

Proof. We need to show that $\sum_{j\in S} P_{ij}(q)w(j) \leq w(i) + L, \forall i \in S, q \in \mathcal{Q}$ for some $L > 0$. We have

$$\begin{aligned} \sum_{j\in S} P_{ij}(q)w(j) &= \sum_{j\in S} P_{ij}(q)rj \\ &= r\lambda(q) + ri - rp(q)i \\ &\leq ri + r\lambda(\bar{q}) = w(i) + L, \end{aligned}$$

where $L \equiv r\lambda(\bar{q})$, the second equation uses the fact that the expected next state from any state i is $i(1-p(q)) + \lambda(q)$, and the inequality is due to $p(q) \geq 0, r > 0, i \geq 0$, and $\lambda(q) \leq \lambda(\bar{q})$. □

From Lemma C.2 we have that the first condition (Assumption 6.10.1) in Theorem C.1 is met. From Lemma C.3 and Proposition C.1, the second condition (Assumption 6.10.2) in Theorem C.1 is also met. Thus, our result follows from Theorem C.1. □

Proof of Lemma 4.1. Let $z(q) = \frac{r-c(q)}{1-\beta(1-p(q))}$. By definition $\alpha \equiv \max_{q\in\mathcal{Q}}\{z(q)\}$. Let $q_{\max} = \min\{\tilde{q}, \bar{q}\}$, where $\tilde{q}$ is defined as the value of q at which the per patient profit is zero i.e., $r - c(\tilde{q}) = 0$. Since providing quality $q > \tilde{q}$ results in negative profits $z(q) < 0$ whenever $q > \tilde{q}$, the feasible set can be reduced to $q \leq q_{\max}$. The following lemma states that there is a unique maximizer of $z(q)$.

Lemma C.4. *$z(q)$ is uniquely maximized at x_0. Furthermore, the feasible set is reduced to $q \leq \bar{q}$.*

Proof. The first derivative $z'(q) = \frac{1}{(1-\beta(1-p(q)))^2}[-c'(q)(1-\beta(1-p(q))) - \beta p'(q)(r-c(q))]$. Let $N(q)$ denote the numerator i.e., $N(q) = -c'(q)(1-\beta(1-p(q))) - \beta p'(q)(r-c(q))$ and its root be x_0, i.e., x_0 solves

$$-\frac{c'(q)}{\beta p'(q)} = \frac{r-c(q)}{1-\beta(1-p(q))}. \tag{C.1}$$

First, we show that x_0 maximizes $z(q)$. The second derivative is $z''(q) = \frac{d}{dq}\left[\frac{N(q)}{(1-\beta(1-p(q)))^2}\right]$, and at $q = x_0$ its value is given by $z''(x_0) = \frac{1}{[1-\beta(1-p(x_0))]^2} \left.\frac{dN}{dq}\right|_{q=x_0}$ because $N(x_0) = 0$. Now $\frac{dN}{dq} = -c''(q)(1-\beta(1-p(q))) - \beta p''(q)(r-c(q)) < 0$ since $c'' > 0, p'' > 0, x_0 \leq \tilde{q}$ and $\beta(1-p(q)) \leq 1$. Thus, $z''(x_0) < 0$ so that x_0 is a maximizer.

Next, we show that x_0 is unique. We already have that $\frac{dN}{dq} < 0$ from above, or $N(q)$ is decreasing. The sign of $z'(q)$ is the same as that of $N(q)$, and $N(q) = 0$ at $q = x_0$. Combined with $N(q)$ being decreasing, this implies that $N(q)$, or equivalently $z'(q)$, is positive for $q < x_0$, zero at $q = x_0$, and negative otherwise. Thus, $z(q)$ is uniquely maximized at x_0.

Finally, the L.H.S of (C.1) is positive at the solution since $c' > 0$ and $p' < 0$ so that the R.H.S. must also be positive. Thus, $r - c(x_0) \geq 0$, or $x_0 \leq \tilde{q}$. This implies that the feasibility constraint can be simplified to $q \leq \bar{q}$. □

The following lemma defines the optimal solution q_0.

Lemma C.5. *The optimal solution to* $\max_{q\in\mathcal{Q}}\{z(q)\}$ *is* $q_0 = \min\{x_0, \bar{q}\}$.

Proof. Given that the feasibility constraint is simplified to $q \leq \bar{q}$ (Lemma C.4), we only need to consider the following two cases to derive q_0:

1. $x_0 \leq \bar{q}$: In this case the solution x_0 is the optimal solution as it does not violate our constraint.
2. $x_0 > \bar{q}$: In this case x_0 is not feasible. From the proof of Lemma C.4 we have that the first derivative $z'(q)$ is positive for $q < x_0$. Since the unique maximizer is not feasible due to the constraint $q \leq \bar{q}$, this implies that the optimal solution is $\bar{q}$.

Therefore, we have our result. □

Next, we will show that q_0 maximizes the maximand in (4.1) and that the return function $v_0(i) = \alpha i + \gamma$ solves the system of simultaneous equations given by (4.1). We will start from the R.H.S. of (4.1) and show that it simplifies to the L.H.S. The maximand $y(i,q)$ in (4.1) corresponding to $v_0(j) = \alpha j + \gamma$ is $y(i,q) = [r - c(q)]i + \beta \sum_{j \in S}(\alpha j + \gamma)P_{ij}(q)$, so that the R.H.S. is $\max_{q \in \mathcal{Q}} y(i,q)$ and the L.H.S. is $v_0(i)$. In other words, we need to show that $\max_{q \in \mathcal{Q}} y(i,q)$ simplifies to $v_0(i)$.

We first simplify $y(i,q)$:

$$\begin{aligned} y(i,q) &= [r - c(q)]i + \beta \sum_{j \in S}(\alpha j + \gamma)P_{ij}(q) && \text{(C.2)} \\ &= [r - c(q)]i + \beta\alpha \sum_{j \in S} jP_{ij}(q) + \beta\gamma \sum_{j \in S} P_{ij}(q) && \text{(C.3)} \\ &= [r - c(q)]i + \beta\alpha \sum_{j \in S} jP_{ij}(q) + \beta\gamma && \text{(C.4)} \\ &= [r - c(q)]i + \beta\alpha[\lambda_0 + i(1 - p(q))] + \beta\gamma && \text{(C.5)} \\ &= [r - c(q) + \beta\alpha(1 - p(q))]i + \beta\gamma + \beta\alpha\lambda_0, && \text{(C.6)} \\ &= [r - c(q) + \beta\alpha(1 - p(q))]i + \gamma, && \text{(C.7)} \end{aligned}$$

where (C.4) uses $\sum_{j \in S} P_{ij}(q) = 1$, and (C.7) uses the definition of γ.

Define $h(q) \equiv r - c(q) + \beta\alpha(1 - p(q))$. Since γ is a constant, we only have to show that q_0 maximizes $h(q)i + \gamma$ subject to $q \in \mathcal{Q}$ and $\max_{q \in \mathcal{Q}} h(q)i + \gamma = v_0(i)$. We consider two cases:

1. $i = 0$: In this case we have $y(0,q) = \gamma$. In this case there is no unique maximizer and $\max_{q \in \mathcal{Q}} h(q)i + \gamma = \gamma$, which equals the L.H.S. of (4.1), i.e., $v_0(0)$.
2. $i > 0$: In this case let the unconstrained maximizer be x_1. The first derivative of $h(q)$ is $h'(q) = -c'(q) - \beta\alpha p'(q)$ and the second derivative is $h''(q) = -c''(q) - \beta\alpha p''(q) < 0$ since $c'' > 0$ and $p'' > 0$. Thus, x_1 solves
$$\alpha = \frac{-c'(q)}{\beta p'(q)}. \qquad \text{(C.8)}$$

   The expression $\frac{-c'(q)}{\beta p'(q)}$ is increasing in q: Its derivative is $-\frac{1}{\beta(p'(q))^2}(p'(q)c''(q) - c'(q)p''(q)) > 0$. Moreover, $h'(q) > 0$ when $q < x_1$ since the second derivative is negative. We will condition on the different values of q_0 relative to $\tilde{q}$:

   (a) $q_0 = x_0$. By definition $\alpha = \frac{-c'(x_0)}{\beta p'(x_0)}$ from (C.1) so that $h(q)$ is maximized at $x_1 = x_0$.

Furthermore, $h(x_1) = \alpha$: This is because $\alpha = \frac{r-c(q_0)}{1-\beta(1-p(q_0))}$ by its definition. Rearranging the terms, we have $\alpha(1-\beta(1-p(q_0))) = r-c(q_0)$, or $h(x_0) = \alpha$. Thus, $\max_{q\in\mathcal{Q}} h(q)i+\gamma = \alpha i + \gamma = v_0(i)$.

(b) $q_0 = \bar{q} < x_0$. In this case we know that $z'(\bar{q}) > 0$ (see Lemma C.5 and proof of Lemma C.4). Thus, by the definition of $z'(q)$ we have $-c'(q)(1-\beta(1-p(q)))-\beta p'(q)(r-c(q)) > 0$ at $q = \bar{q}$, or $\frac{r-c(\bar{q})}{1-\beta(1-p(\bar{q}))} > -\frac{c'(\bar{q})}{\beta p'(\bar{q})}$. By definition $\alpha = \frac{r-c(\bar{q})}{1-\beta(1-p(\bar{q}))}$ so that $\alpha > -\frac{c'(\bar{q})}{\beta p'(\bar{q})}$. Combined with $\frac{-c'(q)}{\beta p'(q)}$ increasing in q and (C.8), this implies that $x_1 > \bar{q}$. Since x_1 violates the feasibility condition and $h'(q) > 0$ for $q < x_1$, the optimal solution is $\bar{q}$. Again, $h(x_1) = \alpha$ by definition of α (see above) so that $\max_{q\in\mathcal{Q}} h(q)i+\gamma = \alpha i+\gamma = v_0(i)$.

From Proposition 4.1 this implies that q_0 is optimal. □

Derivation of q_H in Definition 4.1. We need to evaluate the maximand in (4.1) i.e., $y(i,q) \equiv [r-c(q)]i + \sum_{j\in S}\beta v_0(j)P_{ij}(q), \forall i \in S, q \in \mathcal{Q}$. This is because, by definition, the heuristic maximizes $y(i,q)$ with respect to $q \in [0,\bar{q}]$. Given the starting return function $v_0(j) = \alpha j + \gamma$ and proceeding similar to Proof of Proposition 4.1 and (C.2)-(C.5), we have $y(i,q) = [r-c(q)]i + \beta\alpha[\lambda(q) + i(1-p(q))] + \beta\gamma$. In other words, $y(i,q) = [r-c(q)+\beta\alpha(1-p(q))]i + \beta\alpha\lambda(q) + \beta\gamma$. Note that $\beta\gamma$ is independent of q and i. When $i = 0$ we have $y(i,q) = \beta\alpha\lambda(q) + \beta\gamma$, which is maximized at $q = \bar{q}$ since $\lambda'(q) \geq 0$. The heuristic policy q_H is given by $q_H(i) \equiv \arg\max_{q\in\mathcal{Q}}\left\{r-c(q)+\beta\alpha(1-p(q))+\beta\alpha\frac{\lambda(q)}{i}\right\}, \forall i > 0$ and $q_H(0) = \bar{q}$ as in Definition 4.1. □

Proof of Lemma 4.2. (a) When $i = 0$, we have $q_H(0) = \bar{q}$ so that $q_H(1) \leq \bar{q} = q_H(0)$. From the definition of q_0 (Lemma C.5) we also have that $q_0 \leq \bar{q} = q_H(0)$. Thus, we only need to consider $i > 0$. Let $z(i,q) \equiv r-c(q)+\beta\alpha(1-p(q))+\beta\alpha\frac{\lambda(q)}{i}, i > 0$, so q_H maximizes $z(i,q)$ subject to $q \in \mathcal{Q}$. The derivative $z_1(i,q)$ of $z(i,q)$ w.r.t. q is given by $z_1(i,q) = -c'(q)-\beta\alpha p'(q)+\beta\alpha\frac{\lambda'(q)}{i}$, and its second derivative $z_2(i,q)$ is given by $z_2(i,q) = -c''(q) - \beta\alpha p''(q) + \beta\alpha\frac{\lambda''(q)}{i}$. Since $c'' > 0, p'' > 0$ and $\lambda'' \leq 0$ we have $z_2(i,q) < 0$. Suppose $x_1(i)$ is the value of q that solves the equation $z_1(i,q) = 0$. Since $z_2(i,q) < 0$ we have $z_1(i,q) > 0$ for $q < x_1(i)$, and the heuristic $q_H(i) = \min(\bar{q}, x_1(i))$. From the Proof of Lemma 4.1 (cases (a) and (b)) we know that $\alpha \geq \frac{-c'(q_0)}{\beta p'(q_0)}$, or $-\beta p'(q_0)\alpha - c'(q_0) \geq 0$ since $p'(q_0) < 0$. Thus, $z_1(i,q_0) = -c'(q_0) - \beta\alpha p'(q_0) + \beta\alpha\frac{\lambda'(q_0)}{i} \geq \beta\alpha\frac{\lambda'(q_0)}{i} \geq 0$ since $\lambda'(q) \geq 0$. Combined with the fact that $z_1(i,q) \geq 0$ for $q \leq x_1(i)$, we have $q_0 \leq x_1(i), \forall i \in S$. Since $\bar{q} \geq q_0$ and $q_H(i) = \min(\bar{q}, x_1(i))$, we have $q_0 \leq q_H(i), \forall i$.

It is easy to see that $z_1(i,q)$ is non-increasing in i, i.e., $z_1(i+1,q) \leq z_1(i,q)$. Combined with $z_1(i,q)$ decreasing in q since $z_2(i,q) < 0$, this implies that the value of q at which $z_1(i+1,q) = 0$ is at most that at which $z_1(i,q) = 0$, or $x_1(i+1) \leq x_1(i)$. It is easy to show that $q_H(i+1) \leq q_H(i)$ as well by conditioning on the different cases for $\bar{q}$: (a) $\bar{q} > x_1(i)$: Then, $q_H(i+1) = x_1(i+1) \leq x_1(i) = q_H(i)$; (b) $\bar{q} < x_1(i+1)$: Then, $q_H(i+1) = q_H(i)$; and (c) $x_1(i+1) \leq \bar{q} \leq x_1(i)$: Then, $q_H(i+1) = x_1(i+1) \leq \bar{q} = q_H(i)$.

(b) First we will show that q_0 is non-decreasing in r. Consider $z'(q)$ and x_0 from the proof of Lemma 4.1. Now, since $z'(q)$ is continuous and differentiable everywhere, we can use the Implicit Function Theorem to evaluate the sign of $\frac{dx_0}{dr}$. To do so, we define the implicit function $F(q,r) = z'(q)$, so that $F(x_0, r) = 0$ and $\frac{\partial F(q,r)}{dq} = z''(q)$. Therefore, $\frac{dx_0}{dr} = -\frac{\frac{\partial F(x_0,r)}{dr}}{z''(x_0)}$. Since $z''(x_0) < 0$, the sign of $\frac{dx_0}{dr}$ is the sign of $\frac{\partial F(x_0,r)}{dr}$, which is given by $\frac{\partial F(x_0,r)}{dr} = -\frac{\beta p'(q)}{(1-\beta(1-p(q)))^2} > 0$, where the inequality is because $p'(q) < 0$. Thus, x_0 is increasing in r. Since $q_0 = \min\{\bar{q}, x_0\}$, we have that q_0 is non-decreasing in r.

Next, we show that α is increasing in r. Consider x_0 and α from the proof of Lemma 4.1. Suppose we have two values of the reimbursement rate r_1 and r_2 such that $r_2 > r_1$. Let α_i and $x_0^{(i)}$ be the values of α and x_0 respectively, corresponding to rate r_i. Then, the quality $q_0^{(i)}$ corresponding to r_i is given by $q_0^{(i)} = \min\{\bar{q}, x_0^{(i)}\}$. Since x_0 is increasing in r, we have $x_0^{(2)} > x_0^{(1)}$. We will condition on the value of $\bar{q}$ to show that $\alpha_2 > \alpha_1$:

(i) $\bar{q} > x_0^{(1)}$. In this case $\alpha_2 = \frac{r_2 - c(q_0^{(2)})}{1-\beta(1-p(q_0^{(2)}))} > \frac{r_2 - c(x_0^{(1)})}{1-\beta(1-p(x_0^{(1)}))} > \frac{r_1 - c(x_0^{(1)})}{1-\beta(1-p(x_0^{(1)}))} = \alpha_1$, where the equations are by definition of α_i, the first inequality is due to the suboptimality of $x_0^{(1)}$ when the reimbursement rate is r_2, and the second inequality is due to $r_2 > r_1$.

(ii) $\bar{q} \leq x_0^{(1)}$. In this case $\alpha_2 = \frac{r_2 - c(\bar{q})}{1-\beta(1-p(\bar{q})} > \frac{r_1 - c(\bar{q})}{1-\beta(1-p(\bar{q})} = \alpha_1$, where the equations are by definition of α_i and the inequality is due to $r_2 > r_1$.

To show that $q_H(i)$ is non-decreasing in r, it suffices to show that $x_1(i)$ is non-decreasing in r for $i > 0$. This is because $q_H(0) = \bar{q}$ which is independent of r, and $q_H(i) = \min\{\bar{q}, x_1(i)\}$. From the proof of (a) above, we know that $x_1(i)$ solves $z_1(i,q) = 0$. We can apply the Implicit Function Theorem again with $F(q,r) = z_1(i,q)$ to compute $\frac{dx_1(i)}{dr}$. Since $z_2(i,q) < 0$, the sign of $\frac{dx_1(i)}{dr}$ is the sign of $\frac{\partial F(x_1(i),r)}{dr}$, which is given by $\frac{\partial F(x_1(i),r)}{dr} = \beta\left[-p'(q) + \frac{\lambda'(q)}{i}\right]\frac{d\alpha}{dr} > 0$, where the inequality is because $p'(q) < 0, \lambda'(q) > 0$, and α is increasing in r.

(c) As in (b) we will show that $x_1(i)$ is larger when λ' is larger for $i > 0$ because $q_H(0) = \bar{q}$ which is constant. We know from the proof of (a) above that x_1 solves $z_1(i,q) = 0$, where $z_1(i,q) = -c'(q) - \beta\alpha p'(q) + \beta\alpha\frac{\lambda'(q)}{i}$. Rearranging we have that $x_1(i)$ solves $\beta\alpha\frac{\lambda'(q)}{i} = c'(q) + \beta\alpha p'(q)$. The R.H.S. is increasing in q: Its derivative is $c'' + \beta\alpha p'' > 0$ since $c'' > 0, p'' > 0$ and $\alpha > 0$. The L.H.S. is decreasing in q for $i > 0$: Its derivative is $\beta\alpha\lambda''/i < 0$ since $\lambda'' < 0$. The L.H.S. is larger when $\lambda'(q)$ is larger, so that the intersection point of the two curves i.e., $x_1(i)$, will be larger. Since $q_H(i) = \min\{\bar{q}, x_1(i)\}$, we have our result.

□

C.2 Additional results

Violation of Lemma 4.2(a). We will show that Lemma 4.2(a) may be violated by the optimal policy by constructing a counterexample. The following parameters provide one such counterexample: $\lambda(q) = q$, $p(q) = 1 - 0.9q$, and $c_0 = 4$; the remaining values are as in §4.4. This results in $q_0 = 0.11$.

Figure C.1 shows the quality under the optimal and heuristic policies. While the heuristic still satisfies Lemma 4.2(a), the optimal policy $\mathbf{q}^*$ is increasing in state i for $i \geq 5$, and $\mathbf{q}^*$ chooses values that are smaller than q_0. The counterexample is characterized by high cost of quality and a departure probability function that is sensitive to quality. Revenues are increasing in the number patients, so the hospice ideally prefers to increase its patient census. The high cost of quality incentivizes the hospice to provide high quality when the patient census is zero (so that costs are zero) to attract the most demand without high costs. And, at low patient census, the hospice is better off lowering quality to accelerate its course to the zero state, where it can take advantage of the high quality "for free." However at high patient census, the hospice can afford to increase quality to ensure that not too many patients depart; it is more profitable to try to extract revenue from patients than to immediately drift to zero to bring in a new batch.

Figure C.2 shows the evolution of a sample path over 100 stages under the optimal policy for this example, starting from either the empty state (Figure C.2(a)) or a larger state (Figure C.2(b)). We see that the patient census bounces off the empty state very often: It reaches a non-zero state but does not stay there for a long time. On the other hand, when the starting state is large enough (Figure C.2(b)), the policy offers a slightly higher quality (Figure C.1) and the patient census does not reach zero immediately. To facilitate comparison against more realistic examples, Figure C.3 shows the sample path corresponding to one of our experiments in §4.4: Relative to Figure C.2, the census in this example moves steadily towards a non-zero state.

Figure C.1: Comparison of optimal and heuristic policies for the counterexample in Appendix C.2.

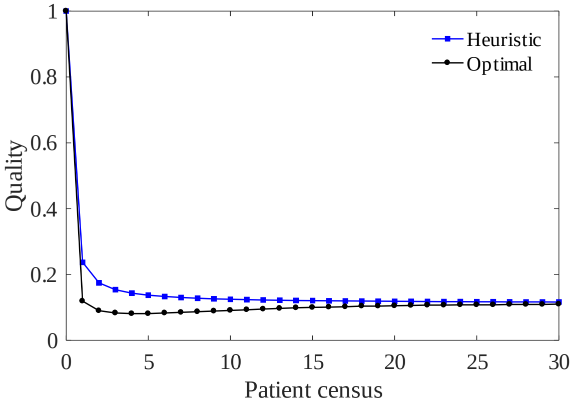
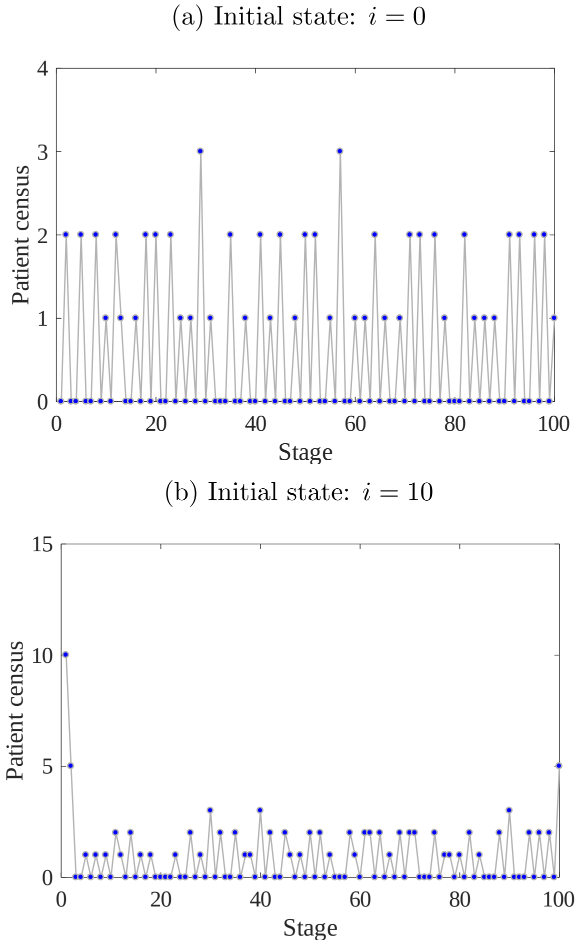
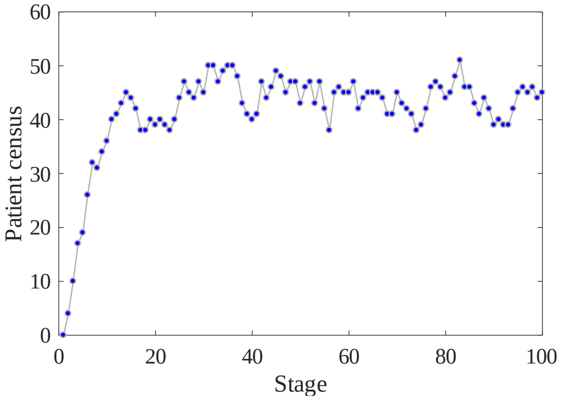


Figure C.2: Sample paths under the optimal policy for the counterexample in Appendix C.2.


(a) Initial state: $i = 0$

(b) Initial state: $i = 10$

Figure C.3: Sample path under the optimal policy for for-profit, $(a_0, a_1) = (3, 2)$, and $(b_0, b_1) = (0.1, 0.04)$ with initial state $i = 0$.



Bibliography

Accenture (2015) U.S. retailers struggling to meet consumer expectations around mobile and in-store experience, Accenture study finds. Available at https://newsroom.accenture.com.

Akgun OT, Righter R, Wolff R (2011) Multiple-server system with flexible arrivals. *Advances in Applied Probability* 43(04):985–1004.

Akgun OT, Righter R, Wolff R (2012) Understanding the marginal impact of customer flexibility. *Queueing Systems* 71(1):5–23.

Altug MS, Aydinliyim T (2016) Counteracting strategic purchase deferrals: The impact of online retailers' return policy decisions. *Manufacturing & Service Operations Management* 18(3):376–392.

Ardekani MS, Orlowski JM (2010) Multiple listing in kidney transplantation. *American Journal of Kidney Diseases* 55(4):717–725.

Armony M, Plambeck EL (2005) The impact of duplicate orders on demand estimation and capacity investment. *Management Science* 51(10):1505–1518.

Ata B, Killaly BL, Olsen TL, Parker RP (2013) On hospice operations under medicare reimbursement policies. *Management Science* 59(5):1027–1044.

Ata B, Skaro A, Tayur S (2016) Organjet: Overcoming geographical disparities in access to deceased donor kidneys in the United States. *Management Science* .

Bell DR, Gallino S, Moreno A (2014) How to win in an omnichannel world. *MIT Sloan Management Review* 56(1):45.

Bell DR, Gallino S, Moreno A (2017) Offline showrooms in omnichannel retail: Demand and operational benefits. *Management Science* .

Bloomberg (2016) Gap tumbles as dismal store traffic drags down july sales. Available at https://www.bloomberg.com/news/articles/2016-08-08/gap-shares-drop-as-dismal-store-traffic-drags-down-july-sales.

Blumstein JF (1989) *Organ transplantation policy: issues and prospects* (Duke University Press).

Body Labs (2016) Apparel & footwear retail survey report. Available at https://www.bodylabs.com/resources/white-papers/2016-apparel-footwear-retail-survey-report/.

Bower AB, Maxham III JG (2012) Return shipping policies of online retailers: Normative assumptions and the long-term consequences of fee and free returns. *Journal of Marketing* 76(5):110–124.

Brynjolfsson E, Hu YJ, Rahman MS (2013) Competing in the age of omnichannel retailing. *MIT Sloan Management Review* 54(4):23.

Carlson MD, Barry C, Schlesinger M, McCorkle R, Morrison RS, Cherlin E, Herrin J, Thompson J, Twaddle

ML, Bradley EH (2011) Quality of palliative care at US hospices: results of a national survey. *Medical care* 49(9):803.

Castle NG, Engberg J (2007) The influence of staffing characteristics on quality of care in nursing homes. *Health Services Research* 42(5):1822–1847.

Cavallo A (2017) Are online and offline prices similar? Evidence from large multi-channel retailers. *The American Economic Review* 107(1):283–303.

Centre for Retail Research (2017) Online Retailing: Britain, Europe, US and Canada 2017. Available at http://www.retailresearch.org/onlineretailing.php.

Chan CW, Yom-Tov G (2014) Balancing admission control, speedup, and waiting in service systems. *Columbia Business School New York, NY. Working Paper.*

Che YK (1996) Customer return policies for experience goods. *The Journal of Industrial Economics* 17–24.

Chen B, Chen J (2017) When to introduce an online channel, and offer money back guarantees and personalized pricing? *European Journal of Operational Research* 257(2):614–624.

Chen S, Sun Y, Kozat UC, Huang L, Sinha P, Liang G, Liu X, Shroff NB (2014) When queueing meets coding: Optimal-latency data retrieving scheme in storage clouds. *INFOCOM, 2014 Proceedings IEEE*, 1042–1050 (IEEE).

Cherlin EJ, Carlson MD, Herrin J, Schulman-Green D, Barry CL, McCorkle R, Johnson-Hurzeler R, Bradley EH (2010) Interdisciplinary staffing patterns: do for-profit and nonprofit hospices differ? *Journal of palliative medicine* 13(4):389–394.

Chu W, Gerstner E, Hess JD (1998) Managing dissatisfaction: How to decrease customer opportunism by partial refunds. *Journal of Service Research* 1(2):140–155.

CMS (2012) Hospice quality reporting program voluntary reporting period: Executive summary of findings. URL https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Hospice-Quality-Reporting/Downloads/ExecutiveSummaryHospiceQRPVoluntaryReportingPeriod.pdf.

CMS (2014) Update to hospice payment rates, hospice cap, hospice wage index, quality reporting program and the hospice pricer for fiscal year (FY) 2015. URL https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNMattersArticles/Downloads/MM8876.pdf.

CMS (2015) Hospice quality reporting. URL https://www.cms.gov.

CMS (2017) Hospice Compare datasets. URL https://data.medicare.gov/data/hospice-compare.

Davis A, Mehrotra S, Friedewald J, Ladner D (2013) Characteristics of a simulation model of the national kidney transplantation system. *Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World*, 2320–2329 (IEEE Press).

Davis S, Gerstner E, Hagerty M (1995) Money back guarantees in retailing: Matching products to consumer tastes. *Journal of Retailing* 71(1):7–22.

Davis S, Hagerty M, Gerstner E (1998) Return policies and the optimal level of hassle. *Journal of Economics and Business* 50(5):445–460.

Durham DD, Rokoske FS, Hanson LC, Cagle JG, Schenck AP (2011) Quality improvement in hospice: adding a big job to an already big job? *American Journal of Medical Quality* 26(2):103–109.

Edelson NM, Hilderbrand DK (1975) Congestion tolls for Poisson queueing process. *Econometrica* 43:81–92.

Flatto L, Hahn S (1984) Two parallel queues created by arrivals with two demands I. *SIAM Journal on Applied Mathematics* 44(5):1041–1053.

Foss S (1989) Comparison of service disciplines in G/GI/m queues. Technical Report RR-1097, INRIA.

Gallino S, Moreno A (2014) Integration of online and offline channels in retail: The impact of sharing reliable inventory availability information. *Management Science* 60(6):1434–1451.

Gallino S, Moreno A, Stamatopoulos I (2016) Channel integration, sales dispersion, and inventory management. *Management Science* .

Gao F, Su X (2016a) Omnichannel retail operations with buy-online-and-pick-up-in-store. *Management Science* .

Gao F, Su X (2016b) Online and offline information for omnichannel retailing. *Manufacturing & Service Operations Management* 19(1):84–98.

Gardner K, Harchol-Balter M, Scheller-Wolf A, Velednitsky M, Zbarsky S (2017) Redundancy-d: The power of d choices for redundancy. *Operations Research* 65(4):1078–1094.

Gardner K, Zbarsky S, Doroudi S, Harchol-Balter M, Hyytiä E, Scheller-Wolf A (2016) Queueing with redundant requests: exact analysis. *Queueing Systems* 83(3-4):227–259.

Guo P, Hassin R (2016) Why do customers place duplicate orders in queues and why do firms allow such behavior? Technical report, Working paper, Hong Kong Polytechnic University.

Harsha P, Subramanian S, Uichanco J (2016) Omni-channel revenue management through integrated pricing and fulfillment planning. Technical report, Working Paper, Ross School of Business, University of Michigan.

Hess JD, Chu W, Gerstner E (1996) Controlling product returns in direct marketing. *Marketing Letters* 7(4):307–317.

Hsiao L, Chen YJ (2014) Return policy: Hassle-free or your money-back guarantee? *Naval Research Logistics (NRL)* 61(5):403–417.

Huang W, Swaminathan JM (2009) Introduction of a second channel: Implications for pricing and profits. *European Journal of Operational Research* 194(1):258–279.

Huskamp HA, Buntin MB, Wang V, Newhouse JP (2001) Providing care at the end of life: do Medicare rules impede good care? *Health Affairs* 20(3):204–211.

Janakiraman N, Syrdal HA, Freling R (2016) The effect of return policy leniency on consumer purchase and return decisions: A meta-analytic review. *Journal of Retailing* 92(2):226–235.

Joshi G, Liu Y, Soljanin E (2014) On the delay-storage trade-off in content download from coded distributed storage systems. *Selected Areas in Communications, IEEE Journal on* 32(5):989–997.

Joshi G, Soljanin E, Wornell G (2015) Efficient replication of queued tasks for latency reduction in cloud systems. *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*, 107–114 (IEEE).

Karp J, Ravi R, Tayur S (2017) Models and methods for omni-channel fulfillment. Working Paper, Carnegie Mellon University.

Kelley AS, Deb P, Du Q, Carlson MDA, Morrison RS (2013) Hospice enrollment saves money for Medicare and improves care quality across a number of different lengths-of-stay. *Health Affairs* 32(3):552–561.

Killaly B, Mukamel D (2007) The cost structure of hospice providers. Technical report, Working paper, University of California, Irvine.

Koole G, Righter R (2008) Resource allocation in grid computing. *Journal of Scheduling* 11(3):163–173.

KPMG (2016) Omnichannel retail survey 2016. Available at `https://assets.kpmg.com/content/dam/kpmg/pdf/2016/02/omnichanel-retail-survey-2016.pdf`.

Krishnan K (1987) Joining the right queue: A Markov decision-rule. *Decision and Control, 26th IEEE Conference on*, volume 26, 1863–1868 (IEEE).

Kumar A, Mehra A, Kumar S (2014a) How physical retail channels impact customers online purchase behavior? Working Paper.

Kumar A, Tandon R, Clancy TC (2014b) On the latency of erasure-coded cloud storage systems. *arXiv preprint arXiv:1405.2833* .

Lal R, Sarvary M (1999) When and how is the internet likely to decrease price competition? *Marketing Science* 18(4):485–503.

Lamas DJ (2014) A private jet is waiting to take you to your kidney transplant. URL `https://www.theatlantic.com/health/archive/2014/10/the-businessman-disrupting-organ-transplantation/381843/`.

Lee K, Pedarsani R, Ramchandran K (2015) On scheduling redundant requests with cancellation overheads. *Proc. of the 53rd Annual Allerton conference on Communication, Control, and Computing.*

Lee N, Kulkarni VG (2014) Optimal arrival rate and service rate control of multi-server queues. *Queueing Systems* 76(1):37–50.

Li B, Ramamoorthy A, Srikant R (2016) Mean-field-analysis of coding versus replication in cloud storage systems. *Computer Communications, IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on*, 1–9 (IEEE).

Li L (1992) The role of inventory in delivery-time competition. *Management Science* 38(2):182–197.

Liang G, Kozat UC (2014) Fast cloud: Pushing the envelope on delay performance of cloud storage with coding. *Networking, IEEE/ACM Transactions on* 22(6):2012–2025.

Liu Z, Nain P, Towsley D (1995) Sample path methods in the control of queues. *Queueing Systems* 21(3-4):293–335.

Lu SF, Lu LX (2016) Do mandatory overtime laws improve quality? Staffing decisions and operational flexibility of nursing homes. *Management Science* 63(11):3566–3585.

Lu SF, Rui H (2017) Can we trust online physician ratings? Evidence from cardiac surgeons in Florida. *Management Science* .

Macy's (2015) Macy's shopper survey yields insights for holiday 2015. Available at `http://investors.macysinc.com`.

Market Force (2017) Nordstrom is nation's favorite fashion retailer, new Market Force study finds. Available at `http://www.marketforce.com/consumers-favorite-apparel-retailers-2017-Market-Force-study`.

Matthews SA, Persico N (2007) Information acquisition and refunds for returns. Working Paper, University of Pennsylvania.

Matthews SA, Persico NG (2005) Information acquisition and the excess refund puzzle. Working Paper, University of Pennsylvania.

McDonald D, Turner S (2000) Comparing load balancing algorithms for distributed queueing networks. *Analysis of Communication Networks: Call Centres, Traffic, and Performance* 28:105.

MEDPAC (2017) Report to the congress: Medicare payment policy. URL `http://www.medpac.gov/-documents-/reports`.

Merion RM, Guidinger MK, Newmann JM, Ellison MD, Port FK, Wolfe RA (2004) Prevalence and outcomes of multiple-listing for cadaveric kidney and liver transplantation. *American Journal of Transplantation* 4(1):94–100.

Moorthy S, Srinivasan K (1995) Signaling quality with a money-back guarantee: The role of transaction costs. *Marketing Science* 14(4):442–466.

Nahmias S (1977) Higher-order approximations for the perishable-inventory problem. *Operations Research* 25(4):630–640.

National Association for Home Care & Hospice (2015) IN FOCUS Part 3: The Proposed FY2016 Hospice Payment Rule HQRP. URL `http://www.nahc.org/NAHCReport/nr150603\_1/`.

Neuts MF (1981) *Matrix-geometric solutions in stochastic models: an algorithmic approach* (Courier Corporation).

NHPCO (2015) FY201 hospice wage index final rule. URL `https://www.nhpco.org/sites/default/files/public/regulatory/RegAlert_FY2016-HospiceWageIndex.pdf`.

NHPCO (2016) NHPCO Facts and Figures: Hospice Care in America .

Ofek E, Katona Z, Sarvary M (2011) "Bricks and clicks": The impact of product returns on the strategies of multichannel retailers. *Marketing Science* 30(1):42–60.

Organ Procurement and Transplantation Network (2015) Waiting list candidates. URL `http://optn.transplant.hrsa.gov/converge/data/default.asp`.

Petersen JA, Kumar V (2009) Are product returns a necessary evil? Antecedents and consequences. *Journal of Marketing* 73(3):35–51.

Puterman ML (1994) Markov decision processes: Discrete dynamic stochastic programming. *New York, NY: John Wiley. doi* 10:9780470316887.

Shah NB, Lee K, Ramchandran K (2016) When do redundant requests reduce latency? *IEEE Transactions on Communications* 64(2):715–722.

Shulman JD, Coughlan AT, Savaskan RC (2009) Optimal restocking fees and information provision in an integrated demand-supply model of product returns. *Manufacturing & Service Operations Management* 11(4):577–594.

Shulman JD, Coughlan AT, Savaskan RC (2011) Managing consumer returns in a competitive environment. *Management Science* 57(2):347–362.

Sparaggis PD, Towsley D (1994) Optimal routing and scheduling of customers with deadlines. *Probability in the Engineering and Informational Sciences* 8(01):33–49.

STAT (2017) Most hospices fare well in first public release of Medicare quality scores. URL `https://www.statnews.com/2017/08/18/hospice-compare-quality-scores/`.

Stock J, Speh T, Shear H (2006) Managing product returns for competitive advantage. *MIT Sloan Management Review* 48(1):57.

Su X (2009) Consumer returns policies and supply chain performance. *Manufacturing & Service Operations Management* 11(4):595–612.

Sun Y, Zheng Z, Koksal CE, Kim KH, Shroff NB (2015) Provably delay efficient data retrieving in storage clouds. *Computer Communications (INFOCOM), 2015 IEEE Conference on*, 585–593 (IEEE).

Transport Topics (2017) Returns of online purchases a key factor in e-commerce boom. Available at `http://www.ttnews.com/articles/returns-online-purchases-key-factor-e-commerce-boom`.

United Network for Organ Sharing (2015) Questions and answers for transplant candidates about multiple listing and waiting time transfer. URL `https://www.unos.org`.

United States Renal Data System (2013) 2013 annual data report: Atlas of chronic kidney disease and end-stage renal disease in the United States. URL `http://www.usrds.org`.

UPS (2014) UPS pulse of the online shopper. A customer experience study. Available at `https://www.ups.com/media/en/2014-UPS-Pulse-of-the-Online-Shopper.pdf`.

UPS (2015) Rethinking online returns. Available at `https://solvers.ups.com/assets/UPS_ReturnsExecutiveSummary.pdf`.

UPS (2016a) 2016 UPS pulse of the online shopper. Available at `https://pressroom.ups.com/assets/pdf/2016_UPS_Pulse%20of%20the%20Online%20Shopper_executive%20summary_final.pdf`.

UPS (2016b) UPS pulse of the online shopper. Tech-savvy shoppers transforming retail. Available at `https://solvers.ups.com/ups-pulse-of-the-online-shopper/`.

van Houtum GJ, Adan I, Wessels J, Zijm WH (2001) Performance analysis of parallel identical machines with a generalized shortest queue arrival mechanism. *OR-Spektrum* 23(3):411–427.

Veeraraghavan S, Debo L (2009) Joining longer queues: Information externalities in queue choice. *Manufacturing & Service Operations Management* 11(4):543–562.

Visschers J, Adan I, Weiss G (2012) A product form solution to a system with multi-type jobs and multi-type servers. *Queueing Systems* 70(3):269–298.

Wachterman MW, Marcantonio ER, Davis RB, McCarthy EP (2011) Association of hospice agency profit status with patient diagnosis, location of care, and length of stay. *JAMA* 305(5):472–479.

White A, Ozminkowski RJ, Hassol A, Dennis JM, Murphy M (1997) The relationship between multiple listing and cadaveric kidney transplantation and the effects of a multiple listing ban. *Transplantation Reviews* 11(2):76–83.

Whitt W (1986) Deciding which queue to join: Some counterexamples. *Operations Research* 34(1):55–62.

Whitt W (2005) Engineering solution of a basic call-center model. *Management Science* 51(2):221–235.

Winston W (1977) Optimality of the shortest line discipline. *Journal of Applied Probability* 181–189.

Wolff RW (1989) *Stochastic modeling and the theory of queues* (Pearson College Division).

WSJ (2013) Rampant returns plague e-retailers. Available at `https://www.wsj.com/articles/rampant-returns-plague-eretailers-1387752786`.

WSJ (2016) Stores gear up for returns – and more shopping. Available at `www.wsj.com/articles/stores-gear-up-for-returnsand-more-shopping-1482769959`.