Optimal Recruitment of Temporary and Permanent Healthcare Workers in Highly Uncertain Environments

Saha Malaki

A THESIS

Submitted to the Faculty of Management

Bayes Business School (formerly Cass), City, University of London

For the Degree of Doctor of Philosophy in Management (Operations and Supply Chain)

Under supervision of:

Dr. Navid Izady

Prof. Lilian M. de Menezes

Dr. Oben Ceryan

September 2022

Abstract

There has been a significant increase in the demand for temporary skilled workers in the health sector. They provide volume flexibility, but are generally more expensive than their permanent counterparts. A balance must therefore be struck between staffing cost and service quality by recruiting the right mix of temporary and permanent healthcare workers. Focusing on periods of highly uncertain demand, in this thesis, we propose optimization models aiming to inform permanent and temporary recruitment decision making for settings in which all patients must be served. We pursue this under two different scenarios, a mid-term planning horizon and a long-term planning horizon.

The first part of the thesis [1] is devoted to recruitment decision making in a mid-term planning horizon. The main trade-off in this case is between recruitment lead times and staffing costs of temporary and permanent workers. More specifically, permanent skilled workers are cheaper for the healthcare provider than equivalent temporary workers, but have a substantially longer recruitment lead time. Longer recruitment lead time of permanent workers implies that providers face a higher level of demand uncertainty when making permanent recruitment decisions and a higher likelihood of not being able to fill the created positions. Considering a single-interval planning horizon, we propose a two-stage stochastic optimization framework to capture this fundamental trade-off. The first stage of our framework identifies the number of permanent positions to advertise, and the second stage determines the number of temporary workers to recruit. Our framework accounts for the uncertainty in the number of permanent vacancies that will be filled, stochasticity of the service delivery process, and imperfect demand information at the time of advertising for permanent positions. Under a general setting of the problem, we characterize the optimal first- and second-stage decisions analytically, propose fast numerical methods for finding their values, and prove some insensitivity and monotonicity properties for the optimal decisions and their corresponding costs. The benefit/loss of delaying the advertisement

[1] This part has been accepted for publication in European Journal of Operational Research.

for permanent positions to obtain a more accurate demand information, at the expense of a higher risk of not filling the advertised positions, is also investigated. A case study based on data from a geriatric ward illustrates the application of our framework to an inpatient department, and further managerial insights are developed using a combination of analytical and numerical results.

The second part of the thesis is dedicated to recruitment decision making in a long-term planning horizon. In addition to the different staffing costs and recruitment lead times of temporary and permanent workers captured in the first part, we consider the difference in their placement durations. This is because permanent workers have substantially longer contracts which may cover periods of low demand, hence in the long run, they are likely to be more expensive to the provider than temporary workers. We capture this by a multi-interval optimization framework which involves a two-stage decision making, similar to the two-stage decision making of the first part, repeated in each interval. The time-varying nature of demand over different intervals is also incorporated into this framework. Using a Markov decision process formulation, we prove that the optimal recruitment policy for permanent healthcare workers in this context has a hire-up-to structure. Numerical experiments then investigate the sensitivity of the hire-up-to value to different system parameters. The potential benefits of using the long-term (multi-interval) recruitment model as compared to the mid-term (single-interval) recruitment model is also evaluated numerically.

Author's Declaration

I, Saha Malaki, declare that this thesis entitled "Optimal Recruitment of Temporary and Permanent Healthcare Workers in Highly Uncertain Environments" is the result of my own work, under the guidance of my PhD supervisors, Dr. Navid Izady, Prof. Lilian M. de Menezes, and Dr. Oben Ceryan. I also certify that the work in this thesis has not been submitted for the award of a higher degree anywhere else.

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Navid Izady, for his time, constructive advice, honest feedback, and kindness. Over these years, he set a real example of academic and research integrity for me. Through his manner, he taught me a great deal about the qualities of a true teacher. I would also like to thank my co-supervisor, Prof. Lilian M. de Menezes, for her guidance, patience, and continuous support. I am very grateful not only for her scientific supervision and active involvement in every aspect of this work, but also for all her sympathetic encouragement. Finally, I would like to thank my other co-supervisor, Dr. Oben Ceryan, for his time, insightful vision, and enthusiastic guidance. I learned a lot as their student and without their help, my PhD studies would not have been as enjoyable as it was.

I would like to sincerely acknowledge my examiners Prof. Inneke Van Nieuwenhuyse and Dr. Dimitris Paraskevopoulos for their time to review my thesis and for providing valuable comments.

I am deeply thankful to Malla Pratt, research programmes operation manager, and Abdul Momin, PhD admission officer, for assisting me in many different ways. Their support has made my academic life at Bayes a joyful experience. During my studies, I was fortunate enough to know and become friends with many great people: Fabienne, Bahar, Parastoo, Adi, and many other friends and colleagues at Bayes Business school, you made my PhD a delightful journey.

My sincere appreciation goes to my parents, Maryam and Saeed. Their unconditional love, encouragement, and guidance in every step of my life have always been the true source of inspiration for me. I am thankful to my brother and best friend, Sahand, who I cannot imagine the world without. At the end, my deepest loves go to my husband, Kiarash. He was always around and constantly motivated me when the research seemed difficult and overwhelming. This work would not exist if it was not for his extreme love, care, and understanding.

To my parents, Maryam and Saeed,

to my husband, Kiarash,

and to my brother, Sahand.

Contents

List of Tables

List of Figures

List of Algorithms

List of Abbreviations

CDF	**C**umulative **D**istribution **F**unction
CV	**C**oefficient of **V**ariation
FTE	**F**ull-**T**ime **E**quivalent
HCW	**H**ealth **C**are **W**orker
HUDP	**H**ighly **U**ncertain **D**emand **P**eriod
i. i. d	**I**ndependent and **I**dentically **D**istributed
MC	**M**arkov **C**hain
MDP	**M**arkov **D**ecision **P**rocess
MMFE	**M**artingale **M**odel of **F**orecast **E**volution
MRMS	**M**ulti-**R**esource **M**ulti-**S**erver
NHS	**N**ational **H**ealth **S**ervice
PDF	**P**robability **D**istribution **F**unction
SRMS	**S**ingle-**R**esource **M**ulti-**S**erver
SRSS	**S**ingle-**R**esource **S**ingle-**S**erver
UK	**U**nited **K**ingdom

List of Notations

Symbol	Description
Parameters	
T	length of planning horizon in long-term recruitment model
t	given advertisement epoch in mid-term recruitment model
t_e	beginning of the HUDP
t'	delayed advertisement epoch
N	number of segments during the HUDP
c	penalty cost for the remained permanent workers at the end of the planning horizon
c_g, c_p, c_o	cost rates of temporary, permanent, and mandatory overtime work
c_w	waiting cost incurred by the patients per unit of time in the system
r_o	percentage of mandatory overtime work by permanent HCWs
n	number of permanent HCWs available at time t who are expected to remain in their jobs during the HUDP
p	total number of permanent HCWs in the system at time t_e
s	number of servers available for service delivery
λ	exact value of demand rate during the HUDP
$\tilde{\lambda}(p)$	unique root of function $\phi_p(x)$ given in (3.4) in the interval $(0, p(1+r_o))$
$\tilde{g}(\lambda, p)$	unique root of function $\theta_{\lambda,p}(g)$ given in (3.5) in the interval $((\lambda - p(1+r_o))^+, \infty)$
α_1, α_2	small positive numbers used in Algorithm 1
λ_u, q_u	upper-bound for the support of h_t and f_t
$\tilde{a}(n)$	unique root of function $\psi_n(a)$ given in (3.10) in the interval $[0, \infty)$
a_u	positive number used in Algorithm 2
ξ	mean demand rate
κ	coefficient of variation for the (random) demand rate
μ_r	mean number of qualified applications received during $(t, t_e]$
κ_r	coefficient of variation for the (random) number of qualified applications
τ	coefficient of variation of the service time
γ	minimum probability of the system being stable in single-stage decision making
B	number of beds in the geriatric ward of our case study

$\lambda^{(b)}$	rate of patient arrival to the bed queueing system
$\mu^{(b)}$	service rate for the bed queueing system
$\lambda^{(n)}$	rate of patients' regular requests arrival to the nursing queueing system
$\mu^{(n)}$	service rate of the regular requests in the nursing queueing system
$\xi^{(b)}$	mean of patients arrival rate to the bed queueing system
$\kappa^{(b)}$	coefficient of variation of patients arrival rate to the bed queueing system
$\boldsymbol{M}$	mean of the multivariate demand rate distribution
$\boldsymbol{\Sigma}$	covariance matrix of the multivariate demand rate distribution
$\rho_{(j,k)}$	correlation between the demand of segments j and k of the HUDP
σ_j	standard deviation of the demand in segment j of the HUDP
α	discount factor
$\boldsymbol{Q}$	transition probability matrix
i	state of the demand rate distribution with state space $\mathcal{S} = \{0, 1, \cdots, k\}$
p_t	total number of permanent HCWs at the beginning of the HUDP of interval t
n_t	number of permanent HCWs in the system before the permanent recruitment decision at the beginning of interval t
λ_t	exact value of demand rate during the HUDP of interval t
ξ^i	mean demand rate at state i
κ^i	coefficient of variation for the (random) demand rate at state i
Decision Variables: chapter 3	
a	number of permanent positions to advertise at time t
g	number of temporary HCWs to recruit at time t_e
$a^*(n)$	optimal number of permanent positions to advertise in the first stage given n
$g^*(\lambda, p)$	optimal number of temporary workers to recruit in the second stage given arrival rate λ and p permanent workers
Decision Variables: chapter 4	
a_t	number of permanent workers to recruit at the beginning of each interval t
g_t	number of temporary workers to recruit at the beginning of each interval t
$a_t^*(n_t, i)$	optimal number of permanent worker to recruit in each interval t given n_t and state i
$g_t^*(\lambda_t, p_t)$	optimal number of temporary workers to recruit in each interval t given arrival rate λ_t and p_t permanent workers
$p_{t,i}^*$	hire-up-to value given time interval t and state i
Random Variables	
Q_t	random number of qualified applications received during the advertisement period with pdf f_t and cdf F_t
Λ_t	random demand rate as predicted at time t with pdf h_t and cdf H_t
$\Lambda_t^{(b)}$	random patient arrival rate to the bed queueing system as predicted at time t
$\boldsymbol{\Lambda}_t$	vector of random demand rates for different segments of HUDP as predicted at time t with joint pdf $h_t(\lambda^1, \cdots, \lambda^N)$
Λ^i	random demand rate at state i with pdf h^i

X_t	random number of qualified applications received during the permanent recruitment period of interval t
Functions	
$v(\lambda, p)$	optimal second-stage cost given arrival rate λ and p permanent workers
$m(n)$	optimal first-stage cost given n existing permanent workers
$l(\lambda, s)$	mean number of requests in the system given arrival rate λ and s servers
$C(\lambda, s)$	continuous extension of the Erlang delay function given in (3.33)
$l^{(n)}(\lambda^{(b)}, s)$	mean number of requests in the nursing system given patient arrival rate λ^b and s nurse
$\phi_p(x)$	function given in Equation (3.4)
$\theta_{\lambda,p}(g)$	function given in Equation (3.5)
$\psi_n(a)$	function given in Equation (3.10)
$\psi_n^{h_t^j}(a)$	function given in Equation (3.44), where h_t^j is the marginal pdf of Λ^j

Chapter 1

Introduction

1.1 Motivation

In the past few decades, the healthcare sector has witnessed significant changes in the way that jobs are structured and work is organized. Chronic staff shortages (Bae et al., 2010), long lead times in recruiting permanent staff (Lu and Lu, 2017), predictable and unpredictable variabilities in patient demand (Seo and Spetz, 2013), and rising absenteeism and turnover among permanent staff (West et al., 2020) have led to a substantial increase in the use of temporary healthcare workers (HCWs) worldwide. In the UK, for example, the total hours of temporary nurses requested by the hospitals within the National Health Service (NHS) doubled from 2011 to 2015 (National Audit Office, 2016).

Temporary workers are a flexible workforce with short-term employment and variable

working hours (Kesavan et al., 2014). They provide volume flexibility, i.e., the ability to adjust the staffing patterns flexibly and quickly in response to variations in patient demand and (un)availability of permanent staff (Qin et al., 2015). However, temporary skilled workers earn higher wages than their permanent counterparts, hence are generally more expensive to the provider. In fact, findings from a recent survey suggest that savings of about half a billion pounds could have been made in the UK's NHS during 2018 if the hours worked under temporary contracts had been covered by permanent staff (The Open University, 2018). There is also mixed evidence in regards to temporary HCWs performance, with some studies linking undesirable outcomes to their deployment (e.g., Roche et al., 2009), and some other studies refuting such links (e.g., Aiken et al., 2013). We review this evidence below.

Analyzing the data from a sample of Canadian hospitals, Estabrooks et al. (2005) conclude that there is a positive correlation between the use of temporary nurses and 30-day in-hospital mortality. However, the statistical analysis conducted by Aiken et al. (2013) on data from a wide range of US hospitals reveals that this correlation becomes insignificant when the quality of work environment is taken into account. Significant associations between temporary staff use and adverse events are also reported in the literature; see, for example, Aiken et al. (1997), Bae et al. (2014), Stratton (2008), and Roseman and Booker (1995). The study by Aiken et al. (2007) on Pennsylvania hospitals, however, indicates that these associations are rendered insignificant after controlling for staffing levels and resources' adequacy. Overall, it can be argued that it is the quality of the working environment in

(some) hospitals with many temporary HCWs, rather than the actual use of temporary workers, which may be linked to undesirable outcomes.

Disruption in continuity of care is another concern against the use of temporary HCWs, see, e.g., Roche et al. (2009) and Cabana and Jee (2004). However, Aiken et al. (2007) argue that continuity of care in hospitals is not ideal, regardless of the use of temporary workers. Discontinuity, as they suggest, is the outcome of 12-hour shifts in hospitals resulting in nurses working 3 to 4 days per week. Bae et al. (2010) contend that a higher use of temporary HCWs increases the administrative burden as such workers may be unfamiliar with policies, procedures, equipments, colleagues, and patients, thus requiring more supervision. However, the interviews conducted by Berg Jansson and Engström (2017) in a Swedish intensive care unit suggest that temporary workers are more likely to develop closer relationships with patients as, in most cases, they do not have to participate in planning or internal training. The authors also explain that, since clear documentation is needed in systems with blended workforce (that is, where both temporary and permanent HCWs are hired), the additional administrative burden results in greater transparency and better communication, which can improve performance.

Our review of the literature as outlined above suggests that the use of temporary workers does not influence the performance negatively in general. The analyses of Aiken et al. (2007) and Xue et al. (2012) also show that temporary nurses are as well educated as permanent nurses, and that nurses in hospitals with larger shares of temporary staff are not more likely to be dissatisfied with their jobs or more burned out. Furthermore, given an appropriate

environment, temporary HCWs can have positive impacts. For example, their use is linked with higher efficiency in the study by Hughes and Marcantonio (1991). As such, healthcare providers need a mix of permanent and temporary HCWs to be able to deliver a quality service in a timely and efficient manner. However, a balance must be struck between staffing costs and service quality by recruiting the right mix of permanent and temporary HCWs.

Finding the right mix of temporary and permanent HCWs is challenging for the following reasons. First, permanent and temporary recruitment decisions are not made at the same time; advertising for permanent workers must typically start well ahead of the service delivery, e.g., a few months in advance, whereas recruitment of temporary workers occurs much later, e.g., a few hours/days in advance. This implies an asymmetry in demand information, i.e., a more accurate demand information is available at the time of temporary recruitment than permanent recruitment. Second, there is uncertainty in recruitment since there is no guarantee that all the required positions can be filled. Third, healthcare providers often experience periods of highly uncertain demand. In the UK's NHS, for example, there is high uncertainty in predicting winter peak demand, with some years such as 2017 having a substantially busier winter than previous years (NHS Improvement, 2018). The recent COVID-19 pandemic has added to demand uncertainty. For example, as illustrated in Thorlby et al. (2020), the emergency care demand in the UK dropped significantly during the first wave of the pandemic (March to June 2020), while it peaked back up in the second wave (September to December 2020). In addition to making the role of demand forecasting and the timing of permanent advertisement more critical, a

highly uncertain demand implies that permanent employees recruited in one year may not be needed in the following year.

In light of the challenges outlined above, the aim of this thesis is to develop optimization frameworks to inform permanent and temporary recruitment decision making for highly uncertain demand periods. We pursue this aim under two different planning horizons, mid term and long term. The rest of this chapter is organized as follows. The thesis objectives and scope are outlined in Section 1.2. The structure of the thesis is presented in Section 1.3.

1.2 Thesis Objectives and Scope

For a mid-term planning horizon, e.g., one year ahead, the optimal blend of temporary and permanent HCWs mainly depends on the following trade-off: permanent HCWs are cheaper, but their recruitment lead time, i.e., the time between advertisement and recruitment, is substantially longer. Longer recruitment lead times for permanent HCWs have two implications for recruitment decision making: limited information about demand is available when permanent positions are advertised; and some (or even all) of these positions may not be filled in the desired time frame. Indeed, 10% of permanent nursing vacancies in the NHS were not filled in 2020 (NHS Vacancy Statistics, 2021). For a long-term planning horizon, e.g., multiple years ahead, the optimal mix of temporary and permanent HCWs is further influenced by the difference in their placement durations. More specifically, permanent

HCWs have substantially longer contracts to the service provider than temporary HCWs, hence the provider must consider their long-term cost when making recruitment decisions. This is particularly important under highly uncertain demand.

The core of this thesis is divided into two main chapters: Chapter 3, which addresses the mid-term recruitment problem, and Chapter 4, which considers the long-term recruitment problem. More specifically, given a single-interval planning horizon, a framework is proposed in Chapter 3 which informs decision making for temporary and permanent workforce recruitment by capturing the trade-off between their staffing costs and recruitment lead times. In Chapter 4, on the other hand, considering a multi-interval planning horizon, a framework is proposed that informs recruitment decision making by capturing the trade-off between staffing costs, recruitment lead times, and contract durations. To simplify the analysis, the randomness of the permanent recruitment process, which is incorporated into the framework of Chapter 3, is excluded from the framework of Chapter 4.

In summary, we investigate the following research questions:

- How the trade-off between recruitment lead times, staffing costs, and placement durations can be captured by stylized analytical frameworks?
- What insights can be derived from such frameworks with regards to the impact of cost elements, demand uncertainty, timing of permanent advertisement, and service uncertainty on the optimal permanent and temporary recruitment decisions and the overall system cost?

- How much savings can potentially be gained by using our proposed frameworks?

1.3 Thesis Structure

An overview of the analytical approaches proposed in the literature for the mid- and long-term recruitment problems is presented in Chapter 2. In Chapter 3, we propose a single-interval stochastic optimization framework for the mid-term recruitment problem. We characterize the optimal recruitment decisions, propose algorithms for calculating their values, derive some monotonicity and insensitivity properties for the optimal decisions and their corresponding costs, and illustrate the implementation of our framework with data from an inpatient ward. Numerical experiments are also conducted to evaluate the likely savings obtained from our proposed framework as compared to two simplified models. While our framework in Chapter 3 is mainly based on a single opportunity for temporary recruitment, an extension is also proposed which considers multiple opportunities for temporary recruitment.

In Chapter 4, we propose a multi-interval stochastic optimization framework for the long-term recruitment problem. We characterize the structure of the optimal policy, perform sensitivity analysis, and evaluate the potential savings that can be obtained from adopting the multi-interval framework as compared to the single-interval framework developed in Chapter 3. Final conclusions, a summary of the contributions, and future areas for research are discussed in Chapter 5.

Chapter 2

Literature Review

2.1 Introduction

In this chapter, an overview of previous works falling within the scope of this thesis is presented. The purpose of this chapter is twofold. First, to introduce recruitment models for blended workforce settings facing uncertain demand. Second, to be more specific about the focus of this research and justify the frameworks developed in the thesis. This review is not bound to healthcare systems as other blended workforce environments are also considered.

We start with an overview of the mid-term recruitment models in §2.2. These models are categorized into single-stage and two-stage optimization models, which are reviewed in §2.2.1 and §2.2.2, respectively. Long-term recruitment models are then reviewed in §2.3. In §2.4, we link this research to dual sourcing problems in the inventory management literature.

§2.5 outlines the gaps in the literature, leading to the focus of this research. This chapter concludes in §2.6.

2.2 Mid-term Recruitment Models

The first part of our review is devoted to analytical recruitment models seeking to determine the optimal mix of temporary and permanent workers assuming a single-interval (typically one year) planning horizon. The studies covered assume that there exist a single opportunity for recruiting permanent employees at the beginning of the planning horizon, and (potentially) multiple opportunities for recruiting temporary employees at given epochs during the planning horizon. Two main streams can be identified in this literature. The first stream uses single-stage optimization models (e.g., Dong and Ibrahim, 2020), whereas the second stream uses two-stage optimization models (e.g., Hu et al., 2021b). Single-stage models assume simultaneous recruitment of permanent and temporary HCWs, thus ignoring the higher levels of demand uncertainty for permanent recruitment. Two-stage models, however, assume that the permanent recruitment decision is made in the first-stage under a limited demand information, and temporary recruitment decision is made in the second-stage when a more accurate demand information is available. We note that a comprehensive review covering the studies up to 2010 is provided in Qin et al. (2015).

2.2.1 Single-Stage Models

Abraham (1988) uses a combination of analytical and empirical studies to investigate the employers' motivations for using temporary workers in a manufacturing setting. In their analytical study, a model is proposed for identifying the number of permanent workers to hire so as to minimize the expected labour cost, assuming that unmet demand will be covered by temporary workers. Demand for output is represented by a probability distribution, and each permanent (temporary) worker is assumed to be able to produce one (less than or equal to one) unit of output. Supply uncertainty is modeled by assuming that a random fraction of permanent workers will not show up. The analytical solutions lead to the hypotheses that variability in demand and uncertainty in the availability of permanent workers are the employers' main motivations for using temporary workers, which are then supported by an empirical study.

Berman and Larson (1994) formulate a model wherein the number of permanent workers is fixed and the optimal pool size of temporary workers at the beginning of each month must be determined. They assume that temporary workers are guaranteed a minimum number of hours per month as an incentive for pool membership. The cost per hour of temporary workers is assumed to be lower than the cost per hour of overtime work, thus the preference is to first use temporary staff and then resort to overtime work. Berman and Larson (1994) represent daily demand for workforce by a probability distribution, and capture the supply uncertainty caused by employees' absenteeism. Their analysis leads to

an exact model for identifying the optimal pool size of temporary workforce.

Jeang (1996) proposes a mixed-integer programming model to determine the number of permanent workers to hire on a weekly basis, as well as the number of temporary and overtime workers needed for each day, in a healthcare setting with uncertain demand. The workforce demand is captured by a probability distribution, and a constraint is included in the optimization model ensuring that the labour supply must exceed the demand average plus/minus a multiple of demand standard deviation. A heuristic enumeration approach is proposed for estimating the optimal number of permanent employees.

Harper et al. (2010) propose a simulation-based model to find the optimal number of different types of permanent nurses to hire at the start of the year, as well as the optimal daily number of temporary nurses. Their model takes the daily demand for nurses from a discrete-event simulation, and converts it to staffing numbers using either the nurse-to-patient ratio method or the dependency-activity-quality method. The former estimates the staffing numbers based on a fixed proportion of occupied beds (de Véricourt and Jennings, 2011), while the latter establishes the number of nurses required based on the care needed for patients of different dependency levels in a ward (Hurst, 2002). The results suggest that increasing the number of permanent nurses is more cost-effective for coping with fluctuations in demand than using temporary nurses.

Dong and Ibrahim (2020) propose stochastic models for situations in which the manager must decide on the optimal numbers of temporary and/or permanent workers so as to minimize the sum of staffing cost and performance cost (including the costs of patients

waiting or leaving without being served). The authors consider two modelling scenarios. In the first one, the number of temporary staff is decided in each period, and in the second one, the number of permanent staff for the entire planning horizon and the number of temporary staff for each period are decided at the beginning of the horizon. The dynamics of service delivery is captured using an abandonment queueing model with a random number of servers, and the demand is assumed to be Poisson with a known and time-varying rate for different periods. Due to analytical intractability of the staffing problem with a random number of servers, the authors consider an asymptotic, many-server, mode of analysis. For the first scenario, four regimes are identified for the optimal policy depending on the magnitude of the variability of the random number of servers. For the second scenario, it is illustrated that when temporary workers are more expensive, the best strategy is to rely solely on permanent workers in low-demand periods, and to use both types of workers in high-demand periods.

The main shortcomings of the models outlined above are as follows. Some studies capture only the permanent recruitment decision (Abraham, 1988) or the temporary recruitment decision (Berman and Larson, 1994) explicitly in their models. Some others (Jeang, 1996; Harper et al., 2010; Dong and Ibrahim, 2020) capture both decisions but assume that they are made simultaneously, and therefore ignore the asymmetry in demand information at the times of permanent and temporary recruitment. These shortcomings are addressed in the two-stage models.

2.2.2 Two-Stage Models

Kao and Queyranne (1985) present a two-stage model for minimizing the overall nursing cost in inpatient departments of a hospital. They divide a yearly planning horizon into periods of equal length, e.g., a month. In the first stage, the number of permanent nurses in each skill class for the entire year is determined. In the second stage, given the number of permanent nurses and the realized demand in each period, the numbers of overtime and temporary nurses to utilize are identified. Patient arrival is modelled as a Normal distribution with a time-varying mean and variance, which are evaluated using an autoregressive integrated moving average model. This is then converted to nursing hours by the nurse-to-patient ratio method, taking into account the patients' random length of stay. The authors propose different simplifications of their model, and demonstrate that, while ignoring the time-varying nature of demand does not lead to substantial errors in nursing estimates, ignoring the demand uncertainty leads to underestimation.

Similarly, Pinker and Larson (2003) divide a yearly planning horizon into periods of equal length. Their model determines the number of permanent workers to hire and the pool size of temporary workers to contract (from a temporary labour supplier) over the planning horizon in order to minimize the expected labour and backlog costs (any unprocessed work in a period is assumed to be backlogged to the next period.) Embedded within their model is a Markov decision process that gives the amount of temporary and overtime workers to utilize in each period. To capture the impact of timing of demand information, the authors

split each period into two segments, odd and even, and model the demand for workforce as the sum to two random variables, one that is realized at the beginning of the odd segments and the other that is revealed at the beginning of the even segments. In an odd segment, based on the revealed demand variable and the number of present permanent workers, the number of temporary workers to use out of the contracted pool is obtained. In an even segment, based on full demand information, the amount of overtime work required from permanent workers is established. The model accounts for different productivity levels of distinct types of workers as well as absenteeism. Their results illustrate that labour flexibility on its own does not provide a better performance, and appropriate demand information is required as a complementary tool.

Lu and Lu (2017) develop a two-stage stochastic optimization model to capture the effect of mandatory overtime laws on staffing ratios in nursing homes. In the first stage of their model, facing an uncertain patient enrolment, the optimal regular hours of registered nurses is determined. In the second stage, given the actual patient enrolment and the regular nursing hours, the optimal contract and overtime nursing hours are decided. The uncertainty in patient enrolment is captured by a random variable whose distribution is known in the first stage, and its exact value is revealed in the second stage. The objective is to minimize the total staffing cost, while ensuring a minimum staff-to-resident ratio. Results from their model support the following hypotheses: (i) mandatory overtime laws increase (decrease) the nursing hours of contract (permanent) workers; (ii) more (fewer) staffing hours of contract (permanent) workers are associated with a lower quality of care;

and (iii) mandatory overtime laws diminish the quality of care in nursing homes.

Recently, Hu et al. (2021b) propose a two-stage model in which the first stage identifies the base staffing levels (i.e., permanent workforce) and the second stage determines the surge staffing levels (i.e., overtime and temporary workforce). They capture the dynamics of service delivery explicitly by an abandonment queueing model, and represent demand as a Poisson mixture model, i.e., a Poisson process with a random rate. They assume that the distribution of the Poisson rate is known in the first stage, and the exact value of the rate is revealed in the second stage. Assuming that the random demand rate follows a specific form, they show that surge staffing is most beneficial when demand rate uncertainty dominates the system stochasticity (as driven by random inter-arrival, service and abandonment times). Taking an asymptotic approach that increases the system scale to infinity, they propose near-optimal two-stage staffing rules minimizing the sum of staffing and performance costs. The authors extend their model to allow for the rate prediction error in the second stage, and also make empirical adjustments to their staffing rules to facilitate their implementation in an emergency department.

Table 2.1 summarizes the models developed in the mid-term recruitment literature.

Table 2.1: Summary of the references in mid-term recruitment models literature

References	Decisions	Features	Method	Context
		Single Stage Models		
Abraham (1988)	# of permanent workers	Demand and supply uncertainty	Stochastic optimization	Manufacturing
Berman and Larson (1994)	Pool size of temporary workers	Demand and supply uncertainty, overtime work	Stochastic optimization	Postal services
Jeang (1996)	# of permanents weekly, # of temporaries and overtime daily	Demand uncertainty	Mixed-integer programming	Healthcare
Harper et al. (2010)	# of permanents for the year, # of temporaries daily	Staffing levels as the output of simulation (using nurse-to-patient ratio or dependency-activity-quality methods)	Stochastic optimization	Healthcare
Dong and Ibrahim (2020)	# of temporary and/or permanent workers	Abandonment queueing system, Poisson arrivals with time-varying rate, random servers	Fluid and stochastic-fluid approximations	Service systems
		Two-Stage Models		
Kao and Queyranne (1985)	# of permanents in the first stage and # of overtime and temporaries in the second stage	Demand uncertainty	Stochastic optimization	Healthcare
Pinker and Larson (2003)	# of permanents and # of temporaries to contract for the entire planning horizon, # of temporaries to use in the first stage and # of overtime to use in the second stage within each interval	Demand and supply uncertainty, backlog of unfinished work	Stochastic optimization using dynamic programming	Delivery service, clerical operations, repair services
Lu and Lu (2017)	# of permanents in the first stage, # of temporaries and overtime in the second stage	Nurse-to-patient ratio model, demand uncertainty	Stochastic optimization	Nursing homes
Hu et al. (2021b)	# of permanent workers in the first stage and # of temporary workers in the second stage	Abandonment queueing system, Poisson mixture arrivals	Stochastic-fluid approximation	Emergency department

2.3 Long-term Recruitment Models

The planning horizon in long-term recruitment models spans across multiple years. This implies that, in contrast with mid-term recruitment models, there are multiple opportunities for hiring and potentially firing permanent employees. We review this literature below.

Gans and Zhou (2002) study a long-term hiring problem with heterogeneous workforce (different skill levels), in which the decision maker must identify the optimal number of workers to hire in each interval so as to minimize the sum of staffing, hiring and operating costs. Employees are hired at skill level 1, and progress through a sequence of skill levels over time. The problem is formulated as a Markov Decision Process (MDP), with the number of employees at different skill levels (before a hiring decision is made) identifying the state of the system at the beginning of each interval. In addition to fixed hiring costs and per-period wages, the value function of the MDP includes an operating cost function, which is the solution to a work scheduling optimization model given the numbers of employees at different skill levels and a (point or distributional) forecast of demand. This function can potentially capture the amount of overtime and outsourcing that must be done in each interval to meet the demand. The authors prove that the optimal policy is of a *hire-up-to* type, assuming that the operating cost is a convex function of the numbers of employees at different levels. They also show that when demand is stationary or stochastically increasing, a myopic policy which optimizes a one-period static problem for each interval is optimal. The authors further propose a computationally efficient heuristic approach which produces

near-optimal policies when there is little learning and relatively inexpensive flexible capacity.

Ahn et al. (2005) investigate a similar problem, but allow for hiring decisions to be made at different levels. They also account for the possibility of employees being fired in each interval. In addition to the current numbers of workers of different levels, a state variable representing the current state of the environment is considered. This additional variable could affect the distribution of the demand, the pool from which the hiring is made, and/or the probabilities of employees leaving the system. The authors prove that when the numbers of employees are non-integral, hiring and firing costs are linear, the operating cost function is convex, and a random fraction of employees of each type may leave at the end of each interval, the optimal policy has a simple *hire-up-to/fire-down-to* structure. However, the optimal policy is more difficult to characterize when the numbers of employees must be integer.

2.4 Analogy with Dual Sourcing in Inventory Management

There are connections between our recruitment decision-making problem and the dual-sourcing problem in inventory management (see Svoboda et al., 2021 for a comprehensive review on multiple sourcing). In this context, firms must decide, in a newsvendor setting, how much to procure from short lead-time suppliers (i.e., *emergency* suppliers) and how much from long lead-time suppliers (i.e., *regular* suppliers). With regular suppliers, retail

prices are typically lower but orders must be placed well in advance of the selling season when demand forecasts are highly inaccurate. The firms may, therefore, delay some orders to closer to the time of sales when more accurate demand information is available, but this involves using more expensive emergency suppliers. We can think of temporary and permanent workers in the service delivery context as emergency and regular suppliers, respectively, in the inventory management context. We therefore review studies on inventory management that investigate the optimal quantity of orders from a portfolio of short and long lead-time suppliers, and assess the similarities and potential differences with our research.

Gurnani and Tang (1999) consider a system with uncertain demand and two ordering opportunities wherein the product unit cost in the second opportunity could be lower or higher than that of the first opportunity. They assume that the demand follows a given distribution in the first opportunity, and this distribution is updated before the second opportunity in light of the most recent market information. They formulate the problem as a nested newsvendor model to identify the optimal order quantities at the two ordering opportunities. Given specific assumptions for updated demand distribution, they characterize the situations under which it is optimal to delay the ordering until the second opportunity.

Yan et al. (2003) study the same problem as in Gurnani and Tang (1999) but with two simplifications: (i) they assume the unit price in the second ordering opportunity is given and always larger than that of the first one, and (ii) the demand forecast standard deviation reduces as a linear (and potentially random) function of time. This enables them

to fully characterize the optimal ordering policy and also evaluate the marginal benefit of improved forecasting. They further generalize the model to multiple intervals and show that the policy is myopically optimal for some demand distribution functions.

Wang and Tomlin (2009) develop a model in which the firm has only a single opportunity to place orders but it can be any time before the selling season. The unit price is assumed constant but lead-time is random with a known distribution. The trade-off is therefore between ordering early to reduce the risk of late arrival, and ordering late to improve demand forecast accuracy. The dynamics of demand forecasting is captured by a martingale model of forecast evolution (MMFE) where successive forecasts have a Markovian property. The authors characterize the optimal procurement time and quantity, and prove that, with a multiplicative MMFE model, the timing decision is independent of forecast evolution but the quantity is not.

Wang et al. (2012) study a system in which the unit purchase price and forecasting accuracy increase as we get closer to the start of the selling season. They propose two different ordering strategies, single ordering and multi-ordering. In single ordering, the firm is restricted to a single ordering opportunity. The decision maker must decide when to order and how much to order. The single-ordering strategy is divided into static and dynamic models. In the static model, the timing decision is made at the start of the planning horizon but the quantity decision is delayed until the selected time. In the dynamic model, the decision maker must decide in each period whether to order (if she has not ordered before) or wait, and how much to order (if she has decided to order). The authors propose analytical

solutions for both dynamic and static models. In multi-ordering, the firm can order multiple times in response to the available stock and most up-to-date demand information. The optimal policy in the multi-ordering strategy is proved to be a *base-stock* policy, where the optimal base-stock level is a function of the current demand information.

The trade-offs considered in the frameworks we develop in Chapters 3 and 4 of this thesis are similar to those of the inventory models cited above. Our context is, however, fundamentally different. This is because (i) we are recruiting employees, not ordering products, so the dynamics and uncertainties of recruitment must be captured explicitly in modelling, and (ii) measuring performance in service delivery is more complex than in selling products as it is a function of interactions between random inter-arrival and service times. Hence, a completely different setup is needed in our models.

2.5 The Focus of the Thesis

Mid-term Recruitment Models. Chapter 3 focuses on the trade-off between staffing costs and recruitment lead times of temporary and permanent HCWs in a mid-term planning horizon. Given the asymmetry of demand information at the points of permanent and temporary recruitment, we follow a two-stage modelling approach in this chapter. The two-stage models in the literature do not typically capture the dynamics of service delivery explicitly; they either work directly with the workforce demand distribution (e.g., Pinker and Larson, 2003), or assume a linear relationship between demand and the number of servers required

to meet this demand (e.g., Lu and Lu, 2017). The former is difficult to measure in service environments, and the latter amounts to the nurse-to-patient ratio method in healthcare. As argued in Yankovic and Green (2011), the nurse-to-patient ratio method may lead to under- or over-staffing as factors such as the unit size and the variability in service durations are not explicitly accounted for. The work of Hu et al. (2021b) is the only two-stage study that captures the dynamics of service delivery explicitly, hence it is the closest study to our research. Our study is different from that of Hu et al. (2021b) in the following ways.

First, Hu et al. (2021b) (and other two-stage studies reviewed in §2.2.2) assume that the desired permanent staffing level can always be achieved, whereas we will consider the uncertain nature of permanent recruitment and thus account for the possibility of some positions not being filled. In addition to making our models more realistic, this would allow us to investigate the benefit of a lower demand uncertainty as a result of a later permanent advertisement than scheduled, versus the associated risk of a shorter advertisement window. Second, Hu et al. (2021b) focus on abandonment queues. This implies that some customers may leave the queue before their service begins. This captures the reality of some service systems such as emergency departments where some patients may leave without being seen. It does not, however, capture the reality of inpatient departments, care homes, or residential care settings where, once patients are admitted, all of their requests must be served. The same applies to diagnostic services in hospitals. As such, we focus on delay queues which are appropriate for situations where all customers joining the queue must be served. Third, Hu et al. (2021b) derive their staffing rules by taking an asymptotic

approach which increases the system scale to infinity. This approach may lead to significant errors in small systems as illustrated in Tables 2 and 3 in Hu et al. (2021b). We aim to address this problem by taking an exact approach. This is important given that the systems representing residential or inpatient care settings are relatively small, as we will illustrate in our case study in §3.6.

Long-term Recruitment Models. Chapter 4 focuses on the trade-off between staffing costs, recruitment lead times, and contract durations of temporary and permanent HCWs in a long-term planning horizon. We found only two related studies, as reported in §2.3. Neither capture the co-existence of temporary and permanent workers explicitly. The fundamental trade-off prevailed in such settings are therefore not considered. Our aim is to fill this gap.

Focusing on settings where all patients must be served, i.e., there is no loss or abandonment, we capture the dynamics of service delivery by generic delay queueing models in both Chapters 3 and 4. A cost-minimization approach, including the cost of workforce plus the waiting cost incurred by patients, is also followed throughout. Further, we model the patient demand as a Poisson mixture model, i.e., a Poisson process with a random rate. As illustrated in Jongbloed and Koole (2001) and Maman (2009), this captures the higher variability relative to the standard Poisson process that is typically observed in patients' arrival data. It also allows us to represent the asymmetry in demand information at the times of permanent and temporary recruitment. In particular, we assume that the distribution of the Poisson rate is available at the time of permanent recruitment and

the exact value of the rate (not the demand itself) is revealed at the time of temporary recruitment. This follows the assumption made in Hu et al. (2021b), and implies that there remains a degree of demand uncertainty at the time of temporary recruitment, which matches the reality of service systems.

2.6 Summary

The literature addressing the recruitment of blended workforce in the presence of uncertain demand was divided into mid-term and long-term categories. Assuming a single-interval planning horizon, studies in the first category consider a single opportunity for permanent recruitment and single or multiple opportunities for temporary recruitment. We further divided this category into single-stage and two-stage streams, and reviewed the studies in both streams. Special attention was made to the way that demand and supply uncertainties were represented in different studies. It was highlighted that a two-stage modelling approach will be followed in this thesis as it captures the asymmetry in demand information available for temporary and permanent recruitment of skilled workers.

The study of Hu et al. (2021b) was identified as the only study that captures the dynamics of service delivery explicitly in the two-stage stream. We set out to fill the gaps in this study by focusing on delay systems (instead of abandonment systems), capturing the uncertainty in permanent recruitment, and proposing a solution approach that produces accurate recruitment levels irrespective of the system size. This will be followed in Chapter

3.

The studies on long-term planning horizon focus on multiple intervals, with a permanent recruitment opportunity and one or more temporary recruitment opportunities in each interval. We reviewed two major studies in this category, and highlighted that neither consider temporary recruitment explicitly. Our objective in Chapter 4 will therefore be to fill this gap by proposing a multi-interval blended recruitment framework.

Chapter 3

A Mid-Term Recruitment Model

3.1 Introduction

The purpose of this chapter is to capture the key trade-off inherent in the mid-term blended workforce planning via a stylized analytical model. This trade-off is between the permanent HCWs, who are cheaper but have a longer recruitment lead time, versus the temporary HCWs, who are as qualified and more expensive but have a substantially shorter recruitment lead time. We specifically consider permanent and temporary recruitment decision making for a highly uncertain demand period (HUDP).

We consider a setting in which patients' requests arrive to the system during the HUDP, and wait in a queue until they are served by a member of a pool of HCWs. The provider must decide how many permanent HCW positions to advertise well ahead of the HUDP,

e.g., several months in advance, when only partial information about demand is available. We refer to this decision as the first-stage decision. Once the permanent positions are advertised, applications arrive and offers are made to qualified applicants. At the start of the HUDP, the provider must then decide how many temporary HCWs to recruit given the number of permanent HCWs recruited and the latest demand information. We refer to this as the second-stage decision, and propose a two-stage stochastic optimization framework to capture the dependence of the second-stage decision on that of the first-stage. The objective is to minimize the expected cost of workforce plus the cost incurred by patients while their requests are in the system. Our framework is based on the assumption that advertisement for permanent positions must begin at an exogenously given time. However, we also investigate the benefit/loss of delaying advertisement to obtain more accurate demand information at the expense of a higher risk of not filling the advertised positions. We further consider an extension of the main framework wherein the HUDP is divided into multiple segments with potentially correlated demand.

As mentioned in Chapter 2, we model the patient demand process as a Poisson mixture model. This is the model proposed in Jongbloed and Koole (2001) and Maman (2009) for capturing the higher variability relative to the standard Poisson process that is commonly observed in patients arrival data. Similar to Hu et al. (2021b), we assume that only the distribution of the Poisson rate is available to the decision maker in the first stage, while the true value of this rate becomes known in the second stage. We model the uncertainty in the permanent recruitment process by considering a probability distribution for the number

of qualified applicants. The dynamics of service delivery in the HUDP are captured by a generic delay queueing model, which evaluates the expected system size, i.e., the mean number of requests waiting or being served, in steady state.

The remainder of this chapter is organized as follows. We start off in §3.2 by proposing our two-stage framework assuming a generic delay queueing model for representing service delivery in the HUDP. Under this general setting, we provide analytical characterization of the optimal first- and second-stage decisions, investigate their monotonicity and insensitivity properties, and propose numerical algorithms for estimating their values. In §3.3, we specialize our framework for three specific queueing models to derive further analytical insight. §3.4 is devoted to conducting numerical experiments, aiming to evaluate the potential savings that could follow from using our proposed framework. In §3.5, we investigate the potential benefit/loss of delaying advertisement for permanent positions. We then develop a unified approach to guide implementation of our two-stage framework in an inpatient department of a hospital in §3.6. In §3.7, we propose an extension to our framework, where we allow multiple opportunities for temporary recruitment. Using this extension, we evaluate the impact of demand correlation between different segments of a HUDP on the optimal permanent recruitment decision. A summary of the findings is presented in §3.8.

3.2 The Two-Stage Framework

Consider a HUDP during which patients' requests arrive for HCW services according to a Poisson process with rate λ. The requests wait in a queue until they are served by a member of the pool of HCWs (permanent and temporary). It takes a random amount of time to serve each request, and the average of this time is set as the time unit so that the rate of service delivery is equal to one. As depicted in Figure 3.1, the HUDP is preceded by a permanent recruitment period of length t_e, during which advertising and recruitment for permanent HCWs occur.

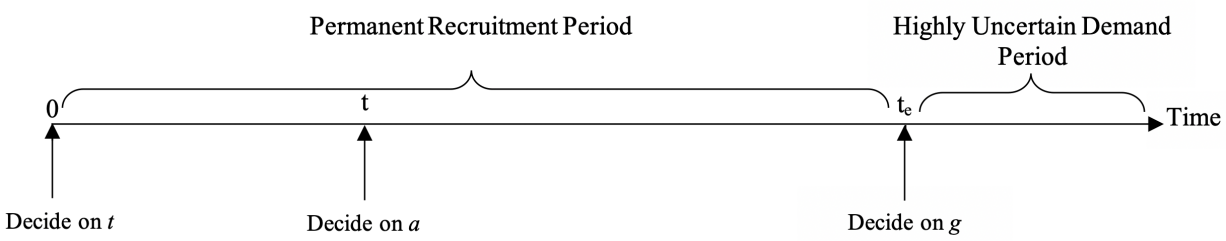


Figure 3.1: Schematic diagram of permanent and temporary recruitment decision making process

The rate of Poisson arrivals is unknown to the service provider during the permanent recruitment period. As such, it is denoted by the random variable Λ_d, for $d \in [0, t_e]$. We assume that the provider knows the distribution (and thus the mean) of Λ_d. This can be estimated from historical data as we will illustrate in §3.6. We further assume that the provider gains full knowledge of the rate at time t_e, i.e., $\Lambda_{t_e} = \lambda$. This can be achieved using a forecasting model such as the one proposed in Hu et al. (2021a). Suppose a decision has been made to advertise for permanent positions at time $t \in [0, t_e)$. This decision is

exogenous as it depends on external factors, e.g., the timing of nurses graduation. In §3.5, however, we explore the potential benefit/risk of a later advertisement. The first-stage problem would then involve the number of permanent full-time equivalent (FTE) positions to advertise at time t, denoted by $a \in \mathbb{R}^+$, whereas the second-stage problem would concern the number of temporary FTEs to recruit at time t_e, denoted by $g \in \mathbb{R}^+$ ($\mathbb{R}^+$ represents the set of non-negative real numbers.) Note that, following the blended workforce literature (e.g., Kao and Queyranne, 1985; Abraham, 1988), we represent the quantity of HCWs (and their positions) with FTEs, which are non-negative real numbers, thus facilitating the derivation of analytical results.

We start with the formulation for the second-stage problem. Following Lu and Lu (2017), we assume that each permanent employee must provide an additional $r_o \geq 0.0$ percentage of mandatory overtime work. Let c_p, c_o, and c_g be the cost rates of permanent, mandatory overtime, and temporary work, respectively. Similar to Lu and Lu (2017), we assume that $c_p < c_o < c_g$. The first inequality is in line with the UK's NHS overtime payment, which is typically 1.5 times of the standard hourly rate (Royal College of Nursing, 2021). The second inequality is supported by various surveys indicating that temporary nurses cost the highest to the employers; see, e.g., Vovak (2010) and National Audit Office (2006). Let c_w be the waiting cost incurred by patients per unit of time in the system, i.e., waiting in the queue and/or being served. We opt to cost the total time the requests stay in the system, instead of the time they are waiting in the queue, in order to reflect the nature of services provided by HCWs. For example, a patient cannot be considered

admitted or discharged until the admission or discharge process is fully completed in an inpatient ward. We normalize the cost rates so that $c_p = 1.0$. Suppose there exists a total of $p \in \mathbb{R}^+$ permanent HCWs in the system at time t_e (including those employed in the most recent recruitment period plus those recruited previously). Given request arrival rate λ and p permanent HCWs, the second-stage problem is then formulated as

$$v(\lambda, p) = \min_g \{p(1 + r_o c_o) + g c_g + l(\lambda, p(1 + r_o) + g) c_w : g \in \mathbb{R}^+, g > \lambda - p(1 + r_o)\}, \quad (3.1)$$

where $l(\lambda, s)$ denotes the system-size function representing the mean number of requests in the system given the rate of requests' arrival is $\lambda > 0$ and the size of the HCW pool is $s \in \mathbb{R}^+$ with $s > \lambda$. In problem (3.1), the first two terms in the objective function yield the total staffing cost, the last term in the objective function gives the performance cost, and the second constraint ensures the stability of the system (recall that service rate is set to one). We denote the optimal solution to problem (3.1) by $g^*(\lambda, p)$.

To formulate the first-stage problem, let Q_t represent the (random) number of qualified applications received during the advertisement period $(t, t_e]$ following $a \in \mathbb{R}^+$ permanent FTE positions being advertised at time t. We assume that offers are made to, and accepted by, qualified applicants on a *sequential* basis. Sequential recruitment is a search strategy, in which each applicant is screened immediately upon arrival, and an offer is made if the applicant is sufficiently qualified (Van Ommeren and Russo, 2014). Recruitment continues until a maximum of a permanent FTEs are recruited or t_e is reached. Given $n \in \mathbb{R}^+$ the

exact FTE of permanent HCWs available in the system at time t who are expected to remain in their jobs during the HUDP, the first-stage problem is formulated as

$$m(n) = \min_{a}\{\mathbb{E}[v\left(\Lambda_t, n + \min\{Q_t, a\}\right)] : a \in \mathbb{R}^+\}, \tag{3.2}$$

where $v(\lambda, p)$ is evaluated through the second-stage problem given in (3.1), and the dependence of $m(n)$ to Λ_t and Q_t is suppressed to simplify the notation. Note that since permanent HCWs are typically expected to give notices if they intend to resign, it is reasonable to assume that n is known to the provider at time t. We denote the optimal solution to problem (3.2) by $a^*(n)$.

In order to characterize the optimal solutions to the first- and second-stage problems, given in Equations (3.2) and (3.1), respectively, we need to make the following set of assumptions concerning the system-size function, $l(\lambda, s)$.

Assumption 1. *$l(\lambda, s)$ satisfies the following properties on its domain $\{(\lambda, s) : \lambda > 0, s \in \mathbb{R}^+, s > \lambda\}$:*

A(i) It is continuous and twice differentiable on λ and s;

A(ii) $\lim_{\lambda\downarrow 0} l(\lambda, s) = 0$, $\lim_{s\downarrow\lambda} l(\lambda, s) = \lim_{\lambda\uparrow s} l(\lambda, s) = \infty$, and $\lim_{s\to\infty} l(\lambda, s)$ is finite;

A(iii) It is strictly increasing in λ, and strictly decreasing in s;

A(iv) It is strictly convex in s;

A(v) Its first order partial derivative with respect to s is strictly decreasing in λ.

Note that $x \uparrow y$ and $x \downarrow y$ denote x approaching y from left and right, respectively. Since the number of servers is typically an integer value in queueing models, an extension to non-integral server numbers is needed for Assumption 1. As we illustrate in §3.3, such extensions exist for some common queueing models. These extensions are continuous and twice differentiable, i.e., property A(i) is met. The first two limits in property A(ii) are trivial and naturally hold. For the last limit in the same property, note that when the number of servers tends to infinity, there will not be a queue in the system, and thus the mean number of requests in the system will be finite. Property A(iii) is trivial. Property A(iv) implies diminishing returns in queueing systems, i.e., the amount of improvement achieved in performance as a result of one additional server reduces as the number of servers increases. Property A(v) implies economies of scale in queueing systems, which can be seen by changing the order of differentiation and noting that congestion always increases with the arrival rate, but this increase reduces with the number of servers. In §3.3, we formally prove these properties for three common queueing models.

Given Assumption 1, we propose

Proposition 1. *For the second-stage problem given in* (3.1)*,*

$$g^*(\lambda, p) = \begin{cases} 0 & \text{if } \lambda \leq \tilde{\lambda}(p), \\ \tilde{g}(\lambda, p) & \text{if } \lambda > \tilde{\lambda}(p), \end{cases} \tag{3.3}$$

where $\tilde{\lambda}(p)$ *is the unique root of function*

$$\phi_p(x) \triangleq c_g + c_w \frac{\partial l(x,s)}{\partial s}\Big|_{s=p(1+r_o)}, \tag{3.4}$$

in the interval $(0, p(1+r_o))$ *when* $p > 0$*, and* $\tilde{\lambda}(0) = 0$*.* $\tilde{g}(\lambda, p)$ *in* (3.3) *is the unique root of function*

$$\theta_{\lambda,p}(g) \triangleq c_g + c_w \frac{\partial l(\lambda,s)}{\partial s}\Big|_{s=p(1+r_o)+g}, \tag{3.5}$$

in the interval $((\lambda - p(1+r_o))^+, \infty)$*, where* $(x)^+ = \max\{0, x\}$ *.*

To prove Proposition 1, we first need the following Lemmas.

Lemma 3.2.1. $\lim_{\lambda \downarrow 0} \dfrac{\partial l(\lambda,s)}{\partial s} = 0.$

Proof. We need to prove that for every $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\left|\frac{\partial l(\lambda,s)}{\partial s}\right| = -\frac{\partial l(\lambda,s)}{\partial s} < \epsilon, \tag{3.6}$$

when $0 < \lambda < \delta$, where the equality above is due to property A(iii). Since $\lim_{\lambda \downarrow 0} l(\lambda, s) = 0$ by property A(ii), we can find a $\delta' > 0$ such that $-l(\lambda, s) > -\epsilon h/2$ for any $h > 0$. Also, we always have $l(\lambda, s+h) > -\epsilon h/2$. Combining these two inequalities, we obtain

$$\frac{l(\lambda, s+h) - l(\lambda, s)}{h} > -\epsilon. \tag{3.7}$$

Taking the limit as h goes to zero and setting $\delta = \delta'$, the proof is complete. □

Lemma 3.2.2. $\lim_{s\downarrow\lambda} \dfrac{\partial l(\lambda,s)}{\partial s} = -\infty$ *and* $\lim_{\lambda\uparrow s} \dfrac{\partial l(\lambda,s)}{\partial s} = -\infty$.

Proof. To show that $\lim_{s\downarrow\lambda} \dfrac{\partial l(\lambda,s)}{\partial s} = -\infty$, we first prove that $\dfrac{\partial l(\lambda,s)}{\partial s}$ is unbounded on $s \in (\lambda, b]$ for any $b > \lambda$. Supposing that it is not true, i.e., $\dfrac{\partial l(\lambda,s)}{\partial s}$ is bounded for all $\lambda < s < b$. Let's call this bound B. Then, by Mean Value theorem (Thomas, 2014), there exists an $\epsilon \in (s, b)$ such that

$$\frac{l(\lambda,b) - l(\lambda,s)}{b-s} = \frac{\partial l(\lambda,s)}{\partial s}\Big|_{s=\epsilon}.$$

Thus,

$$l(\lambda,s) = l(\lambda,b) - \frac{\partial l(\lambda,s)}{\partial s}\Big|_{s=\epsilon}(b-s),$$

and so,

$$\begin{aligned}
|l(\lambda,s)| &= \left| l(\lambda,b) - \frac{\partial l(\lambda,s)}{\partial s}\Big|_{s=\epsilon}(b-s) \right| \\
|l(\lambda,s)| &\leq |l(\lambda,b)| + \left| \frac{\partial l(\lambda,s)}{\partial s}\Big|_{s=\epsilon} \right| |b-s| \\
&\leq |l(\lambda,b)| + B|b-s|,
\end{aligned}$$

for all $s \in (\lambda, b)$. However, this implies that $l(\lambda, s)$ is bounded for all $s \in (\lambda, b)$, which is not true since $\lim_{s\downarrow\lambda} l(\lambda,s) = \infty$ by property A(ii). Hence, $\dfrac{\partial l(\lambda,s)}{\partial s}$ is unbounded on $s \in (\lambda, b]$ for all $b > \lambda$. Now, since $l(\lambda, s)$ is strictly convex in s by property A(iv), $\dfrac{\partial l(\lambda,s)}{\partial s}$ strictly decreases as s approaches λ from above. Following unboundedness of $\dfrac{\partial l(\lambda,s)}{\partial s}$ on $(\lambda, b]$, we must have $\lim_{s\downarrow\lambda} \dfrac{\partial l(\lambda,s)}{\partial s} = -\infty$.

To prove $\lim_{\lambda\uparrow s} \dfrac{\partial l(\lambda,s)}{\partial s} = -\infty$, we need to find an $\epsilon > 0$ for any $M > 0$ such that $\dfrac{\partial l(\lambda,s)}{\partial s} < -M$, whenever $s - \lambda < \epsilon$. Since $\lim_{s\downarrow\lambda} \dfrac{\partial l(\lambda,s)}{\partial s} = -\infty$, there exists an $\epsilon' > 0$ such that $\dfrac{\partial l(\lambda,s)}{\partial s} < -M$, whenever $s - \lambda < \epsilon'$. We can set $\epsilon = \epsilon'$. □

Lemma 3.2.3. $\lim_{s\to\infty} \dfrac{\partial l(\lambda,s)}{\partial s} = 0.$

Proof. Suppose that $\lim_{s\to\infty} \dfrac{\partial l(\lambda,s)}{\partial s} = L \neq 0$. Then, for any $\epsilon > 0$, there is an $M > 0$ such that $\left|\dfrac{\partial l(\lambda,s)}{\partial s} - L\right| < \epsilon$ when $s > M$. Now, consider an arbitrary $s > M$. By the Mean Value theorem, there is a point $\delta_s \in (s, s+1)$ such that

$$l(\lambda, s+1) - l(\lambda, s) = \frac{\partial l(\lambda,s)}{\partial s}\Big|_{s=\delta_s}.$$

Since $M < s < \delta_s$, we have $\left|\dfrac{\partial l(\lambda,s)}{\partial s}\Big|_{s=\delta_s} - L\right| < \epsilon$, and so $|l(\lambda, s+1) - l(\lambda, s) - L| < \epsilon$. Taking the limit as $s \to \infty$, and noting that $\lim_{s\to\infty} l(\lambda, s)$ is finite by property A(ii), we obtain $|L| < \epsilon$, which cannot be true if $L \neq 0$, hence $\lim_{s\to\infty} \dfrac{\partial l(\lambda,s)}{\partial s} = 0$. □

We now prove Proposition 1 as follows.

Proof. The Lagrange function of the optimization problem given in (3.1) is obtained as

$$\mathcal{L}(\lambda, p, \beta; g) = p(1 + r_o c_o) + g c_g + l(\lambda, p(1+r_o) + g) c_w - \beta g,$$

where β is the Karush-Kuhn-Tucker (KKT) multiplier. Note that constraint $g > \lambda - p(1+r_o)$ is not included in the Lagrange function as it is always active and so its multiplier is equal to zero. This leads to the following scenarios:

(i) $\beta = 0$: the first-order condition is

$$\frac{\partial \mathcal{L}(\lambda, p, \beta; g)}{\partial g} = c_g + c_w \frac{\partial l(\lambda, s)}{\partial s}\Big|_{s=p(1+r_o)+g} = 0. \tag{3.8}$$

From the primal feasibility conditions, we must also have $g \geq 0$.

(ii) $\beta > 0$: the first order condition is

$$\frac{\partial \mathcal{L}(\lambda, p, \beta; g)}{\partial g} = c_g + c_w \frac{\partial l(\lambda, s)}{\partial s}\Big|_{s=p(1+r_o)+g} - \beta = 0,$$

and so,

$$\beta = c_g + c_w \frac{\partial l(\lambda, s)}{\partial s}\Big|_{s=p(1+r_o)},$$

where $g = 0$ is obtained from the complementary slackness condition for the non-negativity constraint. We must also have $\lambda - p(1+r_o) < g$, or equivalently $\lambda < p(1+r_o)$.

Given the convexity of $l(\lambda, s)$, the values of g meeting the conditions in scenarios (i) or (ii) will be optimal. To find these values, focusing initially on the situation with $p > 0$, we first show that the function

$$\phi_p(x) \triangleq c_g + c_w \frac{\partial l(x, s)}{\partial s}\Big|_{s=p(1+r_o)},$$

has a unique root in the interval $(0, p(1+r_o))$. By properties A(i) and A(v), $\dfrac{\partial l(x, s)}{\partial s}$ is contin-

uous and strictly decreasing in x. By Lemma 3.2.1, we also have $\lim_{x\downarrow 0} \dfrac{\partial l(x,s)}{\partial s} = 0$, and so $\lim_{x\downarrow 0} \phi_p(x) = c_g$, which is always positive. Further, by Lemma 3.2.2, $\lim_{x\uparrow s} \dfrac{\partial l(x,s)}{\partial s} = -\infty$, and so $\lim_{x\uparrow p(1+r_o)} \phi_p(x) = -\infty$. As such, by the Intermediate Value theorem and Rolle's theorem (Thomas, 2014), there exists a unique solution to $\phi_p(x) = 0$, which we denote by $\tilde{\lambda}(p)$. Then, for values of $\lambda \in (0, \tilde{\lambda}(p))$, $\phi_p(\lambda) = \beta > 0$, and also $\lambda < \tilde{\lambda}(p) < p(1+r_o)$, hence, the conditions of scenario (ii) are met for $g = 0$, and so $g^*(\lambda, p) = 0$ for $\lambda \in (0, \tilde{\lambda}(p))$.

For values of $\lambda \in [\tilde{\lambda}(p), p(1+r_o))$, $\phi_p(\lambda) = \beta \leq 0$, and so conditions of scenario (ii) do not hold. For $\lambda = \tilde{\lambda}(p)$, however, $\phi_p(\tilde{\lambda}(p)) = 0$. Hence, defining

$$\theta_{\lambda,p}(g) \triangleq c_g + c_w \frac{\partial l(\lambda, s)}{\partial s}\bigg|_{s=p(1+r_o)+g},$$

we obtain $\theta_{\tilde{\lambda}(p),p}(0) = 0$. This implies that conditions of scenario (i) are met for $g = 0$ and so $g^*(\tilde{\lambda}(p), p) = 0$. For $\lambda \in (\tilde{\lambda}(p), p(1+r_o))$, on the other hand, since $c_g + c_w \dfrac{\partial l(\lambda, s)}{\partial s}$ is strictly increasing in s (by property A(iv)), there exists a unique value $\tilde{g}(\lambda, p) > 0$ such that $\theta_{\lambda,p}(\tilde{g}(\lambda, p)) = 0$, thus meeting conditions of scenario (i). That is $g^*(\lambda, p) = \tilde{g}(\lambda, p)$ for $\lambda \in (\tilde{\lambda}(p), p(1+r_o))$.

Finally, for $\lambda \in [p(1+r_o), \infty)$, we show that a $\tilde{g}(\lambda, p) \in (\lambda - p(1+r_o), \infty)$ can still be found to satisfy $\theta_{\lambda,p}(g) = 0$. We know that $\theta_{\lambda,p}(g)$ is continuous and strictly increasing in g. By Lemma 3.2.2, we also have $\lim_{s\downarrow\lambda} \dfrac{\partial l(\lambda, s)}{\partial s} = -\infty$, and so $\lim_{g\downarrow\lambda - p(1+r_o)} \theta_{\lambda,p}(g) = -\infty$. Further, by Lemma 3.2.3, $\lim_{s\to\infty} \dfrac{\partial l(\lambda, s)}{\partial s} = 0$, and so $\lim_{g\to\infty} \theta_{\lambda,p}(g) = c_g$, which is always positive. As a result, based on Intermediate Value and Rolle's theorems, there exists a

unique value $\tilde{g}(\lambda, p) > \lambda - p(1 + r_o)$ satisfying $\theta_{\lambda,p}(g) = 0$. The conditions of scenario (i) are therefore met for $\tilde{g}(\lambda, p)$, and so $g^*(\lambda, p) = \tilde{g}(\lambda, p)$ for $\lambda \in (\tilde{\lambda}(p), \infty)$. Similarly, for the situation with $p = 0$, we can find a value $\tilde{g}(\lambda, 0) \in (\lambda, \infty)$ satisfying the equation $\theta_{\lambda,0}(\tilde{g}(\lambda, 0)) = 0$, and thus the conditions of scenario (i). Therefore $g^*(\lambda, 0) = \tilde{g}(\lambda, 0)$, and the proof is complete. □

Algorithm 1 outlines the steps for evaluating $g^*(\lambda, p)$ based on Proposition 1. This algorithm includes a function for evaluating $\tilde{\lambda}(p)$ as the unique root of $\phi_p(x)$ (given in Equation (3.4)) in the interval $(0, p(1 + r_o))$. As shown in the proof of Proposition 1, $\phi_p(x)$ is continuous and strictly decreasing in x, with a positive value when $x \downarrow 0$, and a negative value when $x \uparrow p(1 + r_o)$. Hence, its root can be obtained by a bracketing method, such as Brent's method (Brent, 1973), with the bracketing interval set to $[\alpha_1, p(1 + r_o) - \alpha_2]$, where α_1 and α_2 are small positive numbers. Note that lines 18 to 25 in Algorithm 1 choose α_1 and α_2 such that $\phi_p(\alpha_1) > 0$ and $\phi_p(p(1 + r_o) - \alpha_2) < 0$, thus ensuring the existence of a unique root in the interval $[\alpha_1, p(1+r_o)-\alpha_2]$. Algorithm 1 also evaluates the unique root of function $\theta_{\lambda,p}(g)$ (given in Equation (3.5)) in the interval $((\lambda - p(1 + r_o))^+, \infty)$. As shown in the proof, $\theta_{\lambda,p}(g)$ is continuous and strictly increasing in g, negative when $g \downarrow (\lambda - p(1 + r_o))^+$, and positive when $g \to \infty$. Its root can therefore be obtained by Brent's method with the bracketing interval set as outlined in Algorithm 1. Note that lines 4 to 11 in Algorithm 1 choose α and g_u such that $\theta_{\lambda,p}((\lambda - p(1 + r_o))^+ + \alpha) < 0$ and $\theta_{\lambda,p}(g_u) > 0$, thus ensuring a unique root can be found in the interval $[(\lambda - p(1 + r_o))^+ + \alpha, g_u]$.

Algorithm 1 Numerical method for evaluating the optimal number of temporary HCWs, $g^*(\lambda, p)$

Require: $l(x, y)$, λ, p, c_g, c_w, r_o
if $\lambda \leq \tilde{\lambda}(p)$ **then**
  $g^*(\lambda, p) \leftarrow 0$
else
  $\alpha \leftarrow 0.001$
  while $\theta_{\lambda,p}((\lambda - p(1 + r_o))^+ + \alpha) > 0$ **do**
    $\alpha \leftarrow \alpha/10.0$
  end while
  $g_u \leftarrow ((\lambda - p(1 + r_o))^+ + 10.0$
  while $\theta_{\lambda,p}(g_u) < 0$ **do**
    $g_u \leftarrow g_u \times 10$
  end while
  $g^*(\lambda, p) \leftarrow$ root of $\theta_{\lambda,p}(g)$ in the interval $[(\lambda - p(1 + r_o))^+ + \alpha, g_u]$
end if
function $\tilde{\lambda}(p)$
  if $p = 0$ **then**
    return 0.0
  else
    $\alpha_1 \leftarrow 0.001$
    while $\phi_p(\alpha_1) < 0$ **do**
      $\alpha_1 \leftarrow \alpha_1/10.0$
    end while
    $\alpha_2 \leftarrow 0.001$
    while $\phi_p(p(1 + r_o) - \alpha_2) > 0$ **do**
      $\alpha_2 \leftarrow \alpha_2/10.0$
    end while
    return root of $\phi_p(x)$ in the interval $[\alpha_1, p(1 + r_o) - \alpha_2]$
  end if
end function

Before proceeding to the first-stage problem, we provide the monotonicity properties of the optimal second-stage decision with respect to λ and p.

Corollary 1. *The optimal second-stage decision, $g^*(\lambda, p)$, is increasing in λ, and decreasing in p.*

To prove Corollary 1, we first need the following lemma.

Lemma 3.2.4. *$\tilde{\lambda}(p)$ is strictly increasing in p.*

Proof. By property A(iv), $\frac{\partial l(x,s)}{\partial s}\Big|_{s=p(1+r_o)}$ is strictly increasing in p. This implies that $\phi_p(x)$ is also strictly increasing in p for all values of $x \in (0, p(1+r_o))$. From this, and the fact that $\phi_p(x)$ is a strictly decreasing function of x, we conclude that the root of this function, i.e., $\tilde{\lambda}(p)$, increases strictly with p. □

We now prove Corollary 1.

Proof. Since function $\theta_{\lambda,p}(g)$ is strictly increasing in g as shown in the proof of Proposition 1, and strictly decreasing in λ by property A(v), its root, i.e., $\tilde{g}(\lambda, p)$, must increase as λ increases. This implies that $g^*(\lambda, p)$ is increasing in λ. Further, since $\theta_{\lambda,p}(g)$ is strictly increasing in p by property A(iv), its root, i.e., $\tilde{g}(\lambda, p)$, must decrease when p increases. This, combined with the fact that $\tilde{\lambda}(p)$ is strictly increasing in p by Lemma 3.2.4, proves that $g^*(\lambda, p)$ is decreasing in p. □

Following Algorithm 1, we can obtain the optimal second-stage decision $g^*(\lambda, p)$, and thus the corresponding cost $v(\lambda, p)$, for any values of λ and p. In theory, this should enable us to evaluate the objective function of the first-stage problem given in (3.2) for different values of $a \geq 0$, providing an estimate for $a^*(n)$. More specifically, let $h_t(\cdot)$ and $f_t(\cdot)$ be the probability density functions (pdfs) of Λ_t and Q_t supported in intervals $[0, \lambda_u)$ and $[0, q_u)$,

respectively, where λ_u and q_u could be infinitely large. Expanding the first-stage objective function, we then have

$$\mathbb{E}[v\left(\Lambda_t, n + \min\{Q_t, a\}\right)] = \int_0^{\lambda_u} \int_0^{q_u} v(\lambda, n + \min\{q, a\}) f_t(q) h_t(\lambda) dq\, d\lambda. \tag{3.9}$$

Evaluating Equation (3.9) for a given a would require calculating a double integral over (potentially) infinite intervals. This calculation would require evaluating the integrand for many pairs of (λ, p), which requires evaluating two pdf functions and the optimal second-stage cost $v(\lambda, p)$, which in turn requires evaluating $g^*(\lambda, p)$. The computation time would therefore be significant, thus making this approach impractical. Instead, we consider the structural properties of the objective function as will be elaborated later in the proof of Proposition 2. In short, let us denote by $\psi_n(a)$ the derivative of the objective function conditioned on $Q_t = q$, $q > a$ with respect to a, i.e., $\partial \mathbb{E}[v\left(\Lambda_t, n + a\right)]/\partial a$. Following integral differentiation rules, we then have

$$\begin{aligned}\psi_n(a) = 1 + r_o c_o + c_w(1 + r_o) \int_0^{\tilde{\lambda}(n+a)} \frac{\partial l(\lambda, s)}{\partial s}\Big|_{s=(n+a)(1+r_o)} h_t(\lambda) d\lambda \\ - c_g(1 + r_o)\left(1 - H_t\left(\tilde{\lambda}(n + a)\right)\right),\end{aligned} \tag{3.10}$$

where $H_t(\cdot)$ is the cumulative distribution function of Λ_t. We now propose

Proposition 2. *For the first-stage problem given in* (3.2)*,*

$$a^*(n) = \begin{cases} 0.0 & \text{if } \psi_n(0.0) \geq 0.0, \\ \min\{\tilde{a}(n), q_u\}, & \text{otherwise}, \end{cases} \tag{3.11}$$

where $\tilde{a}(n)$ *is the unique root of function* $\psi_n(a)$ *in the interval* $(0, \infty)$*.*

We use the following lemmas to prove Proposition 2.

Lemma 3.2.5. $\lim_{a\to\infty} \psi_n(a)$ *is positive.*

Proof. By Lemma 3.2.3, $\lim_{s\to\infty} \dfrac{\partial l(\lambda, s)}{\partial s} = 0$. From Lemma 3.2.4, we know that $\tilde{\lambda}(p)$ is strictly increasing in p, so $\lim_{p\to\infty} \tilde{\lambda}(p) = \infty$. Hence, $\lim_{a\to\infty} \psi_n(a) = 1 + r_o c_o$, which is positive. □

Lemma 3.2.6. $\psi_n(a)$ *is continuous, and strictly increasing in* a *and* n*.*

Proof. Continuity is trivial. Taking the derivative from $\psi_n(a)$ with respect to a gives

$$\begin{aligned} \frac{\partial \psi_n(a)}{\partial a} &= c_w(1+r_o)\frac{\partial \tilde{\lambda}(n+a)}{\partial a}\frac{\partial l(\tilde{\lambda}(n+a), s)}{\partial s}\Big|_{s=(a+n)(1+r_o)} h_t(\tilde{\lambda}(n+a)) \\ &\quad + c_w(1+r_o)^2 \int_0^{\tilde{\lambda}(n+a)} \frac{\partial^2 l(\lambda, s)}{\partial s^2}\Big|_{s=(a+n)(1+r_o)} h_t(\lambda)d\lambda + c_g(1+r_o)\frac{\partial \tilde{\lambda}(n+a)}{\partial a} h_t(\tilde{\lambda}(n+a)) \\ &= (1+r_o)\frac{\partial \tilde{\lambda}(n+a)}{\partial a} h_t(\tilde{\lambda}(n+a)) \left(c_g + c_w \frac{\partial l(\tilde{\lambda}(n+a), s)}{\partial s}\Big|_{s=(a+n)(1+r_o)} \right) \\ &\quad + c_w(1+r_o)^2 \int_0^{\tilde{\lambda}(n+a)} \frac{\partial^2 l(\lambda, s)}{\partial s^2}\Big|_{s=(a+n)(1+r_o)} h_t(\lambda)d\lambda \\ &= c_w(1+r_o)^2 \int_0^{\tilde{\lambda}(n+a)} \frac{\partial^2 l(\lambda, s)}{\partial s^2}\Big|_{s=(a+n)(1+r_o)} h_t(\lambda)d\lambda, \end{aligned} \tag{3.12}$$

where the last equality is because $\tilde{\lambda}(n+a)$ is the unique root of function $\phi_{n+a}(x)$ given in Equation (3.4). By property A(iv), the expression obtained for $\dfrac{\partial \psi_n(a)}{\partial a}$ is always positive. For the last part, taking the derivative from $\psi_n(a)$ with respect to n gives

$$\begin{aligned}
\frac{\partial \psi_n(a)}{\partial n} &= c_g(1+r_o)\left(\frac{\partial \tilde{\lambda}(n+a)}{\partial n}\right) h_t(\tilde{\lambda}(n+a)) \\
&\quad + c_w(1+r_o)\left(\frac{\partial l(\tilde{\lambda}(n+a),s)}{\partial s}\Big|_{s=(n+a)(1+r_o)}\left(\frac{\partial \tilde{\lambda}(n+a)}{\partial n}\right) h_t(\tilde{\lambda}(n+a))\right) \\
&\quad + c_w(1+r_o)\int_0^{\tilde{\lambda}(n+a)} \frac{\partial^2 l(\lambda,s)}{\partial s \partial n}\Big|_{s=(n+a)(1+r_o)} h_t(\lambda)d\lambda \\
&= (1+r_o)\left(\frac{\partial \tilde{\lambda}(n+a)}{\partial n}\right) h_t(\tilde{\lambda}(n+a))\left(c_g + c_w \frac{\partial l(\tilde{\lambda}(n+a),s)}{\partial s}\Big|_{s=(a+n)(1+r_o)}\right) \\
&\quad + c_w(1+r_o)\int_0^{\tilde{\lambda}(n+a)} \frac{\partial^2 l(\lambda,s)}{\partial s \partial n}\Big|_{s=(n+a)(1+r_o)} h_t(\lambda)d\lambda \\
&= c_w(1+r_o)\int_0^{\tilde{\lambda}(n+a)} \frac{\partial^2 l(\lambda,s)}{\partial s \partial n}\Big|_{s=(n+a)(1+r_o)} h_t(\lambda)d\lambda \\
&= c_w(1+r_o)^2\int_0^{\tilde{\lambda}(n+a)} \frac{\partial^2 l(\lambda,s)}{\partial s^2}\Big|_{s=(n+a)(1+r_o)} h_t(\lambda)d\lambda
\end{aligned}$$

which is positive by property A(iv). □

Now, we prove Proposition 2.

Proof. Denoting the expected value in the first-stage problem given in (3.2) by $y(n,a)$, and conditioning on Q_t, we obtain

$$y(n,a) \triangleq \mathbb{E}[v\left(\Lambda_t, n+\min\{Q_t,a\}\right)] = \int_0^a \mathbb{E}\left[v(\Lambda_t, n+q)\right] f_t(q)dq + \mathbb{E}\left[v(\Lambda_t, n+a)\right](1-F_t(a)),$$

where F_t is the cumulative distribution function of Q_t. Taking the derivative of $y(n,a)$ with

respect to a and simplifying, we arrive at

$$\frac{\partial y(n,a)}{\partial a} = \frac{\partial \mathbb{E}[v(\Lambda_t, n+a)]}{\partial a}\left(1 - F_t(a)\right) = \mathbb{E}\left[\frac{\partial v(\Lambda_t, n+a)}{\partial a}\right]\left(1 - F_t(a)\right). \tag{3.13}$$

For an arbitrary λ, using the Envelope theorem (Takayama, 1985), we then have

$$\begin{aligned}\frac{\partial v(\lambda, n+a)}{\partial a} &= \frac{\partial}{\partial a}\min\left\{(n+a)(1+r_o c_o) + g c_g + c_w l(\lambda, (n+a)(1+r_o)+g) : \right. \\ &\qquad \left. g \in \mathbb{R}^+, g > \lambda - (n+a)(1+r_o)\right\} \\ &= \frac{\partial}{\partial a}\Big[(n+a)(1+r_o c_o) + g c_g + c_w l(\lambda, (n+a)(1+r_o)+g)\Big]\Big|_{g=g^*(\lambda,n+a)}, \end{aligned} \tag{3.14}$$

where $g^*(\lambda, n+a)$ is the optimal solution to the second-stage problem given in (3.1) with $p = n+a$. From (3.14), we obtain

$$\frac{\partial v(\lambda, n+a)}{\partial a} = 1 + r_o c_o + c_w(1+r_o)\frac{\partial l(\lambda, s)}{\partial s}\Big|_{s=(n+a)(1+r_o)+g^*(\lambda,n+a)}. \tag{3.15}$$

Taking the expectation of the above expression, it follows that

$$\mathbb{E}\left[\frac{\partial v(\Lambda_t, n+a)}{\partial a}\right] = 1 + r_o c_o + c_w(1+r_o)\mathbb{E}\left[\frac{\partial l(\Lambda_t, s)}{\partial s}\Big|_{s=(n+a)(1+r_o)+g^*(\Lambda_t,n+a)}\right]. \tag{3.16}$$

Replacing $g^*(\Lambda_t, n+a)$ from Proposition 1, we obtain

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial v(\Lambda_t, n+a)}{\partial a}\right] &= 1 + r_o c_o + c_w(1+r_o)\left(\mathbb{E}\left[\frac{\partial l(\Lambda_t, s)}{\partial s}\Big|_{s=(n+a)(1+r_o)}, \Lambda_t \leq \tilde{\lambda}(n+a)\right]\right. \\
&\qquad \left. + \mathbb{E}\left[\frac{\partial l(\Lambda_t, s)}{\partial s}\Big|_{s=(n+a)(1+r_o)+\tilde{g}(\Lambda_t, n+a)}, \Lambda_t > \tilde{\lambda}(n+a)\right]\right) \\
&= 1 + r_o c_o + c_w(1+r_o)\left(\mathbb{E}\left[\frac{\partial l(\Lambda_t, s)}{\partial s}\Big|_{s=(n+a)(1+r_o)}, \Lambda_t \leq \tilde{\lambda}(n+a)\right]\right. \\
&\qquad \left. + \mathbb{E}\left[-\frac{c_g}{c_w}, \Lambda_t > \tilde{\lambda}(n+a)\right]\right) \\
&= 1 + r_o c_o + c_w(1+r_o)\int_0^{\tilde{\lambda}(n+a)} \frac{\partial l(\lambda, s)}{\partial s}\Big|_{s=(n+a)(1+r_o)} h_t(\lambda)d\lambda - c_g(1+r_o)\int_{\tilde{\lambda}(n+a)}^{\infty} h_t(\lambda)d\lambda,
\end{aligned} \tag{3.17}$$

where the last term in the second equality is because $\tilde{g}(\lambda, n+a)$ is the root of function $\theta_{\lambda,n+a}(g)$ given in Expression (3.5) for all values of $\lambda > \tilde{\lambda}(n+a)$. Denoting the expression obtained above for $\mathbb{E}\left[\frac{\partial v(\Lambda_t, n+a)}{\partial a}\right]$ by $\psi_n(a)$, we have

$$\frac{\partial y(n,a)}{\partial a} = \psi_n(a)(1 - F_t(a)). \tag{3.18}$$

By Lemma 3.2.6, $\psi_n(a)$ is a continuous and strictly increasing function of a. Further, by Lemma 3.2.5, $\lim_{a\to\infty} \psi_n(a)$ is always positive. Hence, if $\psi_n(0.0) < 0.0$, by the Intermediate Value theorem and Rolle's theorem, there exists a unique solution to $\psi_n(a) = 0.0$, which we denote by $\tilde{a}(n)$. $\psi_n(a)$ is then negative (positive) for $a \in [0.0, \tilde{a}(n))$ ($a \in (\tilde{a}(n), \infty)$). On the other hand, if $\psi_n(0.0) \geq 0.0$, then $\psi_n(a) > 0.0$ for all $a > 0.0$.

First, consider the situation where $\psi_n(0.0) < 0.0$. Then, when $\tilde{a}(n) < q_u$, $\psi_n(a)(1-F_t(a))$ starts from a negative value at $a = 0.0$, and increases strictly with a until it becomes 0.0 at

$a = \tilde{a}(n)$. It then increases to a positive value and decreases back to 0.0 at $a = q_u$, and remains equal to 0.0 from that point onwards. This implies that the minimum of $y(n, a)$ occurs at $a = \tilde{a}(n)$. On the other hand, when $\tilde{a}(n) \geq q_u$, $\psi_n(a)(1 - F_t(a))$ starts from a negative value at $a = 0.0$, increases strictly with a until it becomes 0.0 at $a = q_u$, and remains 0.0 for $a > q_u$. This implies that the minimum of $y(n, a)$ occurs at $a = q_u$ (Note that any $a \geq q_u$ will be optimal in this case.) These lead to Equation (3.11).

Next, consider the situation where $\psi_n(0.0) \geq 0.0$. Then $\psi_n(a)(1 - F_t(a))$ starts from a non-negative value at $a = 0.0$, and remains non-negative for all $a > 0.0$. This implies that the minimum of $y(n, a)$ occurs at $a = 0.0$. □

Proposition 2 leads to two important corollaries:

Corollary 2. $a^*(0.0) > 0.0$.

Proof. Setting $n = a = 0.0$ in Equation (3.10), we obtain $\psi_0(0.0) = 1 + r_o c_o - c_g(1 + r_o) = 1 - c_g + r_o(c_o - c_g) < 0.0$, where the inequality is because $1 < c_o < c_g$ by assumption. It then follows from Proposition 2 that $a^*(0.0) > 0.0$. □

Corollary 3. *$a^*(n)$ is insensitive to changes in f_t as long as its support remains the same. Further, when $q_u \to \infty$, $a^*(n)$ will be a hire-up-to policy evaluated as $a^*(n) = (\tilde{a}(0.0) - n)^+$.*

Proof. The first part is because f_t does not appear in the expression for $\psi_n(a)$. For the second part, first note that, when $q_u \to \infty$, $a^*(n) = 0.0$ if $\psi_n(0.0) \geq 0.0$, and $a^*(n) = \tilde{a}(n)$ if $\psi_n(0.0) < 0.0$, by Equation (3.11). Next, we consider two situations. First, suppose

that $\psi_n(0.0) \geq 0.0$. Then, since $\psi_n(a)$ increases strictly with n by Lemma 3.2.6, we have $\psi_{n+x}(0.0) > \psi_n(0.0) \geq 0.0$, implying that

$$a^*(n+x) = a^*(n) = 0.0, \tag{3.19}$$

for any $x \geq 0.0$. Second, suppose that $\psi_n(0.0) < 0.0$ and so $a^*(n) = \tilde{a}(n)$ which is a positive value. Then, since a and n only appear as $(a+n)$ in the expression for $\psi_n(a)$, we have $\psi_n(a) = \psi_{n+x}(a-x)$ for $0 \leq x \leq a$. Setting $a = \tilde{a}(n)$, we obtain $\psi_{n+x}\left(\tilde{a}(n) - x\right) = \psi_n\left(\tilde{a}(n)\right) = 0$, where the second equality is by definition. This implies that

$$a^*(n+x) = \tilde{a}(n) - x, \tag{3.20}$$

for $0 \leq x \leq \tilde{a}(n)$. In particular, for $x = \tilde{a}(n)$, we obtain $\psi_{n+\tilde{a}(n)}(0.0) = 0.0$ and $a^*(n + \tilde{a}(n)) = 0.0$. Further, since by Lemma 3.2.6, $\psi_n(a)$ increases strictly with n, we have $\psi_{n+x}(0) > \psi_{n+\tilde{a}(n)}(0) = 0$, implying that

$$a^*(n+x) = 0.0, \tag{3.21}$$

for $x > \tilde{a}(n)$. Combining Equations (3.19), (3.20), and (3.21), we arrive at

$$a^*(n+x) = \begin{cases} 0.0 & \text{if } \psi_n(0) \geq 0.0, \\ \left(\tilde{a}(n) - x\right)^+, & \text{otherwise,} \end{cases} \tag{3.22}$$

for any $x \geq 0$. Setting $n = 0.0$ in Equation (3.22) and a change of variable yield

$$a^*(n) = \begin{cases} 0.0 & \text{if } \psi_0(0.0) \geq 0.0, \\ (\tilde{a}(0.0) - n)^+, & \text{otherwise.} \end{cases} \tag{3.23}$$

But $\psi_0(0.0)$ is always negative as shown in the proof of Corollary 2, and thus $a^*(n) = (\tilde{a}(0.0) - n)^+$. □

Corollary 2 implies that it is never cost-effective to serve customer requests with only temporary HCWs. Corollary 3 implies that, when $q_u \to \infty$, evaluating $\tilde{a}(0.0)$ is sufficient for characterizing $a^*(n)$ for any value of n.

Algorithm 2 outlines the steps for obtaining $a^*(n)$ based on Proposition 2. The algorithm needs to evaluate $\tilde{a}(n)$ as the unique root of function $\psi_n(a)$ in the interval $(0, \infty)$. As shown in the proof of Proposition 2, $\psi_n(a)$ is a continuous and strictly increasing function in a with a negative value at $a = 0$ and a positive value when $a \to \infty$. The root of this function can therefore be obtained by Brent's method given a value $a_u > 0$ such that $\psi_n(a_u) > 0$. Indeed, lines 5 to 8 in Algorithm 2 choose a_u such that $\psi_n(a_u) > 0$ to ensure the existence of a unique root in the interval $(0, a_u]$. Note that each step of Brent's method would also require to evaluate $\psi_n(a)$ for different values of a, which in turn requires evaluating of $\tilde{\lambda}(n + a)$ using the function provided in Algorithm 1.

It is important that Algorithm 2 provides an accurate estimate for the optimal first-stage decision by using function $\psi_n(a)$, which involves only single integrals over finite

Algorithm 2 Numerical method for evaluating the optimal number of permanent positions to advertise, $a^*(n)$.

Require: $l(x, y)$, h_t, c_g, c_o, c_w, r_o, n, q_u, and function $\tilde{\lambda}(p)$ from Algorithm 1

Evaluate $\psi_n(0)$ from Equation (3.10) using $\tilde{\lambda}(n)$ as input
if $\psi_n(0) \geq 0$ **then**
    $a^*(n) \leftarrow 0$
else
    $a_u \leftarrow 10$
    while $\psi_n(a_u) < 0$ **do**
        $a_u \leftarrow a_u \times 10$
    end while
    $\tilde{a}(n) \leftarrow$ root of $\psi_n(a)$ in the interval $(0, a_u]$
    $a^*(n) \leftarrow \min\{\tilde{a}(n), q_u\}$
end if

intervals, and does not require evaluating the optimal second-stage decision and its cost. The calculations are fast as a result, leading to the optimal decision in less than a second in all the numerical experiments we conducted. It is also noteworthy that the methodology proposed for evaluating $a^*(n)$ lends itself to further analytical investigation, leading to results such as the monotonicity properties given in the corollaries 4 and 5 below.

Corollary 4. *$a^*(n)$ increases with c_g and c_w, and decreases with c_o and n.*

Proof. (i) Taking the derivative of $\psi_n(a)$ with respect to c_g, we get

$$\begin{aligned}
\frac{\partial \psi_n(a)}{\partial c_g} &= -(1+r_o)\int_{\tilde{\lambda}(n+a)}^{\infty} h_t(\lambda)d\lambda + \left(\frac{\partial \tilde{\lambda}(n+a)}{\partial c_g}\right) h_t(\tilde{\lambda}(n+a))c_g(1+r_o) \\
&\quad + c_w(1+r_o)\left(\frac{\partial l(\tilde{\lambda}(n+a),s)}{\partial s}\Big|_{s=(n+a)(1+r_o)}\left(\frac{\partial \tilde{\lambda}(n+a)}{\partial c_g}\right) h_t(\tilde{\lambda}(n+a))\right) \\
&= -(1+r_o)\int_{\tilde{\lambda}(n+a)}^{\infty} h_t(\lambda)d\lambda \\
&\quad + \left(\frac{\partial \tilde{\lambda}(n+a)}{\partial c_g}\right) h_t(\tilde{\lambda}(n+a))(1+r_o)\left(c_g + c_w\frac{\partial l(\tilde{\lambda}(n+a),s)}{\partial s}\Big|_{s=(a+n)(1+r_o)}\right) \\
&= -(1+r_o)\int_{\tilde{\lambda}(n+a)}^{\infty} h_t(\lambda)d\lambda,
\end{aligned}$$

which is negative. This, along with the fact that $\psi_n(a)$ is strictly increasing in a by Lemma 3.2.6, implies that $\tilde{a}(n)$ and thus $a^*(n)$ is increasing in c_g.

(ii) Taking the derivative of $\psi_n(a)$ with respect to c_w yields

$$\begin{aligned}
\frac{\partial \psi_n(a)}{\partial c_w} &= c_g(1+r_o)\left(\frac{\partial \tilde{\lambda}(n+a)}{\partial c_w}\right) h_t(\tilde{\lambda}(n+a)) + (1+r_o)\int_0^{\tilde{\lambda}(n+a)} \frac{\partial l(\lambda,s)}{\partial s}\Big|_{s=(n+a)(1+r_o)} h_t(\lambda)d\lambda \\
&\quad + c_w(1+r_o)\left(\frac{\partial l(\tilde{\lambda}(n+a),s)}{\partial s}\Big|_{s=(n+a)(1+r_o)}\left(\frac{\partial \tilde{\lambda}(n+a)}{\partial c_w}\right) h_t(\tilde{\lambda}(n+a))\right) \\
&= (1+r_o)\left(\frac{\partial \tilde{\lambda}(n+a)}{\partial c_w}\right) h_t(\tilde{\lambda}(n+a))\left(c_g + c_w\frac{\partial l(\tilde{\lambda}(n+a),s)}{\partial s}\Big|_{s=(a+n)(1+r_o)}\right) \\
&\quad + (1+r_o)\int_0^{\tilde{\lambda}(n+a)} \frac{\partial l(\lambda,s)}{\partial s}\Big|_{s=(n+a)(1+r_o)} h_t(\lambda)d\lambda \\
&= (1+r_o)\int_0^{\tilde{\lambda}(n+a)} \frac{\partial l(\lambda,s)}{\partial s}\Big|_{s=(n+a)(1+r_o)} h_t(\lambda)d\lambda,
\end{aligned}$$

which is negative by property A(iii). This, along with the fact that $\psi_n(a)$ is strictly increasing in a by Lemma 3.2.6, implies that $\tilde{a}(n)$ and thus $a^*(n)$ is increasing in c_w.

(iii) $\psi_n(a)$ is clearly increasing in c_o. This, along with the fact that $\psi_n(a)$ is strictly increasing in a by Lemma 3.2.6, implies that $\tilde{a}(n)$ and thus $a^*(n)$ is decreasing in c_o.

(iv) By Lemma 3.2.6, $\psi_n(a)$ is increasing in n. This, along with the fact that $\psi_n(a)$ is strictly increasing in a by Lemma 3.2.6, implies that $\tilde{a}(n)$ and thus $a^*(n)$ is decreasing in n. □

Corollary 5. *$m(n)$ increases with c_g, c_w, and c_o.*

Proof. Taking the derivative of $m(n)$ given in Equation (3.2) with respect to c_g, we arrive at

$$\begin{aligned}
\frac{\partial m(n)}{\partial c_g} &= \frac{\partial}{\partial c_g} \min\left\{\mathbb{E}[v(\Lambda_t, n + \min\{Q_t, a\})] : a \in \mathbb{R}^+\right\} \\
&= \frac{\partial}{\partial c_g} \mathbb{E}\left[v(\Lambda_t, n + \min\{Q_t, a\})\right]\Big|_{a=a^*(n)} \\
&= \frac{\partial}{\partial c_g}\left[\int_0^a \mathbb{E}\left[v(\Lambda_t, n+q)\right] f_t(q)dq + \mathbb{E}\left[v(\Lambda_t, n+a)\right](1 - F_t(a))\right]\Big|_{a=a^*(n)} \\
&= \left[\int_0^a \frac{\partial}{\partial c_g}\mathbb{E}\left[v(\Lambda_t, n+q)\right] f_t(q)dq + \frac{\partial}{\partial c_g}\mathbb{E}\left[v(\Lambda_t, n+a)\right](1 - F_t(a))\right]\Big|_{a=a^*(n)}.
\end{aligned} \tag{3.24}$$

For an arbitrary λ, we then have

$$\begin{aligned}
\frac{\partial v(\lambda, n+q)}{\partial c_g} &= \frac{\partial}{\partial c_g} \min\{(n+q)(1+r_o c_o) + g c_g + l(\lambda, (n+q)(1+r_o) + g)c_w \\
&\qquad\qquad : g \in \mathbb{R}^+, g > \lambda - (n+q)(1+r_o)\Big\} \\
&= \frac{\partial}{\partial c_g}\left[(n+q)(1+r_o c_o) + g c_g + l(\lambda, (n+q)(1+r_o) + g)c_w\right]\Big|_{g=g^*(\lambda, n+q)} \\
&= g^*(\lambda, n+q)
\end{aligned}$$

Similarly, $\partial v(\lambda, n+a)/\partial c_g = g^*(\lambda, n+a)$. Substituting these derivatives into Equation 3.24, we obtain

$$\begin{aligned}\frac{\partial m(n)}{\partial c_g} &= \left[\int_0^a \mathbb{E}\left[g^*(\Lambda_t, n+q)\right] f_t(q)dq + \mathbb{E}\left[g^*(\Lambda_t, n+a)\right](1-F_t(a))\right]\Big|_{a=a^*(n)} \\ &= \int_0^{a^*(n)} \mathbb{E}\left[g^*(\Lambda_t, n+q)\right] f_t(q)dq + \mathbb{E}\left[g^*(\Lambda_t, n+a^*(n))\right](1-F_t(a^*(n))),\end{aligned}$$

which is clearly non-negative.

Similarly, since

$$\begin{aligned}\frac{\partial v(\lambda, n+q)}{\partial c_w} &= \frac{\partial}{\partial c_w}\left[(n+q)(1+r_o c_o) + g c_g + l(\lambda, (n+q)(1+r_o)+g)c_w\right]\Big|_{g=g^*(\lambda,n+q)} \\ &= l(\lambda, (n+q)(1+r_o) + g^*(\lambda, n+q)),\end{aligned}$$

and

$$\frac{\partial v(\lambda, n+a)}{\partial c_w} = l(\lambda, (n+q)(1+r_o) + g^*(\lambda, n+a)),$$

we have

$$\begin{aligned}\frac{\partial m(n)}{\partial c_w} = \int_0^{a^*(n)} &\mathbb{E}\left[l(\Lambda_t, (n+q)(1+r_o) + g^*(\Lambda_t, n+q))\right] f_t(q)dq \\ &+ \mathbb{E}\left[l(\Lambda_t, (n+a^*(n))(1+r_o) + g^*(\Lambda_t, n+a^*(n)))\right](1-F_t(a^*(n))),\end{aligned}$$

which is non-negative.

For c_o, since

$$\frac{\partial v(\lambda, n+q)}{\partial c_o} = (n+q)r_o,$$

and

$$\frac{\partial v(\lambda, n+a)}{\partial c_o} = (n+a)r_o,$$

we have

$$\frac{\partial m(n)}{\partial c_o} = \int_0^{a^*(n)} (n+q)r_o f_t(q)dq + (n+a^*(n))r_o(1-F_t(a^*(n))),$$

which is non-negative and the proof is complete. □

3.3 Special Cases

In this section, we consider three common queueing models for the system serving patients' requests. We have chosen these models as their system-size functions are available in closed form. Further, extensions of these functions to non-integral server numbers are already available. In the first model, the system is represented by an $M/M/1$ queue — with Exponential independent and identically distributed (i. i. d.) inter-arrival times and services times that are independent, and a single server — whose service rate is inflated by

the number of servers. This is a common approximation in queueing optimization models; see, e.g., Mandelbaum and Reiman (1998) and Anily and Haviv (2010). The single-server approximation model behaves exactly as the original multi-server system when the number of customers in the system is equal to or larger than the number of servers. When this is not the case, the single-server approximation overestimates the system performance because it consolidates all service capacity into one server. This is less likely to happen when traffic intensity, i.e., the ratio of the arrival rate to service rate, is high. The advantage of this approximation is that it leads to explicit equations for congestion measures that can be applied with non-integral server numbers. In particular, for the $M/M/1$ approximation model,

$$l(\lambda, s) = \frac{\lambda}{(s - \lambda)}, \tag{3.25}$$

with $\lambda > 0$ and $s \in \mathbb{R}^+$ with $s > \lambda$. All properties of Assumption 1 are easily verified for this model. We then have the following proposition.

Proposition 3. *For the special case of $M/M/1$ approximation,*

$$\tilde{\lambda}(p) = p(1 + r_o) + \frac{c_w - \sqrt{4c_g c_w p(1 + r_o) + c_w^2}}{2c_g}, \tag{3.26}$$

$$\tilde{g}(\lambda, p) = \lambda + \sqrt{\frac{c_w \lambda}{c_g}} - p(1 + r_o), \tag{3.27}$$

and

$$v(\lambda, p) = \begin{cases} p(1 + r_o c_o) + \dfrac{\lambda c_w}{p(1 + r_o) - \lambda} & \text{if } \lambda \leq \tilde{\lambda}(p) \ , \\ \left(-c_g(1 + r_o) + 1 + r_o c_o\right) p + c_g \lambda + 2\sqrt{c_g c_w \lambda} & \text{if } \lambda > \tilde{\lambda}(p). \end{cases} \tag{3.28}$$

Proof. Taking the derivative of $l(\lambda, s)$ given in Equation (3.25) with respect to s, plugging it into Equation (3.4), and setting the result equal to zero yields

$$c_g - \frac{c_w x}{(p(1 + r_o) - x)^2} = 0.$$

Solving the above equation for $x \in (0, p(1 + r_o))$, we obtain

$$x = \tilde{\lambda}(p) = p(1 + r_o) + \frac{c_w - \sqrt{4 c_g c_w p(1 + r_o) + c_w^2}}{2 c_g}.$$

Inserting the derivative of $l(\lambda, s)$ with respect to s in Equation (3.5), setting the result equal to zero and solving for g, we obtain the value for $\tilde{g}(\lambda, p)$ given in Equation (3.27). Equation (3.28) is then obtained by evaluating the objective function in problem (3.1) for $g = g^*(\lambda, p)$. □

Equation (3.27) implies that the number of temporary workers when $\lambda > \tilde{\lambda}(p)$ is obtained from an expression analogous to the square-root staffing law (see, e.g., Halfin and Whitt, 1981), according to which the staffing requirement is equal to the offered load (λ in our setting) plus a service quality coefficient multiplied by the square-root of the offered load. The service-quality coefficient appears as $\sqrt{c_w / c_g}$ in our formula. An adjustment is also made to account for the number of permanent workers. The expression given in (3.27)

replaces the numerical procedure for obtaining $\tilde{g}(\lambda, p)$ in Algorithm 1, and the expression given in (3.26) replaces the function provided in Algorithm 1 for evaluating $\tilde{\lambda}(p)$.

For the second model, we assume that the dynamics of service delivery are captured by an inflated $M/G/1$ queue, with G representing a general distribution for service times. The system-size function in this setting is

$$l(\lambda, s) = \frac{1+\tau^2}{2}\frac{\lambda^2}{s(s-\lambda)} + \frac{\lambda}{s}, \tag{3.29}$$

where τ is the coefficient of variation (CV), i.e., the ratio of standard deviation to mean, of the service time distribution (Gross et al., 2008). We then have the following proposition for $M/G/1$ queues.

Proposition 4. *The system-size function $l(\lambda, s)$ given in Equation* (3.29) *meets the properties given in Assumption 1.*

Proof. Properties A(i), A(ii), and A(iii) are easy to verify. For property A(iv), we obtain the second derivative of $l(\lambda, s)$ given in Equation (3.29) with respect to s as

$$\frac{\partial^2 l(\lambda, s)}{\partial s^2} = \left(\frac{1+\tau^2}{2}\right)\left(\frac{2\lambda^2(3s(s-\lambda)+\lambda^2)}{(s^2-\lambda s)^3}\right) + \frac{2\lambda}{s^3},$$

which is positive when $s > \lambda$. For property A(v), we have

$$\frac{\partial^2 l(\lambda, s)}{\partial \lambda \partial s} = \left(\frac{1+\tau^2}{2}\right)\left(\frac{-\lambda s[(2s-\lambda)^2+\lambda s]}{(s^2-\lambda s)^3}\right) - \frac{1}{s^2},$$

which is negative when $s > \lambda$. □

We prove an important result for $M/G/1$ queues in the following corollary. It implies that, for a given n, a higher variability in service time distribution is compensated with a larger number of permanent positions advertised. Similarly, for given λ and p, a higher variability in service time results in a larger number of temporary HCWs. The first- and second-stage optimal cost functions also increase with τ.

Corollary 6. *In $M/G/1$ queues, $g^*(\lambda, p)$ and its corresponding cost function, i.e., $v(\lambda, p)$, as well as $a^*(n)$ and its corresponding cost function, i.e., $m(n)$, all increase with τ.*

Proof. Taking the derivative of $l(\lambda, s)$ given in Equation (3.29) with respect to τ, we obtain

$$\frac{\partial l(\lambda, s)}{\partial \tau} = \frac{\lambda^2 \tau}{s\,(s-\lambda)}. \tag{3.30}$$

This gives

$$\frac{\partial v(\lambda, p)}{\partial \tau} = \frac{c_w \tau\, \lambda^2}{\left(p\,(1+r_o) + g^*(\lambda, p)\right)\left(p\,(1+r_o) + g^*(\lambda, p) - \lambda\right)},$$

which, given the stability constraint, is always positive, and so $v(\lambda, p)$ is increasing in τ. Taking the derivative of $l(\lambda, s)$ given in Equation (3.29) with respect to τ and s, we obtain

$$\frac{\partial^2 l(\lambda, s)}{\partial \tau \partial s} = \frac{\lambda^2 \tau\,(\lambda - 2s)}{s^2\,(\lambda - s)^2} \tag{3.31}$$

which is negative. This implies that $\tilde{\lambda}(p)$, the route of function $\phi_p(x)$ given in (3.4) in the

interval $(0, p(1+r_o))$, is strictly decreasing in τ. It also implies that $\tilde{g}(\lambda, p)$, the unique root of function $\theta_{\lambda,p}(g)$ given in (3.5) in the interval $((\lambda - p(1+r_o))^+, \infty)$, is strictly increasing in τ. We conclude that, for given λ and p, $g^*(\lambda, p)$ is increasing in τ. A similar argument applies for $m(n)$ and $a^*(n)$. □

The third model that we consider is an $M/M/s$ queueing model. The mean number of requests in this system is evaluated as

$$l(\lambda, s) = \frac{\lambda C(\lambda, s)}{s - \lambda} + \lambda, \tag{3.32}$$

where $C(\lambda, s)$ is a continuous extension of the Erlang delay function such as

$$C(\lambda, s) = \left(\int_0^\infty \lambda e^{-\lambda x} (1+x)^{s-1} x dx \right)^{-1}, \tag{3.33}$$

for each $\lambda > 0$ and $s \in \mathbb{R}^+$ with $s > \lambda$ as defined by Jagers and van Doorn (1991). For this system, we propose

Proposition 5. *The system-size function $l(\lambda, s)$ given in Equation* (3.32) *meets the properties given in Assumption 1.*

Proof. Property A(i) is easy to verify. For property A(ii), note that $\lim_{\lambda \downarrow 0} C(\lambda, s) = 0$, $\lim_{s \to \infty} C(\lambda, s) = 0$, and

$$\lim_{\lambda \uparrow s^-} C(\lambda, s) = \lim_{s \downarrow \lambda^+} C(\lambda, s) = \lim_{\lambda \uparrow s^-} \frac{B(\lambda, s)}{1 - \lambda(1 - B(\lambda, s))/s} = \frac{B(\lambda, s)}{B(\lambda, s)} = 1, \tag{3.34}$$

where the second equality is by the relation between delay probability $C(\lambda, s)$ in $M/M/s$ queues and blocking probability $B(\lambda, s)$ in the associated $M/M/s/0$ loss queues (with 0 representing the waiting space). For the first part of property A(iii), note that

$$\frac{\partial l(\lambda, s)}{\partial \lambda} = \frac{\left(C(\lambda, s) + \lambda \frac{\partial C(\lambda, s)}{\partial \lambda}\right)(s - \lambda) + \lambda C(\lambda, s)}{(s - \lambda)^2} + 1. \tag{3.35}$$

Using the relation between $C(\lambda, s)$ and $B(\lambda, s)$ given in Equation (3.34), we then have

$$\frac{\partial C(\lambda, s)}{\partial \lambda} = \frac{(\partial B(\lambda, s)/\partial \lambda)(1 - \lambda/s) + 1/s(1 - B(\lambda, s))B(\lambda, s)}{[1 - (\lambda/s)(1 - B(\lambda, s))]^2},$$

which is non-negative because $B(\lambda, s)$ is increasing in λ (Pacheco, 1993), $\lambda < s$, and $B(\lambda, s) \leq 1$. Hence, the derivative given in Equation (3.35) is positive. For the second part of property A(iii) and property A(iv), Karsten et al. (2015) prove that the expected sojourn time, denoted by $w(\lambda, s)$, is strictly decreasing and strictly convex in s for $M/M/s$ queues. Since $l(\lambda, s) = \lambda w(\lambda, s)$, the same properties apply for $l(\lambda, s)$. Property A(v) is in fact economies of scale as explained in §3.2. This has already been proved in the extant literature; see, e.g., Karsten et al. (2015). □

The $M/M/1$ queue is useful for obtaining rough estimates of optimal decisions with minimum computational effort. The $M/G/1$ queue captures the impact of service time variability, while the $M/M/s$ queue represents the impact of system scale accurately. More computational effort is needed for the last two models, however, as closed-form expressions for $\tilde{\lambda}(p)$ and $\tilde{g}(\lambda, p)$ cannot be provided due to the complexity of the derivatives

of corresponding system-size functions.

3.4 Savings Evaluation

In this section, we assess the savings obtained from our model when compared to a single-stage model with no temporary recruitment and a two-stage model in which the uncertainty in demand rate is ignored. In our experiments, we assume that Λ_t follows a Gamma distribution with mean ξ and CV κ. This assumption is motivated by the study of Jongbloed and Koole (2001), and is verified empirically in our case study in §3.6. For Q_t, Pinker and Tilson (2013) propose a Poisson distribution. Since we need a continuous distribution, however, we use a Log-Normal distribution with mean μ_r and CV κ_r instead.

3.4.1 Comparison with a Single-Stage Model with No Temporary Recruitment

Consider a single-stage model in which the provider has to decide the number of permanent positions to advertise at time t knowing that there is no opportunity for temporary recruitment. This is formulated as

$$m_{single}(n) = \min_{a} \Big\{ \mathbb{E}\left[(n + \min\{Q_t, a\})(1 + r_o c_o) + \right.$$
$$\left. l(\Lambda_t, (n + \min\{Q_t, a\})(1 + r_o)) c_w | \mathbb{S}\right] :$$
$$\mathbb{P}(\mathbb{S}) \geq \gamma; a \in \mathbb{R}^+ \Big\}, \quad (3.36)$$

where $\mathbb{S}$ is the event of the system being stable, and γ is the minimum probability of this event as set by the decision maker. In (3.36), we condition the expected value in the objective function on $\mathbb{S}$ and add the corresponding constraint to the optimization model as in the single-stage decision making with uncertain demand rate, there is a likelihood that the system becomes unstable for any value of a (unless h_t has a bounded support). The stability condition is represented mathematically as $\Lambda_t < (n + \min\{Q_t, a\})(1 + r_o)$, and its probability can be evaluated for any given a by the law of total probability.

We evaluate the savings obtained from our two-stage model as compared to the single-stage model with no temporary recruitment. In particular, we investigate the impact on savings of demand rate uncertainty, as measured by its CV κ, for three different scale scenarios, $\xi = 10.0$, $\xi = 50.0$, and $\xi = 100.0$, using an $M/M/s$ queue. We use the same queue to also investigate the impact of c_g, c_w, c_o, r_o, and n on savings. We set $\mu_r = 10\xi$ and $\kappa_r = 0.5$ in our experiments to minimize the impact on savings of recruitment restrictions as these will be investigated separately. For each set of parameters, we evaluate the optimal cost of the two-stage model by inserting the optimal first-stage decision returned by Algorithm 2 in the objective function of the first-stage problem given in (3.9). The optimal solution to (3.36) is estimated by complete enumeration over values of $a \in [0.0, 0.1, \ldots, 5\xi]$. The cost calculations are performed with 30 digits numerical precision. The results are plotted in Figure 3.2 for $\gamma = 0.95$.

The plots in panel (a) of Figure 3.2 suggest that savings from our model typically increase with the system scale and the level of uncertainty in demand rate, exceeding 10.0%

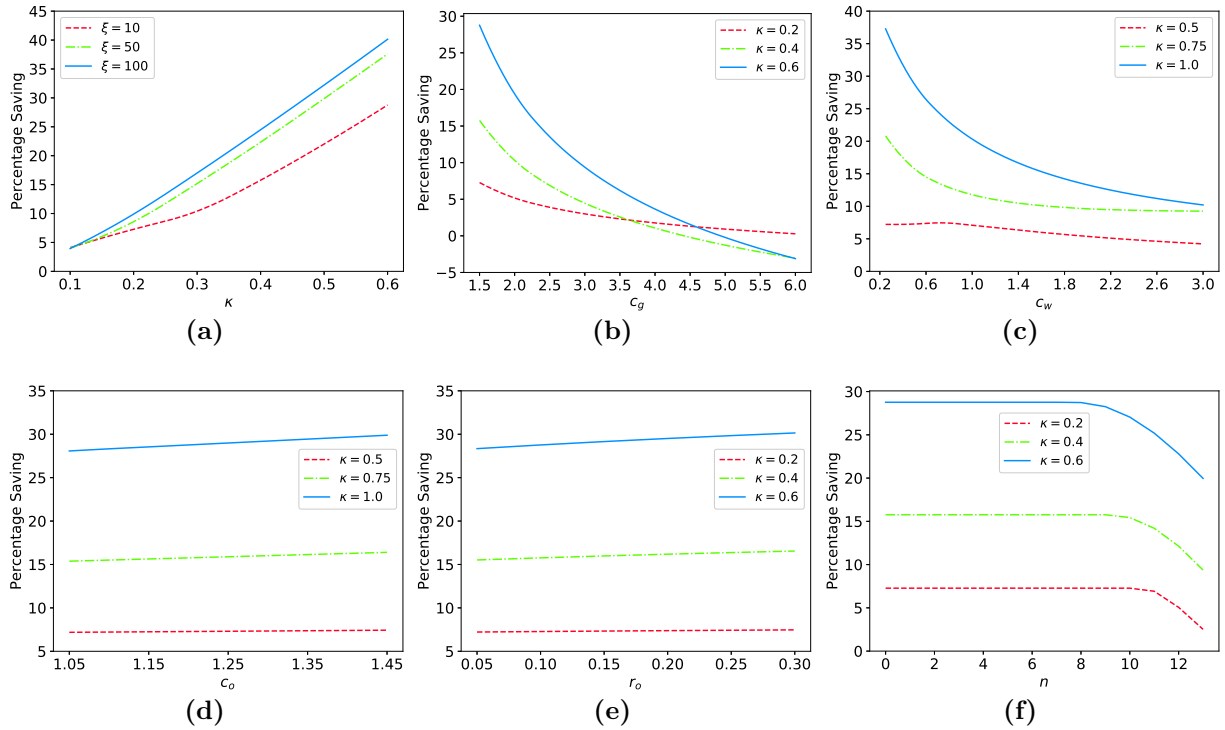


Figure 3.2: The savings of our model as compared to the single-stage model using an $M/M/s$ queue. The parameters not given in the plots are $\xi = 10.0$, $c_g = 1.5$, $c_w = 0.5$, $c_o = 1.2$, $r_o = 0.1$, $n = 0$, $\mu_r = 10\xi$, and $\kappa_r = 0.5$.

for $\kappa \geq 0.3$ and $\xi \geq 10$. They also indicate that savings of at least 3.9% are likely to be gained with all three scale scenarios even when demand rate uncertainty is very low, i.e., $\kappa \approx 0.1$. This is a substantial amount of saving given the high share of staffing cost in healthcare expenditure (see, e.g., The Kings Fund, 2021). Panel (b) suggests that savings reduce with the cost rate of temporary HCWs, becoming negative for $c_g \geq 4.5$ and $c_g \geq 5.0$ with $\kappa = 0.4$ and $\kappa = 0.6$, respectively. This implies that the single-stage model may result in a lower cost than the two-stage model when c_g is very large (bear in mind that there always exists a risk of the system becoming unstable with the single-stage model.) Panel (c) implies that savings typically decrease, but remain positive, as c_w increases. Panels (d)

and (e) show mildly increasing trends for savings with respect to c_o and r_o, respectively. Panel (f) suggests that savings are initially stable with respect to n, but then decrease as n goes beyond a threshold. This is because for n larger than a threshold, there is no need for temporary recruitment and so the difference between the two models reduces.

We perform another set of experiments with an $M/G/1$ queue to investigate the impact of service time variability, as measured by its CV τ, as well as recruitment parameters, μ_r and κ_r. The results are presented in Figure 3.3 for $\xi = 10$. Panel (a) of this figure shows that savings vary from 2.8% for $\kappa = 0.2$ and $\tau = 5.0$ to 39.4% for $\kappa = 0.6$ and $\tau = 0.0$. Panels (b) and (c) suggest that savings are almost insensitive to μ_r and κ_r. This is mainly because, in order to meet the stability constraint with reasonably small values of a, we must have $\mu_r \geq 20$ for the set of parameters considered. As the starting value for μ_r is already large, savings do not change as μ_r or κ_r increase.

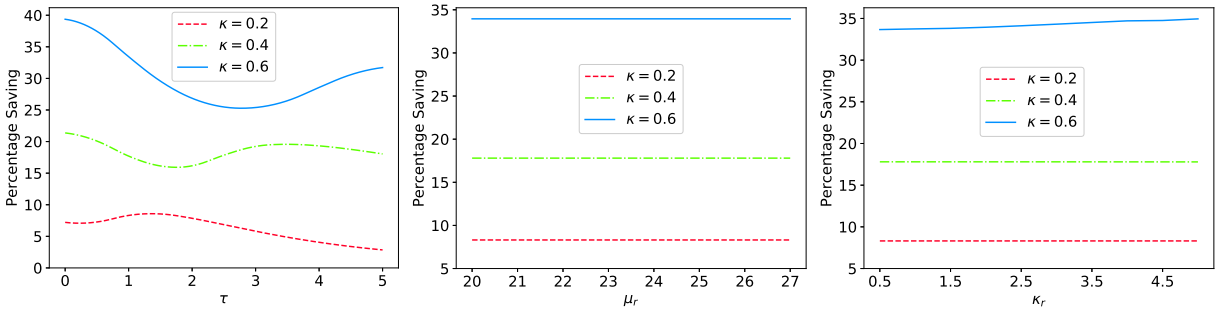


Figure 3.3: The savings of our model as compared to the single-stage model using an $M/G/1$ queue. The parameters are $\xi = 10.0$, $c_g = 1.5$, $c_w = 0.5$, $c_o = 1.2$, $r_o = 0.1$, $n = 0$, and (a) $\mu_r = 10\xi$, $\kappa_r = 0.5$, (b) $\tau = 1.0$, $\kappa_r = 0.5$, and (c) $\tau = 1.0$ and $\mu_r = 20.0$.

3.4.2 Comparison with a Two-Stage Model with No Demand Rate Uncertainty

We consider a two-stage optimization framework similar to §3.2, but assume that the decision maker ignores the demand rate uncertainty at time t, and works with the expected demand rate, denoted by ξ. The first-stage problem then simplifies to

$$m(n) = \min_{a}\{\mathbb{E}[v\left(\xi, n + \min\{Q_t, a\}\right)] : a \in \mathbb{R}^+\}, \tag{3.37}$$

and its solution is obtained through the following proposition.

Proposition 6. *The optimal solution to the first-stage problem with no demand rate uncertainty as given in* (3.37) *is obtained from Proposition 2 with* $\psi_n(a)$ *simplified as*

$$\psi_n(a) = 1 + r_o c_o + \begin{cases} -c_g(1+r_o), & \tilde{\lambda}(a+n) < \xi, \\ c_w(1+r_o)\dfrac{\partial l(\xi, s)}{\partial s}\Big|_{s=(n+a)(1+r_o)}, & \tilde{\lambda}(a+n) \geq \xi, \end{cases} \tag{3.38}$$

In addition, when the service delivery is represented by an $M/M/1$ *queue, we have*

$$\tilde{a}(n) = \frac{\sqrt{\dfrac{\xi c_w(1+r_o)}{1+r_o c_o}} + \xi}{1+r_o} - n. \tag{3.39}$$

Proof. For the first part, we apply Proposition 2 noting that ignoring demand rate uncertainty is equivalent to assuming $\mathbb{P}(\Lambda_t = \xi) = 1.0$. This implies that when ξ is smaller than

$\tilde{\lambda}(n+a)$, the third term in Equation (3.10) simplifies to

$$c_w(1+r_o)\frac{\partial l(\xi,s)}{\partial s}\Big|_{s=(n+a)(1+r_o)},$$

and the fourth term simplifies to zero. Similarly, when ξ is larger than $\tilde{\lambda}(n+a)$, the fourth term in Equation (3.10) simplifies to $-c_g(1+r_o)$ and the third term simplifies to zero. These yield the expression in Equation (3.38).

For the second part, the expression given for $\tilde{a}(n)$ is obtained by replacing the derivative of $l(\lambda, s)$ given in (3.25) with respect to s in Equation (3.38) and solving $\psi_n(a) = 0.0$ for a. □

We evaluate the savings obtained from our model as compared to the model with no demand rate uncertainty. In particular, we investigate the impact on savings of κ for three different values of ξ using an $M/M/s$ queue. We use the same queue to also investigate the impact of c_g, c_w, c_o, r_o, and n on savings. For the same reason as in §3.4.1, we set $\mu_r = 10\xi$ and $\kappa_r = 0.5$. For the model with demand rate uncertainty, the cost is evaluated as explained in §3.4.1. For the model with no demand rate uncertainty, the optimal cost is evaluated by inserting the optimal a produced by Proposition 6 in the objective function of the first-stage problem given in (3.9) The results are plotted in Figure 3.4.

Panel (a) of Figure 3.4 suggests that savings will be small when demand rate uncertainty is low and system scale is small. As the scale and/or demand rate uncertainty grow, however, the savings are likely to increase, exceeding 2.5% for a moderate demand rate

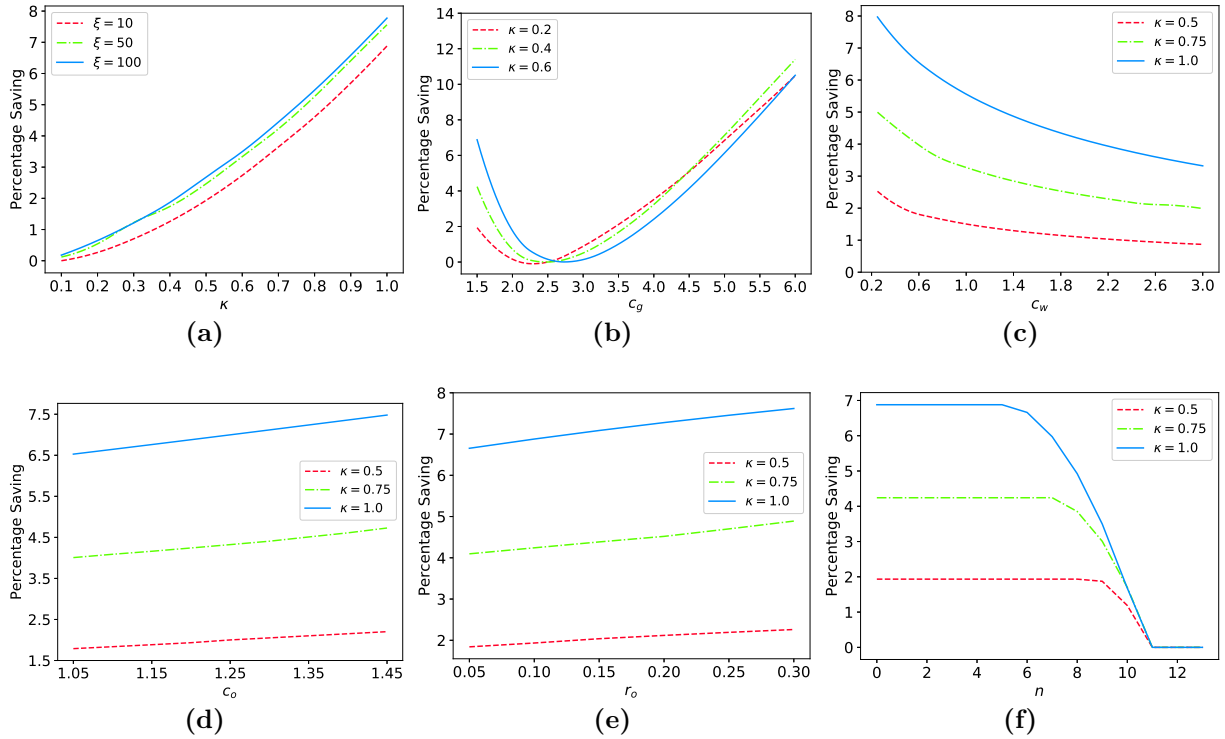


Figure 3.4: The savings of our model as compared to the model with no demand rate uncertainty using an $M/M/s$ queue. The parameters not given in the plots are $\xi = 10.0$, $c_g = 1.5$, $c_w = 0.5$, $c_o = 1.2$, $r_o = 0.1$, $n = 0$, $\mu_r = 10\xi$, and $\kappa_r = 0.5$.

uncertainty, i.e., $\kappa \approx 0.5$, and a medium system, i.e., $\xi \approx 50.0$. Panel (b) of Figure 3.4 suggests that savings show a non-monotone behaviour with respect to c_g, initially decreasing but then increasing. The increase of savings for values of c_g larger than a threshold can be explained by noticing that obtaining the demand rate distribution at the time of permanent advertisement leads to a more successful permanent recruitment, which in turn, helps with relying less on a highly expensive temporary HCWs at the second stage. Panel (c) illustrates a decreasing trend for savings with respect to c_w. Panels (d) and (e) suggest slowly increasing trends for c_o and r_o, respectively. Panel (f) implies that savings reduce with n, becoming 0.0 for $n \geq 11.0$. This is because for larger values of n, there are already

enough workers in the system to respond to the possibility of a higher than expected demand rate, and thus the two models become closer.

We perform another set of experiments with an $M/G/1$ queue to investigate the impact of τ, μ_r and κ_r. The results are presented in Figure 3.5 for $\xi = 10.0$. Panel (a) of this figure shows that savings decrease with τ, while panel (b) suggests that they increase with μ_r up to a threshold, then stabilize. Panel (c) shows a mild decreasing trend for κ_r.

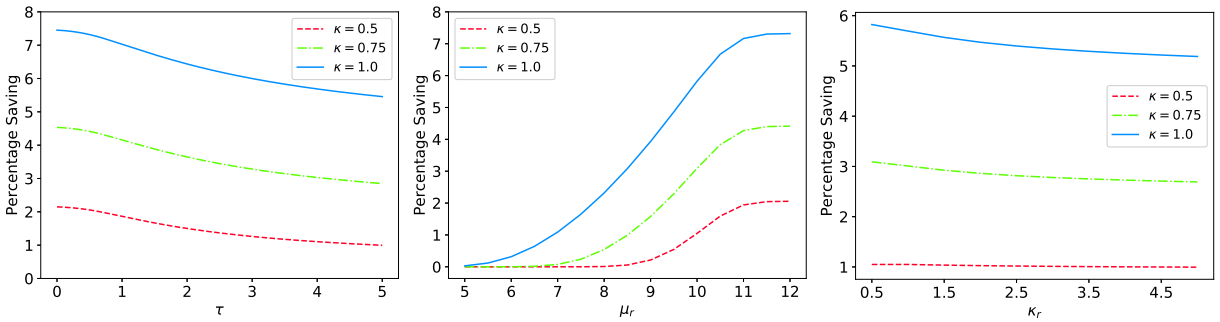


Figure 3.5: The savings of our model as compared to the model with no demand rate uncertainty using an $M/G/1$ queue. The parameters are $\xi = 10.0$, $c_g = 1.5$, $c_w = 0.5$, $c_o = 1.2$, $r_o = 0.1$, $n = 0$, and (a) $\mu_r = 10\xi$, $\kappa_r = 0.5$, (b) $\tau = 1.0$, $\kappa_r = 0.5$, and (c) $\tau = 1.0$ and $\mu_r = 10.0$.

Overall, when demand rate uncertainty is moderate to high, our two-stage approach is likely to be beneficial. When demand rate uncertainty is low, on the other hand, the simplified version of our two-stage approach which uses only the average demand rate (as illustrated in Proposition 6) would suffice. Furthermore, except in situations where c_g is extremely high, temporary recruitment is likely to provide value, even if demand rate uncertainty is very low. This value is likely to increase with the system scale, but decrease with c_w.

3.5 Delaying Advertisement

In §3.2, we assumed that advertisement for permanent HCWs must occur at time t due to external factors. In this section, we consider the possibility of delaying the advertisement beyond t. This is because one would expect that, as advertisement is delayed, there will be a lower level of uncertainty for demand rate. The risk, however, is that with a shorter window for advertising, a smaller number of qualified applications may be received. We investigate this trade-off. We note that this is an important investigation, which to the best of our knowledge, has not been covered in the literature.

We consider the reduction in the number of applications and the reduction in demand rate uncertainty by assuming that, for $t \leq t' < t_e$, $Q_{t'} \leq_{st} Q_t$ and $\Lambda_{t'} \leq_{cx} \Lambda_t$, respectively, where $X \leq_{st} Y$ denotes that X is smaller than Y in the *usual stochastic order*, and $X \leq_{cx} Y$ denotes that X is smaller than Y in the *convex order* (see, e.g., Shaked and Shanthikumar, 2007). To avoid unnecessary complication, we further assume that the pdfs of $Q_{t'}$ and $\Lambda_{t'}$ have the same support as those of Q_t and Λ_t, respectively. Roughly, $Q_{t'} \leq_{st} Q_t$ states that Q_t is more likely to take on large values than $Q_{t'}$, whereas $\Lambda_{t'} \leq_{cx} \Lambda_t$ implies that Λ_t is more likely to take on extreme values than $\Lambda_{t'}$. $\Lambda_{t'} \leq_{cx} \Lambda_t$ also implies that $\mathbb{E}[\Lambda_t] = \mathbb{E}[\Lambda_{t'}] = \xi$, which is consistent with §3.2. In order to show the dependence of the optimal first-stage decision and its cost on Q_t and Λ_t, we expand the corresponding notations defined in §3.2 to $a^*(n, Q_t, \Lambda_t)$ and $m(n, Q_t, \Lambda_t)$, respectively. We first analyze the impact of reduction in application numbers and demand rate uncertainty separately.

From corollary 3, we know that the optimal first-stage decision is not affected by the distribution of Q_t, i.e., $a^*(n, Q_t, \Lambda_t) = a^*(n, Q_{t'}, \Lambda_t)$. The optimal cost, however, decreases as a result of Q_t increasing in the usual stochastic order by the following proposition.

Proposition 7. *Suppose* $Q_{t'} \leq_{st} Q_t$, *then* $m(n, Q_t, \Lambda_t) \leq m(n, Q_{t'}, \Lambda_t)$.

Proof. First, note that $Q_{t'} \leq_{st} Q_t$ implies that $\mathbb{E}\left[\omega(Q_{t'})\right] \geq \mathbb{E}[\omega(Q_t)]$ for any decreasing function $\omega(x)$. Also, note that, by insensitivity of $a^*(n, Q_t, \Lambda_t)$ to the pdf of Q_t, we have

$$m(n, Q_t, \Lambda_t) = \mathbb{E}\left[v(\Lambda_t, n + \min\{Q_t, a^*(n, Q_t, \Lambda_t)\})\right],$$

and

$$m(n, Q_{t'}, \Lambda_t) = \mathbb{E}\left[v(\Lambda_t, n + \min\{Q_{t'}, a^*(n, Q_t, \Lambda_t)\})\right].$$

Hence, it suffices to show that

$$\omega(q) \triangleq \mathbb{E}\left[v(\Lambda_t, n + \min\{q, a^*(n, Q_t, \Lambda_t)\})\right],$$

is decreasing in q. For $q \geq a^*(n, Q_t, \Lambda_t)$, $\dfrac{\partial \omega(q)}{\partial q} = \dfrac{\partial}{\partial q}\mathbb{E}[v(\Lambda_t, n + a^*(n, Q_t, \Lambda_t))] = 0$. For $0 \leq q < a^*(n, Q_t, \Lambda_t)$, on the other hand, we have

$$\begin{aligned}
\frac{\partial \omega(q)}{\partial q} &= \frac{\partial}{\partial q}\mathbb{E}\left[v(\Lambda_t, n + q)\right] \\
&= 1 + r_o c_o + c_w(1 + r_o) \int_0^{\tilde{\lambda}(n+q)} \frac{\partial l(\lambda, s)}{\partial s}\Big|_{s=(n+q)(1+r_o)} h_t(\lambda) d\lambda \\
&\quad - c_g(1 + r_o) \int_{\tilde{\lambda}(n+q)}^{\infty} h_t(\lambda) d\lambda, \qquad (3.40)
\end{aligned}$$

where the second equality is obtained by replacing a with q in Equation (3.17). The expression given in (3.40) is in fact $\psi_n(q)$, which by Proposition 2 is negative when $0 \leq q < a^*(n, Q_t, \Lambda_t)$. Hence, $\omega(q)$ is decreasing in q and the proof is complete. □

The impact of Λ_t is more complex. When λ_u is finite, the following propositions set out the conditions under which the optimal first-stage decision and the corresponding cost show a monotone behaviour as Λ_t increases in the convex order.

Proposition 8. *Suppose* $\Lambda_{t'} \leq_{cx} \Lambda_t$. *Then* $a^*(n, Q_t, \Lambda_{t'}) \leq a^*(n, Q_t, \Lambda_t)$ *if*

(a) $\lambda_u \leq \tilde{\lambda}(n)$, *and*

(b) $\dfrac{\partial^3 l(\lambda, s)}{\partial s \partial \lambda^2} \leq 0.$

Proof. $\Lambda_{t'} \leq_{cx} \Lambda_t$ implies that $\mathbb{E}[\omega(\Lambda_t)] \leq \mathbb{E}[\omega(\Lambda_{t'})]$ for any concave function $\omega(\lambda)$. Define $\omega(\lambda) \triangleq \dfrac{\partial v(\lambda, n+a)}{\partial a}$, and note that $\psi_n^{\Lambda_t}(a) = \mathbb{E}\left[\omega(\Lambda_t)\right]$, where we have expanded the notation for $\psi_n(a)$ defined in Equation (3.10) to indicate its dependence to Λ_t. From (3.15), (3.3), and the fact that $\tilde{g}(\lambda, p)$ is the roof of $\theta_{\lambda,p}(g)$, we now have

$$\omega(\lambda) \triangleq \frac{\partial v(\lambda, n+a)}{\partial a} = \begin{cases} 1 + r_o c_o + c_w(1+r_o) \dfrac{\partial l(\lambda, s)}{\partial s}\Big|_{s=(n+a)(1+r_o)} & \lambda \leq \tilde{\lambda}(n+a), \\ 1 + r_o c_o - c_g(1+r_o) & \tilde{\lambda}(n+a) < \lambda \leq \lambda_u. \end{cases}$$

The first and second derivatives are

$$\frac{\partial \omega(\lambda)}{\partial \lambda} = \begin{cases} c_w(1+r_o) \dfrac{\partial^2 l(\lambda, s)}{\partial s \partial \lambda}\Big|_{s=(n+a)(1+r_o)} & \lambda \leq \tilde{\lambda}(n+a), \\ 0 & \tilde{\lambda}(n+a) < \lambda \leq \lambda_u, \end{cases}$$

and

$$\frac{\partial^2\omega(\lambda)}{\partial\lambda^2} = \begin{cases} c_w(1+r_o)\dfrac{\partial^3 l(\lambda,s)}{\partial s\partial\lambda^2}\Big|_{s=(n+a)(1+r_o)} & \lambda \le \tilde{\lambda}(n+a), \\ 0 & \tilde{\lambda}(n+a) < \lambda \le \lambda_u, \end{cases}$$

respectively. By condition (b) in the proposition, $\dfrac{\partial^2\omega(\lambda)}{\partial\lambda^2}$ is non-positive. By condition (a) and Lemma 3.2.4, $\lambda_u \le \tilde{\lambda}(n) \le \tilde{\lambda}(n+a)$, and so $\lambda \le \tilde{\lambda}(n+a)$ for any value $a \in \mathbb{R}^+$, hence, $\partial\omega(\lambda)/\partial\lambda$ is continuous on $\lambda \in [0, \lambda_u]$. From these, we conclude that $\omega(\lambda)$ is concave when the conditions of the proposition are met (Note that, without condition (a), the first derivative would not be continuous, and so $\omega(\lambda)$ would not be concave.) As such, when $\psi_n^{\Lambda_t}(0) \ge 0$, we will also have $\psi_n^{\Lambda_{t'}}(0) \ge 0$, thus $a^*(n, Q_t, \Lambda_t) = a^*(n, Q_t, \Lambda_{t'}) = 0$. On the other hand, when $\psi_n^{\Lambda_t}(0) < 0$, we will either have $\psi_n^{\Lambda_{t'}}(0) \ge 0$, which implies that $a^*(n, Q_t, \Lambda_{t'}) = 0 < a^*(n, Q_t, \Lambda_t)$, or $\psi_n^{\Lambda_t}(0) \le \psi_n^{\Lambda_{t'}}(0) < 0$, which implies that $a^*(n, Q_t, \Lambda_{t'}) \le a^*(n, Q_t, \Lambda_t)$. □

Proposition 9. *Suppose* $\Lambda_{t'} \le_{cx} \Lambda_t$. *Then* $m(n, Q_t, \Lambda_{t'}) \le m(n, Q_t, \Lambda_t)$ *if*

(a) $\lambda_u \le \tilde{\lambda}(n)$, *and*

(b) $\dfrac{\partial^2 l(\lambda, s)}{\partial\lambda^2} \ge 0$.

Proof. First, note that $\Lambda_{t'} \le_{cx} \Lambda_t$ implies that $\mathbb{E}[\omega(\Lambda_{t'})] \le \mathbb{E}[\omega(\Lambda_t)]$ for any convex function $\omega(\lambda)$. Second, note that $m(n, Q_t, \Lambda_{t'}) \le \mathbb{E}[v(\Lambda_{t'}, n + \min\{Q_t, a\})]$ for all $a \in \mathbb{R}^+$. Hence, it suffices to show that $\omega(\lambda) \triangleq \mathbb{E}[v(\lambda, n + \min\{Q_t, a\})]$ is a convex function of λ when the conditions of the proposition are met. To show the convexity of $\omega(\lambda)$, we obtain the first and second derivatives as

$$\frac{\partial \omega(\lambda)}{\partial \lambda} = \mathbb{E}\left[\frac{\partial}{\partial \lambda} v(\lambda, n + \min\{Q_t, a\})\right]$$
$$= \mathbb{E}\left[c_w \frac{\partial}{\partial \lambda} l(\lambda, (n + \min\{Q_t, a\})(1 + r_o) + g^*(\lambda, n + \min\{Q_t, a\}))\right],$$

and

$$\frac{\partial^2 \omega(\lambda)}{\partial \lambda^2} = \mathbb{E}\left[c_w \left(\frac{\partial^2}{\partial \lambda^2} l(\lambda, (n + \min\{Q_t, a\})(1 + r_o) + g^*(\lambda, n + \min\{Q_t, a\}))\right.\right.$$
$$\left.\left. + \frac{\partial^2 l(\lambda, s)}{\partial \lambda \partial s}\Big|_{s=(n+\min\{Q_t,a\})(1+r_o)+g^*(\lambda,n+\min\{Q_t,a\})} \times \frac{\partial}{\partial \lambda} g^*(\lambda, n + \min\{Q_t, a\})\right)\right].$$

It then follows from Proposition 1 that

$$\frac{\partial^2 \omega(\lambda)}{\partial \lambda^2} = c_w \mathbb{E}\left[\frac{\partial^2}{\partial \lambda^2} l(\lambda, (n + \min\{Q_t, a\})(1 + r_o) + \tilde{g}(\lambda, n + \min\{Q_t, a\}))\right.$$
$$+ \frac{\partial^2 l(\lambda, s)}{\partial \lambda \partial s}\Big|_{s=(n+\min\{Q_t,a\})(1+r_o)+\tilde{g}(\lambda,n+\min\{Q_t,a\})} \times \frac{\partial}{\partial \lambda} \tilde{g}(\lambda, n + \min\{Q_t, a\}),$$
$$\left. \lambda > \tilde{\lambda}(n + \min\{Q_t, a\}) \right]$$
$$+ c_w \mathbb{E}\left[\frac{\partial^2}{\partial \lambda^2} l(\lambda, (n + \min\{Q_t, a\})(1 + r_o)), \lambda \leq \tilde{\lambda}(n + \min\{Q_t, a\})\right].$$

By condition (a) and Lemma 3.2.4, $\lambda_u \leq \tilde{\lambda}(n) \leq \tilde{\lambda}(n + \min\{q, a\})$, and so $\lambda \leq \tilde{\lambda}(n + \min\{q, a\})$ for all $a, q \in \mathbb{R}^+$. This yields

$$\frac{\partial^2 \omega(\lambda)}{\partial \lambda^2} = c_w \mathbb{E}\left[\frac{\partial^2}{\partial \lambda^2} l(\lambda, (n + \min\{Q_t, a\})(1 + r_o))\right],$$

which is non-negative by condition (b) of the proposition. □

Condition (b) of Propositions 8 and 9 can be verified for $M/M/1$ and $M/G/1$ queues

analytically. Our numerical investigations also suggest that they hold for $M/M/s$ queues. Condition (a) is more restrictive as it imposes a relatively short interval for the support of h_t. For the special case of $M/M/1$ queues, for example, Equation (3.26) indicates that λ_u should be less than 4.45 for condition (a) to apply when $n = 5.0, c_g = 2.0, c_w = 0.5$ and $r_o = 0.1$. For smaller values of n or c_g, the upper bound λ_u would have to be even smaller.

For the general situation in which the support of h_t is unbounded, it is difficult to derive analytical results, hence, we resort to numerical experimentation. For this, we assume Λ_t follows a Gamma distribution with mean ξ and CV κ. Assuming an $M/G/1$ queue, we then obtain the optimal first-stage decision and the corresponding cost for increasing values of $\kappa \in [0.1, 3.0]$, while keeping ξ constant to ensure that Λ_t increases in the convex order (Belzunce et al., 2016). Figure 3.6 summarizes the results, and depicts a non-monotonic behaviour for the optimal first-stage decision (panels (a) to (c))) and its cost (panels (d) to (f)).

More specifically, the plots at the top of Figure 3.6 illustrate that there exists a threshold for κ, above (below) which $a^*(n)$ shows a decreasing (increasing) trend as Λ_t increases in the convex order. This highlights the different impact of demand rate uncertainty to that of service time variability. In particular, we proved in Corollary 6 that the optimal first- and second-stage decisions increase with service time variability. The results presented here imply that as the uncertainty in the demand rate increases up to a certain threshold, it is worth to invest in a larger number of permanent positions. Beyond this threshold, however, it is better to advertise a smaller number of permanent positions (and wait for

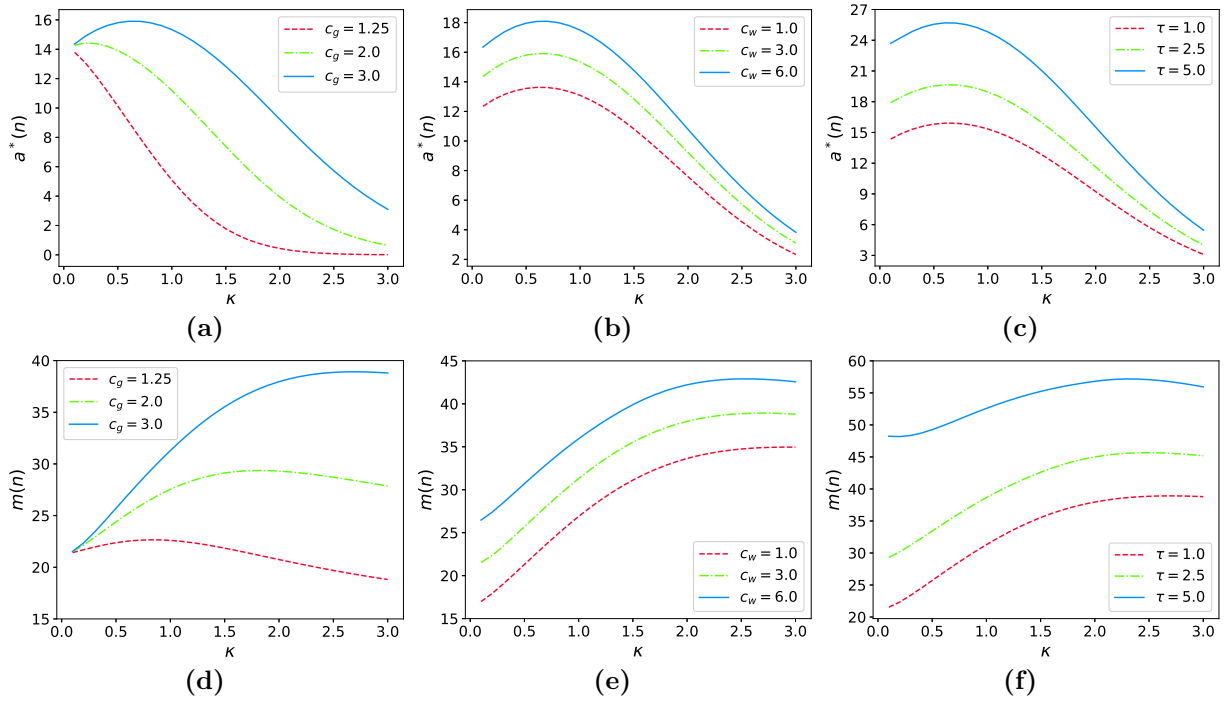
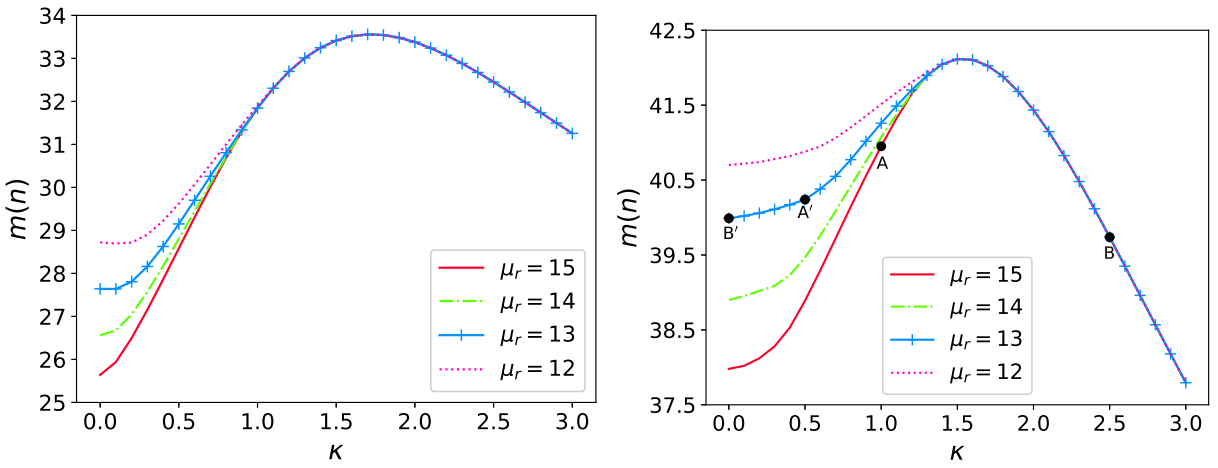


Figure 3.6: Optimal number of permanent positions (top panel) and the corresponding cost (bottom panel) as a function of demand rate uncertainty. The parameters not given in the plots are $c_g = 3.0$, $c_w = 3.0$, $c_o = 1.2$, $r_o = 0.1$, $n = 0.0$, $\tau = 1.0$, and $\xi = 10.0$. For cost evaluations, Q_t is assumed to follow a LogNormal distribution with $\mu_r = 15.0$ and $\kappa_r = 0.3$.

accurate information on demand rate) so as to avoid over-staffing. The plots also show that the value of κ threshold increases with c_g, but is insensitive to c_w and τ. Similarly, the plots in the bottom panels of Figure 3.6 suggest that there exists a threshold for κ, above (below) which the optimal cost function, $m(n)$, shows a decreasing (increasing) trend as Λ_t increases in the convex order. This also contradicts the impact of service time variability as proved in Corollary 6. We further observe in Figure 3.6 that the κ threshold for $m(n)$ (above which the decreasing trend occurs) increases with c_g, decreases slightly with c_w and τ, and is significantly larger than the κ threshold for $a^*(n)$.

The implication of these results is that, when the conditions of Proposition 9 are met,

the savings obtained from the reduction in demand rate uncertainty may be greater than the increase due to fewer applications, thus making a delay in advertising beneficial. When the conditions of Proposition 9 are not met, the situation is more intricate because a reduction in demand rate uncertainty may in fact increase cost, especially if this uncertainty is already high and the cost rate of temporary workers is small relative to the cost rate of waiting. To gain further insight into this situation, we plot the optimal first-stage cost as a function of κ for different levels of μ_r in Figure 3.7, assuming an $M/G/1$ queue. In panel (b) of this figure, c_w and τ are deliberately set to large values to highlight the decreasing trend of cost.



(a) $c_w = 3.0, \tau = 1.0$ **(b)** $c_w = 8.0, \tau = 2.0$

Figure 3.7: Optimal cost as a function of κ for different values of μ_r. The other parameters are $\xi = 10.0$, $c_g = 2.0$, $c_o = 1.2$, $r_o = 0.1$, $n = 0$, and $\kappa_r = 0.3$.

Figure 3.7 shows that for large values of κ, the impact on cost of μ_r becomes negligible. This is because, when κ is large, the optimal number of permanent positions will be small (as illustrated in Figure 3.6), which implies that the number of qualified applications will be less relevant. We further observe in Figure 3.7 that a delayed advertisement is more likely

to be beneficial when the current κ falls on the increasing side of the cost curve than when it is on the decreasing side. For example, assume $\mu_r = 15.0$ for the current advertisement epoch. Panel (b) in Figure 3.7 shows that, if $\kappa = 1.0$, a delayed advertisement leading to a 13% reduction in the mean number of applications and a 50% reduction in demand rate uncertainty would be beneficial (see points A and A' corresponding to the current and delayed advertisement). Yet, if $\kappa = 2.5$, a delayed advertisement with a 13% reduction in mean application numbers would lead to a higher cost, even if the demand rate uncertainty became zero (see points B and B'). In fact, when the current κ falls on the decreasing side of the cost curve, advertising earlier, if feasible, is more likely to be beneficial than later.

The above observations highlight that delaying advertisement beyond t is less likely to be beneficial when demand rate uncertainty is already high and the cost of temporary recruitment is small relative to the cost of patients waiting. It also implies that if the optimal cost with reduced application numbers and no demand rate uncertainty, i.e., $m(n, Q_{t'}, \xi)$, is larger than the current cost, i.e., $m(n, Q_t, \Delta_t)$, there is no benefit from delaying advertisement (Proposition 6 helps in evaluating $m(n, Q_{t'}, \xi)$ by finding the corresponding optimal first-stage decision.) Otherwise, a more detailed investigation is needed as illustrated in the following case study.

3.6 Case Study

We consider the geriatric department of an NHS hospital. The department has a total of $B = 80$ beds and faces significant uncertainty in its winter demand. As an illustration, Figure 3.8 depicts the empirical coefficient of variation (CV) as well as the theoretical CV under the Poisson assumption for daily arrivals using the department's admission data during December and January over the three-year period 2015-2018. The plots indicate a larger variability than expected for a standard Poisson process, hence justifying our use of Poisson mixture models. The department needs to decide how many permanent nursing vacancies to create and advertise for the winter period. Advertising for permanent nurses typically occurs around May/June. Our aim in this section is to illustrate how the framework developed in our study can be applied to guide decision making for nurse recruitment in the department.

We assume that patients arrive to the department according to a Poisson process with a rate whose value is unknown to the decision maker during the permanent recruitment period. This is similar to the assumption made in Hu et al. (2021b) for arrivals to the emergency department. Upon arrival, a patient is admitted to the ward if a bed and a nurse are available. If all beds are taken, the patient joins a queue for beds. If a bed is available but all nurses are busy, an admission request joins a queue for nurses, delaying the admission until a nurse becomes available. The delay in admission to an inpatient department due to unavailability of a nurse is an important factor contributing to the

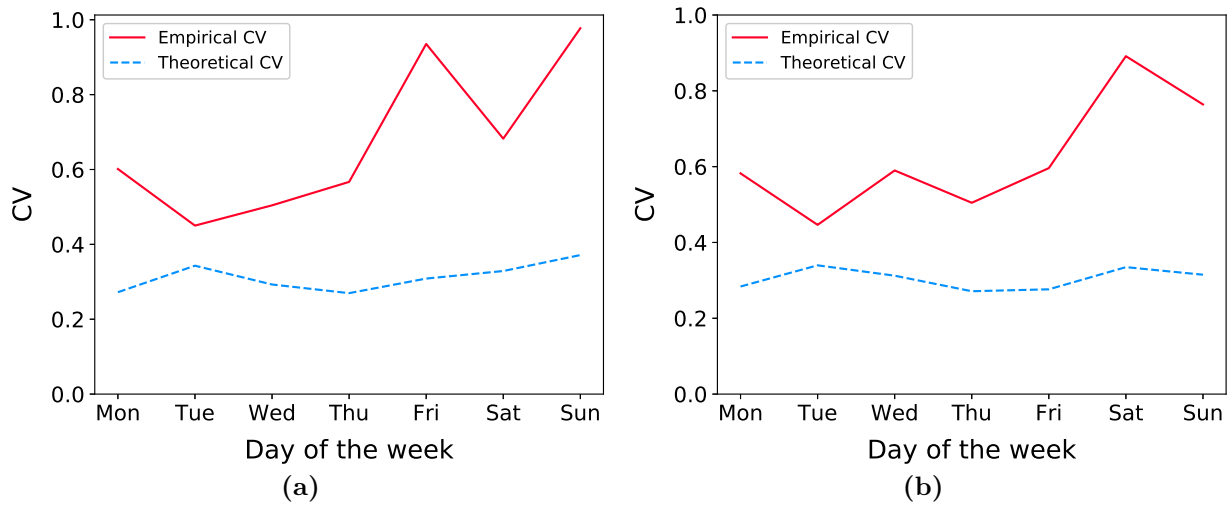


Figure 3.8: Empirical CV and theoretical CV under Poisson assumption for daily admissions in (a) December and (b) January.

so-called "trolley wait" in emergency departments (Abo-Hamad and Arisha, 2013). Whilst in beds, patients generate regular requests for nurses until the end of their length of stay, at which point a discharge request is submitted. The nurse requests are served in the order of regular, discharge, and new admissions by the nursing team. At the end of the discharge process, the patient is discharged from the ward and the bed is cleaned and prepared for the next patient. This workflow implies a nursing queueing system working in conjunction with a bed queueing system.

In §3.6.1, we show how a simulation model capturing the interactions between the bed and nursing queueing systems can be embedded in our two-stage framework to guide recruitment decision making. We refer to this model as the multi-resource multi-server (MRMS) model. In section §3.6.2, we show how a single-resource single-server (SRSS) and a single-resource multi-server (SRMS) approximation, developed based on our analytical

results, could speed up the calculations, and compare their accuracies to the MRMS model. In §3.6.3, we use the SRMS approximation to shed some light on the benefit/loss of delaying advertisement by investigating the trade-off of a more accurate demand information versus the higher risk of not filling permanent positions.

3.6.1 The MRMS Model

We first develop a detailed discrete-event simulation model involving all dynamics of bed and nursing queueing systems. Following Yankovic and Green (2011), the model considers two types of resources, beds and nurses, each of which has its own separate queue. We use the superscript (b) and (n) to represent the association of a parameter to the bed and nursing queueing system, respectively. Let $\lambda^{(b)}$ be the rate of patient arrival during winter, and denote by $\Lambda_t^{(b)}$ the corresponding random variable as predicted at time t when advertisement occurs. As Yankovic and Green (2011), we assume: (i) lengths of stay in the department are i. i. d. as an Exponential distribution with mean $1/\mu^{(b)}$; (ii) each patient generates regular requests, independently of other patients, during her stay according to a Poisson process with a known rate $\lambda^{(n)}$; and (iii) admission, regular, and discharge processing times as well as cleaning times are i. i. d. with known distributions. Given patient arrival rate, $\lambda^{(b)}$, and number of nurses, s, the simulation estimates the mean number of requests in the nursing system, $l^{(n)}(\lambda^{(b)}, s)$.

Next, we adapt the two-stage framework by modifying the first- and second-stage

formulations as

$$m_{sim}(n) = \min_{a} \left\{ \mathbb{E}\left[v_{sim}\left(\Lambda_t^{(b)}, n + \min\{Q_t, a\} \right) \right] : a = 0, \cdots, a_{max} \right\}, \tag{3.41}$$

and

$$v_{sim}(\lambda, p) = \min_{g} \Big\{ p(1 + r_o c_o) + g c_g + l^{(n)}\left(\lambda^{(b)}, p(1 + r_o) + g \right) c_w : \\ g = \lceil \lambda - p(1 + r_o) \rceil, \cdots, g_{max} \Big\}, \tag{3.42}$$

respectively, where $\lceil x \rceil$ is the ceiling function of x, and a_{max} and g_{max} are the respective upper bounds for a and g. The optimal solution to (3.41) is denoted by $a^*_{sim}(n)$ and is obtained by complete enumeration.

The parameters of the model are estimated as follows. For $\Lambda_t^{(b)}$, we test the null hypothesis of a Gamma distribution with shape and scale parameters η and ν, respectively, as per Jongbloed and Koole (2001). This hypothesis implies a Negative Binomial distribution for arrival counts with η experiments and success probability $1/(1 + \nu)$. Using the daily arrival counts of December over the three year period (i.e., 93 observations), we estimate $\hat{\eta} = 2.92$ and $\hat{\nu} = 3.52$ via maximum likelihood. Applying a Kolmogorov-Smirnov goodness-of-fit test and bootstrapping (Jongbloed and Koole, 2001), a p-value of 0.395 is obtained, indicating that the Gamma-distribution hypothesis for arrival rate cannot be rejected. As such, we assume $\Lambda_t^{(b)}$ follows a Gamma distribution with mean $\xi^{(b)} = \hat{\eta}\hat{\nu} = 10.3$ patients

per day and CV $\kappa^{(b)} = \dfrac{1}{\sqrt{\hat{\eta}}} = 0.58$. Based on our findings from §3.4.2, the moderate value obtained for $\kappa^{(b)}$ indicates that there is value in incorporating the demand rate distribution into the two-stage decision making process.

Our data gives a mean length of stay of $1/\mu^{(b)} = 6.48$ days for geriatric patients. This implies a traffic intensity of $\xi^{(b)}/(B\mu^{(b)}) \times 100 = 83.4\%$ for the bed queueing system. The processing times for regular requests are assumed to follow an exponential distribution with rate $\mu^{(n)} = 4$ per hour, based on estimates provided in Lundgren and Segesten (2001) and Dochterman and Bulechek (2004). Following Yankovic and Green (2011), we assume that the admission and discharge processing times are uniformly distributed over intervals $[12, 60]$ and $[10, 60]$ min, respectively, and the time to clean a room after the discharge of a patient is 30 min. These timings were confirmed by the ward's nursing team.

For $\lambda^{(n)}$, Lundgren and Segesten (2001) suggest 0.38 requests per hour, but we consider $\lambda^{(n)} \in \{0.4, 0.5\}$ to cover situations with older and relatively more demanding patients. Following Pinker and Tilson (2013), we assume that Q_t follows a Poisson distribution with mean μ_r. According to the hospital's human resource department, a maximum of 20 qualified applications is likely to arrive over a six-month recruitment period starting from May/June. As such, we consider $\mu_r \in \{10.0, 12.0\}$ so that the probability of receiving more than 20 applications is small. For the remaining parameters, we consider $\kappa^{(b)} \in \{0.58, 1.0\}$, $c_g \in \{2, 3\}$, $c_w \in \{1.5, 3.0\}$, $c_o \in \{1.5, 1.7\}$, $n \in \{0, 1\}$, and $r_o \in \{0.05, 0.1\}$. The values considered for $\kappa^{(b)}$ capture the current level of uncertainty in patient arrival data as well as a situation with a more uncertain arrival rate. The values for r_o and c_g are consistent with

the estimates provided in Lu and Lu (2017), and the values of c_w follow Hu et al. (2021b). The values for c_o capture the current overtime payment in the NHS as well as payments in more expensive private providers. The combinations of these parameters result in 256 scenarios.

For each of the 256 scenarios, we obtain $a^*_{sim}(n)$ and $m_{sim}(n)$ via complete enumeration with $a_{max} = g_{max} = 21$, and $l^{(n)}(\lambda^{(b)}, s)$ estimated by running 50 replications of the simulation model each over 30 days. The values of a_{max} and g_{max} are set based on the maximum value that Q_t may take plus n. The computations are carried out in parallel on a high performance computing system, taking around 3 hours to complete for each scenario. As an example, Figure 3.9 illustrates the first-stage cost as a function of a for two specific scenarios. The plot in the left panel of this figure implies that $a^*_{sim}(n) = 4$, and that underestimating the optimal a may not increase cost substantially, while overestimating it may increase cost by as much as 61.22%. By contrast, the plot in the right panel implies that $a^*_{sim}(n) = 9$, and that overestimating the optimal a may not significantly increase cost, while underestimating it may increase cost by as much as 40.66%. Overall, the results indicate that $a^*_{sim}(n)$ varies between 3 and 9 in the scenarios we considered, and that the difference between optimal and highest first-stage costs (over the range considered for a) exceeds 30.0% in 163 scenarios, and reaches a maximum of 67.0%. These observations highlight the importance of finding the optimal first-stage decision. Note that for larger values of a, the impact of Q on the cost function diminishes, thus the graph flattens out.

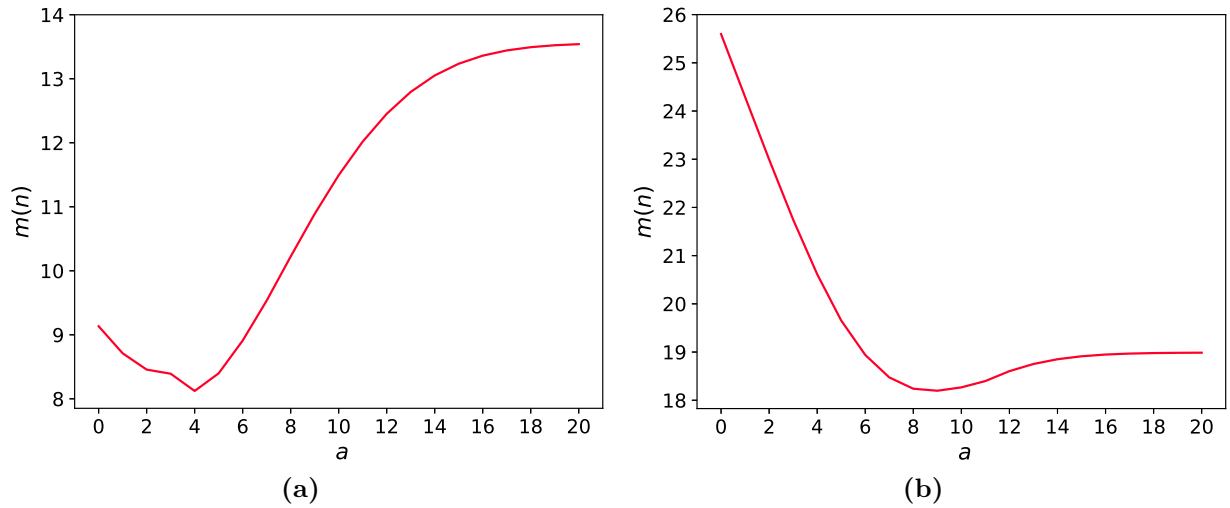


Figure 3.9: First-stage cost as a function of a for the scenario with (a) $c_g = 2.0, c_w = 1.5, c_o = 1.5, n = 1.0, r_o = 0.1, \mu_r = 12.0, \lambda^{(n)} = 0.4$, and $\kappa^{(b)} = 1.0$, and (b) $c_g = 3.0, c_w = 3.0, c_o = 1.5, n = 0.0, r_o = 0.05, \mu_r = 10, \lambda^{(n)} = 0.5$, and $\kappa^{(b)} = 0.58$.

3.6.2 The SRMS and SRSS Approximations

The MRMS model is complex to code and time-consuming to run. To speed up the coding and calculations, we propose SRSS and SRMS approximations by assuming that the dynamics of service delivery in the department are represented by an $M/M/1$ queue and an $M/M/s$ queue, respectively. Focusing on the nursing queueing system, these approximations do not capture the dynamics of the bed system explicitly. In addition, the SRSS approximation estimates the performance of the multi-server nursing queueing system by an inflated single-server queue.

For both approximations, we estimate the demand rate as

$$\lambda = (\lambda^{(n)} + 2\mu^{(b)}) \left(\lambda^{(b)}/\mu^{(b)}\right),$$

where the first term is the overall mean number of requests generated by a single patient per unit of time and the second term is the the average number of patients in the bed system. From this, we obtain $\Lambda_t = (\lambda^{(n)} + 2\mu^{(b)}) \left(\Lambda_t^{(b)}/\mu^{(b)}\right)$, hence, Λ_t follows a Gamma distribution with mean $\xi = (\lambda^{(n)} + 2\mu^{(b)})\xi^{(b)}/\mu^{(b)}$ and CV $\kappa = \kappa^{(b)}$. As an illustration, note that with $\xi^{(b)} = 10.3$ patients per day, $\lambda^{(n)} = 0.5$ requests per hour, and $\mu^{(b)} = 1/6.48$ patients per day, we obtain an average arrival rate of 821.528 requests per day, or equivalently an average offered load of 8.55 (recall that $\mu^{(n)} = 4$ per hour), which is relatively small. For example, the average offered load observed in the emergency department considered in Hu et al. (2021b) exceeds 59.0. This highlights the importance of using an exact approach instead of large-scale asymptotic approximations for inpatient settings.

We use Algorithm 2 to determine $a^*(n)$ for all the 256 scenarios of §3.6.1 with both SRSS and SRMS approximations. We then run the simulation model developed in §3.6.1 with $\lceil a^*(n) \rceil$ to obtain the corresponding cost. Our results indicate that $\lceil a^*(n) \rceil$ obtained from the SRSS approximation is equal to $a^*_{sim}(n)$ in 108 out of 256 scenarios. This figure increases slightly to 112 for the SRMS approximation. The average percentage difference in cost for the SRSS and SRMS approximations, when compared to the MRMS model, are relatively close at about 0.97%. We repeat the same set of experiments with $\xi^{(b)} = 8.64$ patients per day, which yields a traffic intensity of 70.0% for the bed queueing system, to assess the SRMS and SRSS approximations in a less congested department. In this experiment, we observe an average percentage cost difference of 1.33% for the SRSS approximation, and 0.80% for the SRMS approximation, when compared to the MRMS model. The reduction

in the accuracy of the SRSS approximation is because with a lower traffic intensity, it is more likely that some servers become idle. The improvement in accuracy of the SRMS approximation, on the other hand, is because with a lower average patient arrival rate, the impact of the bed constraint on the nursing queueing system diminishes. Overall, both SRSS and SRMS approximations are reasonably accurate under different load conditions.

3.6.3 Delaying Advertisement

As discussed in §3.6.3, delaying the advertisement may reduce the variability in demand rate at the expense of a reduction in the number of qualified applications. We also observed that delaying advertisement is less likely to be beneficial when demand rate uncertainty is already high ($\kappa > 1.0$) and the cost of temporary recruitment is small relative to the cost of patients waiting. To further investigate this, we numerically evaluate the amount of reduction needed in demand rate uncertainty to make the cost of a later advertisement equal to the current cost as a function of the reduction in the mean number of qualified applications. Given the accuracy of the SRMS approximation illustrated in §3.6.2, it is used in the analysis that follows.

We consider the scenario with $\lambda^{(n)} = 0.4, \mu_r = 10.0, c_g = 2.0, c_w = 3.0, c_o = 1.5, n = 0.0$, and $r_o = 0.05$, as the benchmark scenario and evaluate its cost using the SRMS approximation. We then reduce μ_r in steps of 5.0%, and evaluate the minimum reduction in κ that makes the system cost equal to the cost of the benchmark scenario for the resulting

μ_r value. The calculation stops when, for a given percentage reduction in μ_r, the cost with zero demand rate uncertainty falls above the benchmark cost.

The results are presented in Figure 3.10 for different levels of temporary cost rate, c_g, and different levels of current demand rate uncertainty, κ.

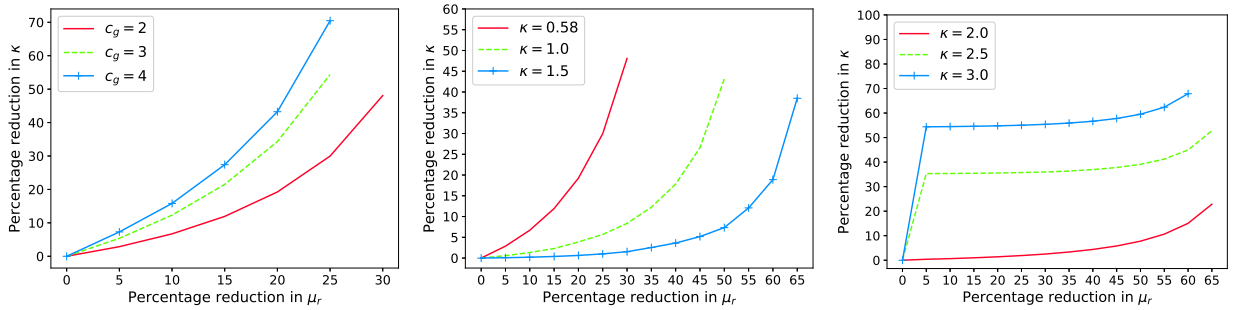


Figure 3.10: The reduction required in demand rate uncertainty as a function of reduction in mean application numbers. For panel (a), $\kappa = 0.58$ and for panels (b) and (c), $c_g = 2.0$.

The plots in panel (a) of Figure 3.10 show that a larger reduction in demand rate uncertainty is needed to make a later advertisement beneficial as c_g grows. They also imply that reductions above 30% (25%) in mean application numbers for $c_g = 2.0$ ($c_g = 3.0$ and $c_g = 4.0$) cannot be compensated even if we knew the demand rate. The plots in panel (b) show that, when the current demand rate uncertainty is less than or equal to 1.5, the reduction required in demand rate uncertainty typically reduces with κ. This corresponds to κ falling on the increasing side of the cost curve. In particular, for $\kappa = 1.5$, the required reduction in demand rate uncertainty is relatively small, even when the mean application number halves. Panel (c) highlights an opposite behaviour when κ is larger than or equal to 2.0, corresponding to κ falling on the decreasing side of the cost curve. Panels (b) and (c)

imply that the maximum reduction in mean application numbers that can be compensated by a reduction in demand rate uncertainty increases (decreases) with κ, when κ is less than or equal to (larger than or equal to) 1.5 (2.0). Overall, Figure 3.10 provides valuable insights on how and when delaying advertisement may create value to the provider.

3.7 Extension to a Multiple-Segment HUDP

We extend our framework by dividing the HUDP into N equally spaced segments indexed by j, and assume there is an opportunity for recruiting temporary HCWs at the beginning of each segment (see Figure 3.11). We consider the possibility of correlated demand in different segments of the HUDP. For example, a low demand segment may follows a high demand segment. There remains a single opportunity for advertising permanent positions at time t, following which applications arrive over the period (t, t_e). A decision is then made at the beginning of each segment j as to how many temporary HCWS to recruit given the exact demand rate of that segment and the number of permanent HCWs recruited. The temporary HCWs recruited for each segment are released at the end of the segment, while permanent HCWs remain in the system until the end of the planning horizon. We assume the system achieves a steady-state during each segment of the HUDP.

Let $\mathbf{\Lambda}_t = (\Lambda_t^1, \ldots, \Lambda_t^N)$ be the vector of random variables representing demand rates of different segments as predicted at time t, and denote by $h_t(\lambda^1, \ldots, \lambda^N)$ the corresponding joint pdf. Let Q_t, n, a, cost parameters, and $v(\lambda, p)$ be as defined in §3.2. The first stage

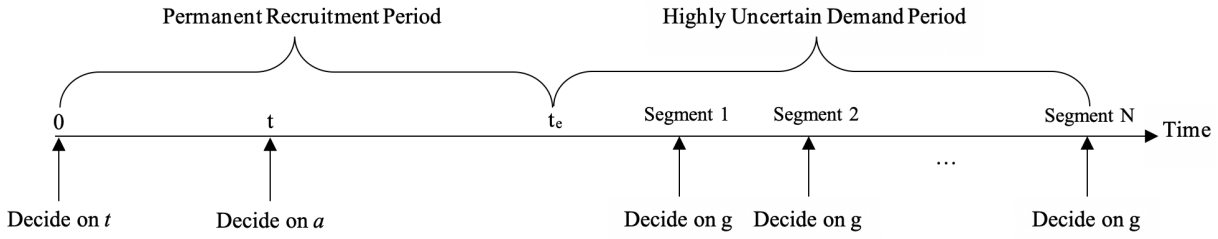


Figure 3.11: Schematic diagram of permanent and temporary recruitment decision making for the multiple-segment HUDP.

problem is then formulated as

$$m(n) = \min_{a} \left\{ \mathbb{E}\left[\sum_{j=1}^{N} v\left(\Lambda_t^j, n + \min\{Q_t, a\}\right) \right] : a \in \mathbb{R}^+ \right\}. \tag{3.43}$$

To obtain the optimal solution to problem (3.43), we revise the definition of function $\psi_n(a)$ as

$$\psi_n^{h_t^j}(a) = 1 + r_o c_o + c_w(1+r_o) \int_0^{\tilde{\lambda}(n+a)} \frac{\partial l(\lambda, s)}{\partial s}\Big|_{s=(n+a)(1+r_o)} h_t^j(\lambda^j) d\lambda \\ - c_g(1+r_o)\left(1 - H_t^j\left(\tilde{\lambda}(n+a)\right)\right), \tag{3.44}$$

where

$$h_t^j(\lambda^j) \triangleq \int_0^\infty \dots \int_0^\infty h_t(\lambda^1, \dots, \lambda^N) d\lambda^1 \dots d\lambda^{j-1} d\lambda^{j+1} \dots d\lambda^N, \tag{3.45}$$

is the marginal pdf of Λ^j, and $H_t^j(\cdot)$ is the corresponding CDF. We now propose

Proposition 10. *For the first-stage problem given in* (3.43)*,*

$$a^*(n) = \begin{cases} 0.0 & \text{if } \sum_{j=1}^{N} \psi_n^{h_t^j}(0) \geq 0, \\ \min\{\tilde{a}(n), q_u\}, & \text{otherwise,} \end{cases} \tag{3.46}$$

where $\tilde{a}(n)$ *is the unique root of function* $\sum_{j=1}^{N} \psi_n^{h_t^j}(a)$ *in the interval* $(0, \infty)$*.*

Proof. Denoting the expected value in Equation (3.43) by $y(n, a)$, and conditioning on Q_t, we obtain

$$y(n,a) = \int_0^a \mathbb{E}\left[\sum_{j=1}^{N} v\left(\Lambda_t^j, n+q\right)\right] f_t(q)dq + \mathbb{E}\left[\sum_{j=1}^{N} v\left(\Lambda_t^j, n+a\right)\right](1 - F_t(a)).$$

Taking the derivative of $y(n, a)$ with respect to a, and simplifying, we arrive at

$$\frac{\partial y(n,a)}{\partial a} = \frac{\partial \mathbb{E}\left[\sum_{j=1}^{N} v(\Lambda_t^j, n+a)\right]}{\partial a}(1 - F_t(a)) \tag{3.47}$$

$$= \left(\sum_{j=1}^{N} \frac{\partial \mathbb{E}\left[v(\Lambda_t^j, n+a)\right]}{\partial a}\right)(1 - F_t(a)) = \left(\sum_{j=1}^{N} \psi_n^{h_t^j}(a)\right)(1 - F_t(a)), \tag{3.48}$$

where the last equality is by Equation (3.17) and the definition of $\psi_n^{h_t^j}(a)$ given in (3.44). The rest of the proof follows the same logic as that of Proposition 2. □

We now investigate the impact of demand correlation between different segments of the HUDP on $a^*(n)$. According to Proposition 10, optimal a is obtained by finding the root of equation $\sum_{j=1}^{N} \psi_n^{h_t^j}(a) = 0$, where $\psi_n^{h_t^j}$ depends only on the marginal pdf $h_t^j(\cdot)$. If the

multivariate distribution is closed under marginalization, the correlation coefficient does not appear in $h_t^j(\lambda^j)$, and so it does not impact $a^*(n)$. Otherwise, the correlation coefficient appears in the marginal distributions and $a^*(n)$ changes with the correlation.

As an illustration, we set $N = 2$ and assume that h_t follows a bivariate Normal distribution with mean $\boldsymbol{M}$ and covariance matrix $\boldsymbol{\Sigma} = [\rho_{(j,k)}\sigma_j\sigma_k]_{2\times 2}$, where $\rho_{(j,k)}$ is the correlation between the demand of segments j and k, and σ_j and σ_k are the corresponding standard deviations. In one set of experiments, we assume an $M/M/1$ queue with $\boldsymbol{M} = (8, 8)$, $(\sigma_j, \sigma_k) \in \{(1, 1), (2, 2), (3, 3)\}$, and find $a^*(n)$ as a function of ρ. As illustrated in panel (a) of Figure 3.12, $a^*(n)$ is independent of ρ for this set of experiment. In the other set of experiments, we set $\boldsymbol{M} = (2, 2)$ and use the same values of (σ_j, σ_k), but truncate the bivariate normal distribution to remove the possibility of arrival rate being negative. The results illustrated in panel (b) of Figure 3.12, shows that $a^*(n)$ does change with ρ. This is because the marginal distribution of a bivariate truncated normal is not a truncated normal, and is in fact a function of the correlation (Horrace, 2005).

3.8 Summary

Given the long lead-time in recruiting permanent workers and the higher cost of temporary skilled workers, it is essential for healthcare providers to know how many permanent positions they need to advertise well before a period of highly uncertain demand starts. By representing the service delivery in such periods as a generic delay queueing model, we

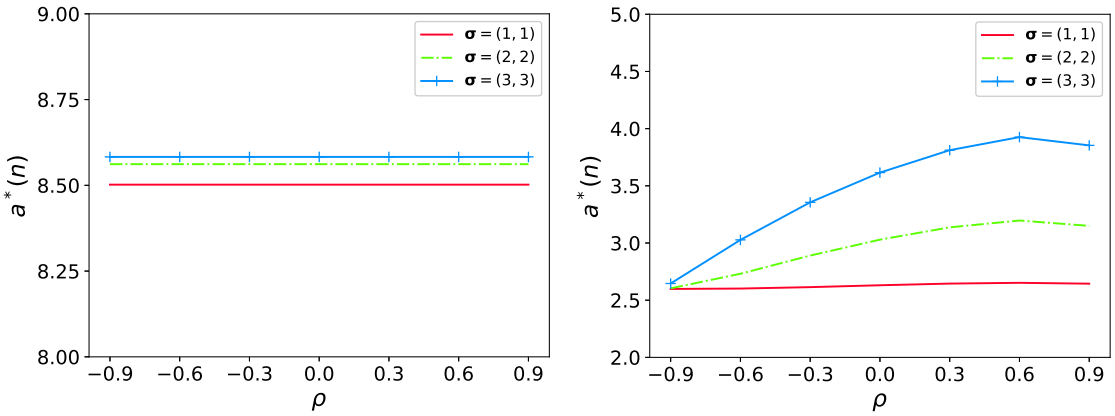


Figure 3.12: $a^*(n)$ as a function of correlation coefficient for (a) h_t following a bivariate normal distribution with $\boldsymbol{M} = (8, 8)$, and (b) h_t following a bivariate truncated normal distribution with $\boldsymbol{M} = (2, 2)$. We set $c_g = 2$, $c_w = 0.5$, $c_o = 1.2$, $n = 0$, and $r_o = 0.2$.

proposed a two-stage stochastic optimization framework to inform recruitment decision making. The first stage focuses on permanent recruitment and the second stage on temporary recruitment.

We analytically characterized the optimal first and second-stage recruitment decisions and proposed fast numerical algorithms for specifying their values. We proved that the optimal first-stage decision is insensitive to the probability distribution for the number of qualified applications as long as the distribution support remains the same. Using stochastic ordering, we also proved that the optimal mean first-stage cost typically decreases when more applications are likely to be received, and set out the conditions under which the optimal first-stage decision and the corresponding cost increase when the demand rate becomes more uncertain. All results were exact and obtained without specific assumptions on the type or scale of the delay queueing model, and remain valid as long as the corresponding

system-size function follows several properties. These properties are intuitive and we proved that they hold for three common queueing models, $M/M/1$, $M/G/1$, and $M/M/s$. We noted that $M/M/1$ and $M/G/1$ models are approximations but are useful as they provide further analytical tractability. In particular, we obtained a closed-form expression for the second-stage decision and the corresponding cost for the $M/M/1$ case, and proved that the optimal first- and second-stage decisions and their corresponding costs increase with service time variability for the $M/G/1$ case.

By combining analytical results with numerical experiments, we derived several managerial insights as follows:

- Except in situations where the cost of temporary recruitment is extremely high, there is value in recruiting temporary staff even if demand rate uncertainty is very low. The amount of this value is likely to increase with the system scale and decrease with the waiting cost.

- When demand rate uncertainty is moderate to high, there is value in obtaining the demand rate distribution and incorporating it into recruitment decision making. Otherwise, a two-stage approach using only the average demand rate would suffice.

- As the uncertainty in demand rate increases up to a certain threshold, the optimal number of permanent positions increases. Above this threshold, however, the optimal number of permanent positions exhibits a decreasing trend. Similarly, the optimal system cost decreases when demand rate uncertainty surpasses a threshold. This

threshold increases with the cost of temporary workers and decreases with the cost of patients waiting. The main implication is that delaying advertisement is less likely to be beneficial when demand rate uncertainty is already high and the cost of temporary recruitment is small relative to the cost of patients waiting.

We conducted a case study using data from a geriatric ward in the UK and demonstrated how our framework can guide nurse recruitment decision making in a complex environment. In particular, we assumed that the system-size function is estimated by a detailed simulation model which captures the complexities of nursing care in inpatient wards, including the wide range of requests from patients and the availability of beds as the second type of resource (in addition to nurses). We also showed that simple single-resource approximation models based on our analytical results are reliable and sufficiently accurate for the permanent recruitment decision. We further illustrated how our models can be applied to evaluate the reduction in demand rate uncertainty that makes a delayed advertisement beneficial as a function of the reduction in mean application numbers.

We further extended our modelling framework to allow multiple opportunities for recruitment of temporary staff during the service delivery. The analytical characterization of the optimal permanent recruitment decision was derived. We showed that the optimal permanent recruitment decision is independent of the demand rate correlation if the (multivariate) demand rate distribution is closed under marginalization.

Chapter 4

A Long-Term Recruitment Model

4.1 Introduction

In this chapter, we aim to capture the difference in placement durations of permanent and temporary HCWs, in addition to the differences in their staffing costs and recruitment lead times. This is because permanent HCWs have substantially longer contracts with the provider than their temporary counterparts. This implies that the provider must consider their longer term cost and benefits when making recruitment decisions. This is particularly important when there are periods of highly uncertain demand as permanent HCWs recruited for one period may not be needed in the next. To simplify the analysis, we ignore the uncertainty in permanent recruitment in this chapter and assume that the desired number of permanent HCWs can always be recruited.

We consider a multi-interval permanent and temporary recruitment problem in a setting where all patient requests must be served. Each interval of the planning horizon is divided into a permanent recruitment period and a HUDP. A two-stage decision making process is repeated for each interval: the number of permanent HCWs are decided at the beginning of the permanent recruitment period (the first-stage decision), and the number of temporary HCWs are decided at the beginning of the HUDP (the second-stage decision). The first-stage decision is made when only partial information about demand in the corresponding HUDP is available, whereas the second-stage decision is made when more accurate demand information becomes available. We assume that the permanent HCWs recruited in each interval stay in the system and continue to provide services in the subsequent intervals until they are dismissed at the end of the planning horizon at a given cost. Temporary workers, however, are contracted for the interval and thus leave the system at the end of each interval.

As in Chapter 3, the dynamics of service delivery during the HUDP are captured by a generic delay queueing model which evaluates the expected system size, i.e., the mean number of requests waiting or being served, in steady state. Similar to Chapter 3, we model demand as a Poisson mixture process, with the distribution of the rate being available at the time of the first-stage decision, and the exact value of the rate when the second-stage decision is made. To capture the variation of demand over time, we assume that the distribution of demand rate evolves according to a discrete-time Markov chain (MC) with a finite state space in consecutive intervals. For instance, we can have two states for the

demand rate distribution, *low* and *high*, such that the parameters (mean and/or variance) of the high distribution are larger than the parameters of the low distribution. This enables us to formulate the problem as an MDP, where the system state at the beginning of each time interval comprises the number of permanent employees already in the system and the state of the demand rate distribution. The objective is to minimize the total expected cost, including staffing costs as well as the cost incurred by patients while their requests are in the system.

The remainder of this chapter is organized as follow. We start with the problem definition in §4.2. This includes the formulation of the problem and the analytical characterization of the optimal permanent recruitment policy. We then conduct numerical experiments using illustrative data in §4.3. This includes investigating the sensitivity of the optimal policy to system parameters in §4.3.1, and evaluating the potential savings obtained from the recruitment policy suggested by our multi-interval model as compared to a myopic policy suggested by a single-interval model (similar to the model developed in Chapter 3) in §4.3.2. A summary of the findings is presented in §4.4.

4.2 Problem Definition

Consider a planning horizon consisting of T intervals of equal length, which is typically a year, but may vary depending on the application. We index time intervals by $t = 1, 2, \ldots, T$, and denote the end of the planning horizon by $T+1$. As depicted in Figure 4.1, each interval

is divided into a permanent recruitment period, during which recruitment of permanent HCWs takes place, and a HUDP, during which patients' requests are served by HCWs. Advertisement for permanent HCW positions starts at the beginning of the permanent recruitment period in each interval, and recruitment of temporary HCWs occurs at the beginning of the HUDP in each interval. Patients' requests arrive according to a Poisson process with rate λ_t during the HUDP of interval t. The requests wait in a queue until they are served by a member of the pool of HCWs (permanent or temporary). The duration of service is random, and its average is set as the time unit so that the rate of service delivery is equal to one.

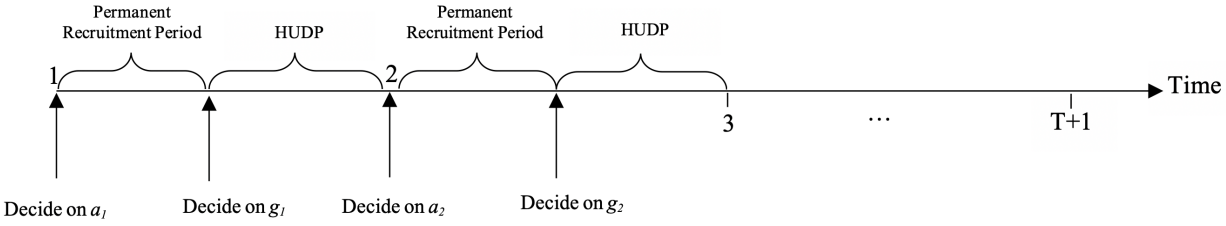


Figure 4.1: Schematic diagram of the long-term recruitment decision making process

The rate of Poisson arrivals is unknown to the service provider at the point of advertising for permanent HCWs. As such, this rate is a random variable, which we denote by Λ_t for $t = 1, \ldots, T$. The distribution of this rate, however, is available to the service provider and evolves in successive intervals according to a MC with state space $\mathcal{S} = \{0, 1, \ldots, k\}$, where each state $i \in \mathcal{S}$ represents a random variable Λ^i with a known pdf $h^i(\cdot)$. We represent the mean and CV of Λ^i by ξ^i and κ^i, respectively. The transition probability matrix of this MC is denoted by $\boldsymbol{Q} = [q_{ij}]_{i,j\in\mathcal{S}}$. The evolution of the demand rate distribution over time

may capture variations in the average demand rate, uncertainty of the demand rate, or a combination of both. The value of the rate becomes known to the service provider at the start of each HUDP.

As shown in Figure 4.1, each interval t involves a two-stage decision making process. The first-stage decision concerns the number of permanent FTE positions to advertise at the beginning of the permanent recruitment period, denoted by $a_t \in \mathbb{R}^+$, and the second-stage decision concerns the number of temporary FTEs to recruit at the beginning of the HUDP, denoted by $g_t \in \mathbb{R}^+$, for $t = 1, \ldots, T$. Once permanent HCWs are recruited in a time interval, they remain in the system and provide services in subsequent intervals until they are dismissed at the end of the planning horizon at a cost c. The temporary HCWs, on the other hand, are contracted for each time interval.

The state of the system at the beginning of interval t is represented by vector (n_t, i), where $n_t \in \mathbb{R}^+$ is the number of permanent HCWs in the system before the permanent recruitment decision is made, and $i \in \mathcal{S}$ is the observed state of the demand rate distribution. Having observed the state of the system at the beginning of interval t, the provider decides on the number of permanent positions, a_t, to advertise. To simplify the analysis, we make the following assumption.

Assumption 2. *The desired number of permanent HCWs can always be recruited.*

Accordingly, there will be a total of p_t permanent HCWs in the system at the beginning of the HUDP of interval t, where $p_t = n_t + a_t$. Given p_t and the revealed value of the

demand rate, λ_t, the provider then decides how many FTEs of temporary HCWs, g_t, to recruit at the beginning of the HUDP of interval t. Let $l(\lambda, s)$, c_p, c_g, c_o, c_w, and r_o be as defined in Chapter 3, and suppose that $l(\lambda, s)$ satisfies the properties given in Assumption 1. Given n_1 permanent HCWs at the beginning of the planning horizon, the multi-interval optimization problem is given by

$$\min_{a_1,a_2,\cdots,a_T} \left\{ \sum_{t=1}^{T} \alpha^t \mathbb{E}\left[v_t(\Lambda_t, n_t + a_t)\right] + \alpha^T c(n_T + a_T) : \right.$$
$$\left. a_t \in \mathbb{R}^+ \text{ and } n_{t+1} = n_t + a_t \text{ for all } t \right\}, \quad (4.1)$$

where $\alpha \in (0, 1]$ denotes the one-interval discount factor and $v_t(\lambda, p)$ is the minimum expected cost rate of the HUDP of interval t given arrival rate λ and p permanent HCWs. Since $v_t(\lambda, p)$ does not vary with time, it follows that

$$v_t(\lambda, p) = v(\lambda, p) =$$
$$\min_g \left\{ p(1 + r_o c_o) + g c_g + l(\lambda, p(1 + r_o) + g) c_w : g \in \mathbb{R}^+, g > \lambda - p(1 + r_o) \right\}, \quad (4.2)$$

for all t, which is the second-stage of the single-interval problem given in (3.1). Hence, given arrival rate λ_t and p_t permanent HCWs, the optimal number of temporary HCWs to be recruited at the beginning of the HUDP of interval t is given by $g_t^*(\lambda_t, p_t) = g^*(\lambda_t, p_t)$, which is obtained via Proposition 1 (or numerically via Algorithm 1).

We reformulate the problem in (4.1) as a discrete-time MDP with state space $\{(n_t, i) \in$

$\mathbb{R}^+ \times \mathcal{S}\}$. The value function of this MDP is represented through the following recursive expression (see Puterman, 1994):

$$V_t(n_t, i) = \min_{a_t} \left\{ \mathbb{E}\left[v(\Lambda^i, n_t + a_t)\right] + \alpha \sum_{j \in \mathcal{S}} q_{i,j} V_{t+1}(n_{t+1}, j) : a_t \in \mathbb{R}^+ \right\}, \tag{4.3}$$

for $t = 1, \ldots, T$, where $n_{t+1} = n_t + a_t$, and

$$V_{T+1}(n, i) = cn, \tag{4.4}$$

for all $n \in \mathbb{R}^+$ and $i \in \mathcal{S}$. Using the structural properties of the MDP formulated in (4.3), we prove in the following proposition that the optimal permanent recruitment policy in each time interval is a state-dependent hire-up-to policy.

Proposition 11. *Given state (n_t, i) at the beginning of interval t, the optimal permanent recruitment policy is given by*

$$a_t^*(n_t, i) = \begin{cases} p_{t,i}^* - n_t & \text{if } n_t \leq p_{t,i}^* \ , \\ 0 & \text{otherwise.} \end{cases} \tag{4.5}$$

To prove Proposition 11, we need the following Lemmas.

Lemma 4.2.1. *If $g(y)$ is convex in y and $f(x)$ is a linear function of x, then $h(x) = g(f(x))$ is also convex in x.*

Lemma 4.2.2. *Let $f(x) = \inf_{y \geq x} \{h(y)\}$. If $h(y)$ is convex in y, then $f(x)$ is convex in x.*

Lemma 4.2.3. *Function $v(\lambda, p)$ given in Equation (4.2) is convex in p.*

For the proofs of Lemmas 4.2.1 and 4.2.2, see the proofs of Lemmas 1 and 2, respectively, in Gans and Zhou (2002, pp.997). The proof of Lemma 4.2.3 is given below.

Proof. The Lagrange function of the optimization model in (4.2) is obtained as

$$\mathcal{L}(\lambda, p, \beta; g) = p(1 + r_o c_o) + g c_g + l(\lambda, p(1 + r_o) + g) c_w - \beta g,$$

where β is the K.K.T multiplier. Note that constraint $g > \lambda - p(1 + r_o)$ is not included in the Lagrange function as it is always active and so its multiplier is equal to zero. Using the Envelope theorem, we then obtain

$$\frac{\partial v(\lambda, p)}{\partial p} = \frac{\partial \mathcal{L}(\lambda, p, \beta; g)}{\partial p}\Big|_{g=g^*(\lambda,p)} = 1 + r_o c_o + c_w(1 + r_o)\frac{\partial l(\lambda, s)}{\partial s}\Big|_{s=p(1+r_o)+g^*(\lambda,p)}, \quad (4.6)$$

which yields

$$\frac{\partial v(\lambda, p)}{\partial p} = \begin{cases} 1 + r_o c_o + c_w(1 + r_o)\dfrac{\partial l(\lambda, s)}{\partial s}\Big|_{s=p(1+r_o)}, & \text{if } \lambda \leq \tilde{\lambda}(p), \\ 1 + r_o c_o + c_w(1 + r_o)\dfrac{\partial l(\lambda, s)}{\partial s}\Big|_{s=p(1+r_o)+\tilde{g}(\lambda,p)}, & \text{if } \lambda > \tilde{\lambda}(p), \end{cases} \quad (4.7)$$

by Proposition 1. Since $\tilde{g}(\lambda, p)$ is the unique root of function $\theta_{\lambda,p}(g)$ given in (3.5), Equation

(4.7) simplifies to

$$\frac{\partial v(\lambda,p)}{\partial p} = \begin{cases} 1 + r_o c_o + c_w(1+r_o)\dfrac{\partial l(\lambda,s)}{\partial s}\Big|_{s=p(1+r_o)} & \text{if } \lambda \leq \tilde{\lambda}(p) \\ 1 + r_o c_o - c_g(1+r_o) & \text{if } \lambda > \tilde{\lambda}(p), \end{cases} \tag{4.8}$$

Taking the derivative of the expression in (4.8) with respect to p, we arrive at

$$\frac{\partial^2 v(\lambda,p)}{\partial p^2} = \begin{cases} c_w(1+r_o)^2\dfrac{\partial^2 l(\lambda,s)}{\partial s^2}\Big|_{s=p(1+r_o)} & \text{if } \lambda \leq \tilde{\lambda}(p) \\ 0 & \text{if } \lambda > \tilde{\lambda}(p), \end{cases}$$

which is non-negative by property A(iv), and so $v(\lambda,p)$ is convex in p. □

We now prove Proposition 11.

Proof. We rewrite the value function given in (4.3) as

$$V_t(n_t,i) = \min_{p_t} \left\{ J_t(p_t,i) : p_t \geq n_t \right\}, \tag{4.9}$$

where

$$J_t(p_t,i) = \mathbb{E}\left[v(\Lambda^i,p_t)\right] + \alpha \sum_{j\in\mathcal{S}} q_{i,j} V_{t+1}(n_{t+1},j), \tag{4.10}$$

and $p_t = n_t + a_t$. We first prove that $J_t(p_t,i)$ is convex in p_t for all $i \in \mathcal{S}$ and $t = 1,\ldots,T$. To show this, we need to prove the convexity of $V_t(n_t,i)$ in n_t for all i and t. We follow proof by induction. It is trivial from Equation (4.4) that $V_{T+1}(n_{T+1})$ is convex in n_{T+1}.

Now, suppose that $V_{t+1}(n_{t+1}, i)$ is convex in n_{t+1} for all i. Then, since $n_{t+1} = n_t + a_t = p_t$, $V_{t+1}(n_{t+1}, i)$ is convex in p_t for all i by Lemma 4.2.1. Since a linear combinations of convex functions is convex, the second term in Equation (4.10) is convex. The first term is also convex because $v(\lambda, p)$ is convex by Lemma 4.2.3, and the integral of a convex function is convex. These imply that $J_t(p_t, i)$ is convex in p_t for all i. Therefore, by Lemma 4.2.2, $V_t(n_t, i)$ given in Equation (4.9) is convex in n_t for all i. Repeating this argument, we conclude that $V_t(n_t, i)$ is convex in n_t, and consequently $J_t(p_t, i)$ is convex in p_t, for all i and t. Next, convexity of $J_t(p_t, i)$ in p_t implies that there exists a point $p^*_{t,i}$ that minimizes $J_t(p_t, i)$ without the constraint $p_t \geq n_t$. Convexity of $J_t(p_t, i)$ further implies that the optimal policy with the constraint is to hire up to $p^*_{t,i}$ when $n_t < p^*_{t,i}$, and not to hire otherwise. □

Proposition 11 implies that the number of existing permanent HCWs, n_t, need not be taken into account when deciding on the optimal hire-up-to value, $p^*_{t,i}$. This reduces the dimension of the search space from three to two, and so speeds up the calculations. Algorithm 3 outlines the steps for finding the optimal hire-up-to value and the corresponding cost given state (n_t, i) for any time $t \in \{1, \ldots, T\}$.

Algorithm 3 Numerical method for evaluating the hire-up-to value $p^*_{t,i}$ for permanent recruitment

Require: $l(x,y)$, $\boldsymbol{Q}$, h^i, α, c, c_g, c_w, c_o, r_o, a_{max}, Δ_a, and $g^*(\lambda,p)$ from Algorithm 1.

function $V(t,n_t,i)$
  if $t = T+1$ **then**
    return cn_t
  else if $p^*_{t,i}$ is not defined **then**
    $cost_{min} \leftarrow \infty$
    for $a_t \in \{0,\Delta_a,\cdots,a_{max}\}$ **do**
      $f = \mathbb{E}\left[v(\Lambda^i,a_t)\right] + \alpha\sum_{j\in\mathcal{S}} q_{i,j}V(t+1,a_t,j)$
      if $cost_{min} < f$ **then**
        $p^*_{t,i} = a_t$
        break
      end if
      $cost_{min} = f$
    end for
  end if
  return $\mathbb{E}\left[v(\Lambda^i,\max(n_t,p^*_{t,i}))\right] + \alpha\sum_{j\in\mathcal{S}} q_{i,j}V(t+1,\max(n_t,p^*_{t,i}),j)$
end function
function $v(\lambda,p)$
  if $g^*(\lambda,p) = 0$ **then**
    return $p(1+r_oc_o) + l(\lambda,p(1+r_o))c_w$
  else
    return $p(1+r_oc_o) + g^*(\lambda,p)c_g + l(\lambda,p(1+r_o)+g^*(\lambda,p))c_w$
  end if
end function

4.3 Numerical Analysis

In this section, we first conduct numerical experiments to investigate the sensitivity of the optimal hire-up-to value with respect to different parameters of the model. We then assess the savings obtained from our multi-interval model when compared to the single-interval model developed in Chapter 3. To simplify the analysis, an inflated $M/M/1$ queue is considered for modelling the dynamics of service delivery during the HUDP. We

consider only two states for the demand rate distribution, i.e., $\mathcal{S} = \{0, 1\}$, where 0 and 1 represent states of low and high demand, respectively. We assume that Λ^i follows a Gamma distribution with given mean ξ^i and CV κ^i. We further assume that there is no permanent employee at the beginning of the planning horizon, i.e., $n_1 = 0$.

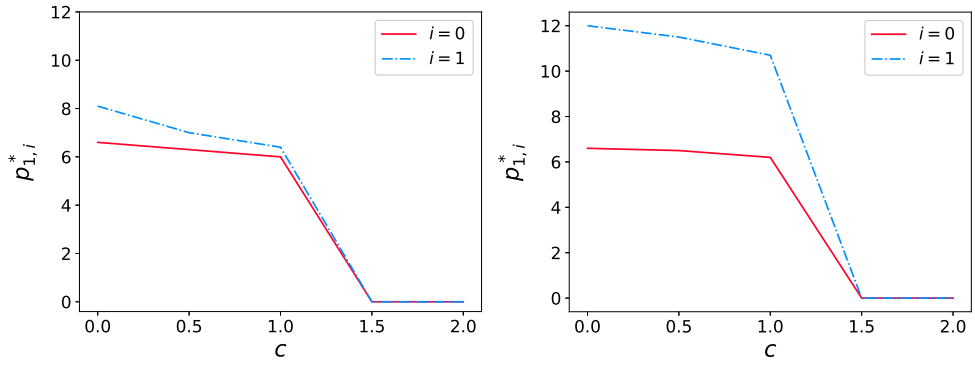
4.3.1 Optimal Policy Illustration

We investigate the impact of c, c_g, c_w, κ^i, and t on the optimal hire-up-to threshold $p^*_{t,i}$. We set $\alpha = 0.8$, $\xi^0 = 5.0$, and $\xi^1 = 10.0$, and assume that there is no mandatory overtime work, i.e., $r_o = 0.0$. For the transition probability matrix, we consider two possibilities, $\boldsymbol{Q}_l$ and $\boldsymbol{Q}_h$, defined as

$$\boldsymbol{Q}_l = \begin{pmatrix} 0.9 & 0.1 \\ 0.9 & 0.1 \end{pmatrix}, \quad \boldsymbol{Q}_h = \begin{pmatrix} 0.1 & 0.9 \\ 0.1 & 0.9 \end{pmatrix},$$

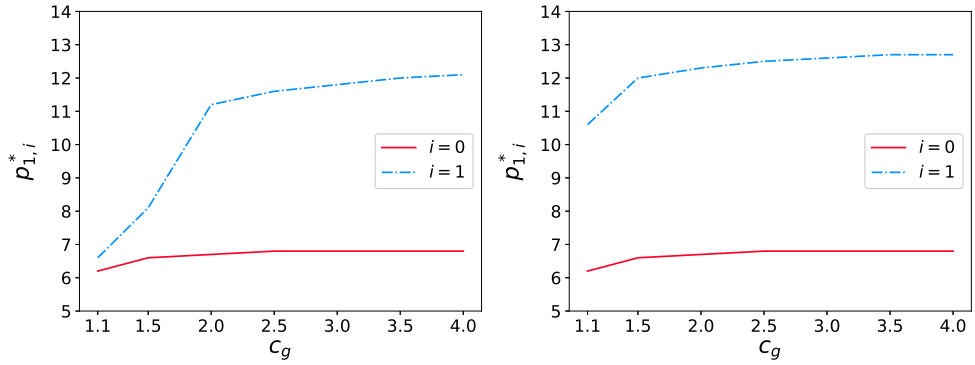
with the former (latter) representing a high probability of ending up in the low (high) demand state.

Figures 4.2, 4.3, and 4.4 show the optimal permanent threshold at the beginning of a two-interval planning horizon ($T = 2$) as function of c, c_g and c_w, respectively, for two initial states of demand ($i = 0, 1$) and two transition probability matrices ($\boldsymbol{Q} \in \{\boldsymbol{Q}_l, \boldsymbol{Q}_h\}$). These figures suggest that $p^*_{1,i}$ decreases with c, and increase with c_g and c_w for all values of i and $\boldsymbol{Q}$. The impact of c_g and c_w is the same as those proved in Corollary 4 for the single-interval model.



(a) Q_l **(b) Q_h**

Figure 4.2: Optimal permanent recruitment threshold at the beginning of a two-interval planning horizon as a function of penalty cost c for $c_g = 1.5$, $c_w = 0.5$, and $\kappa^0 = \kappa^1 = 0.1$.

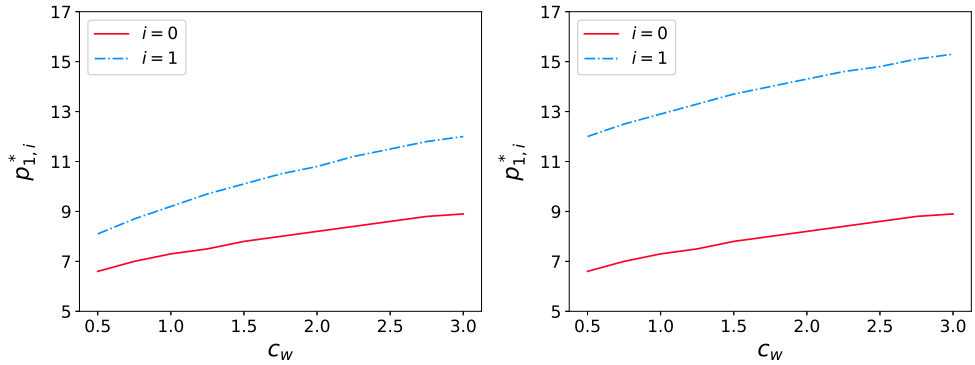


(a) Q_l **(b) Q_h**

Figure 4.3: Optimal permanent recruitment threshold at the beginning of a two-interval planning horizon as a function of temporary cost rate c_g for $c = 0.0$, $c_w = 0.5$, and $\kappa^0 = \kappa^1 = 0.1$.

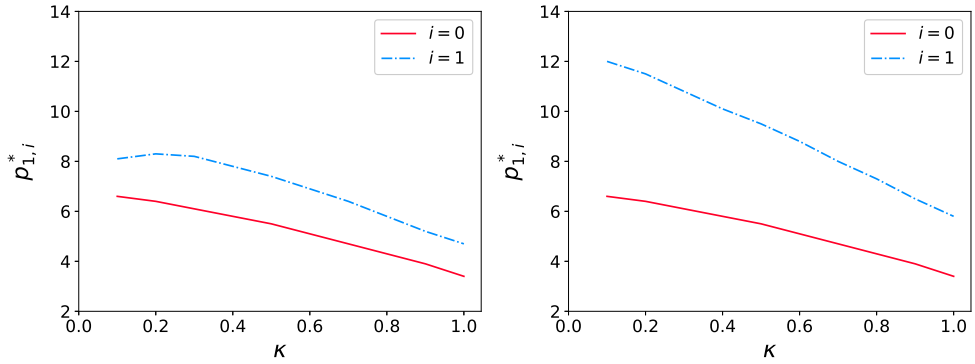
The plots in Figure 4.5 show the impact of $\kappa^0 = \kappa^1 = \kappa$ on the optimal permanent recruitment threshold. They suggest $p^*_{1,i}$ varies in a non-monotone way as κ increases. This is similar to the impact of demand rate uncertainty observed in §3.5 for the single-interval model.

The sensitivity of $p^*_{t,i}$ with respect to t is demonstrated in Figure 4.6 for a five-interval planning horizon ($T = 5$). We observe in this figure that the trend varies depending on



(a) $\boldsymbol{Q}_l$ **(b)** $\boldsymbol{Q}_h$

Figure 4.4: Optimal permanent recruitment threshold at the beginning of a two-interval planning horizon as a function of waiting cost c_w for $c = 0.0$, $c_g = 1.5$, and $\kappa^0 = \kappa^1 = 0.1$.

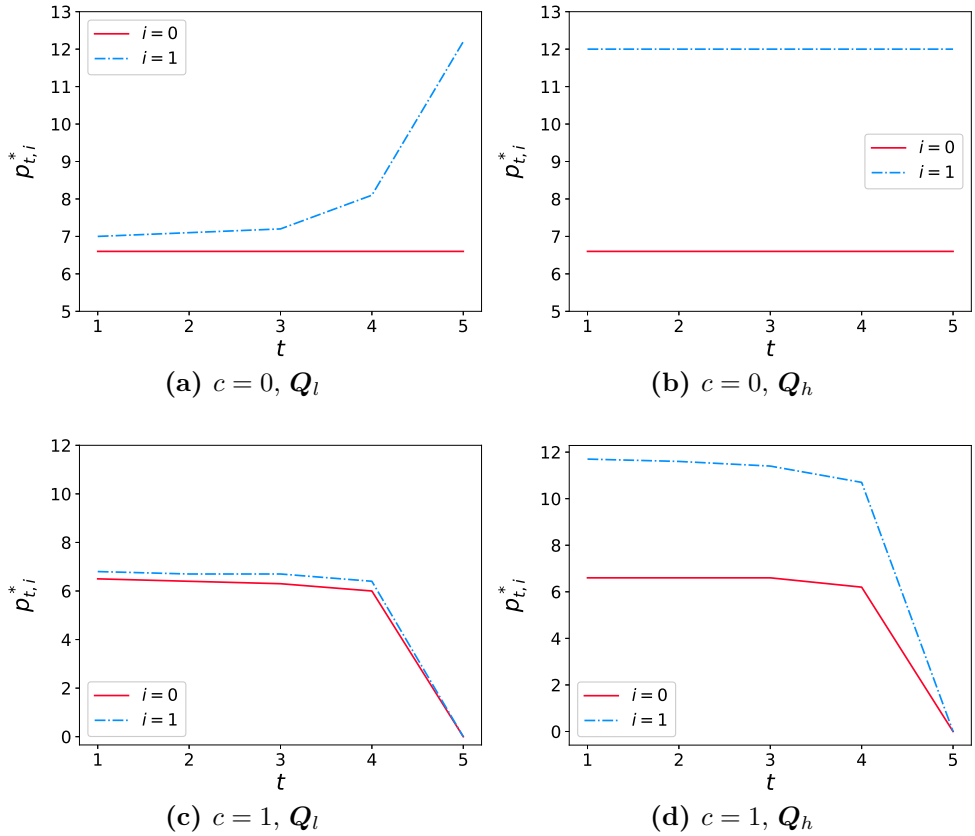


(a) $\boldsymbol{Q}_l$ **(b)** $\boldsymbol{Q}_h$

Figure 4.5: Optimal permanent recruitment threshold at the beginning of a two-interval planning horizon as a function of demand rate uncertainty $\kappa^0 = \kappa^1 = \kappa$ for $c = 0$, $c_g = 1.5$, and $c_w = 0.5$.

the value of c. More specifically, $p^*_{t,i}$ is either constant or increases as we approach the end of planning horizon for $c = 0$, while it typically decreases for $c = 1$. This is because the penalty for dismissing permanent employees discourages the provider from recruiting them towards the end of the planning horizon.

We also conclude from the plots in Figures 4.2 to 4.6 that more permanent employees should be hired when the current state of demand is high than when it is low, i.e., $p^*_{t,1} > p^*_{t,0}$.



(a) $c = 0$, $\boldsymbol{Q}_l$ **(b)** $c = 0$, $\boldsymbol{Q}_h$

(c) $c = 1$, $\boldsymbol{Q}_l$ **(d)** $c = 1$, $\boldsymbol{Q}_h$

Figure 4.6: Optimal permanent recruitment threshold as a function of time t for a five-interval planning horizon with $c_g = 1.5$, $c_w = 0.5$, and $\kappa^0 = \kappa^1 = 0.1$.

We further observe that $p^*_{t,i}$ corresponding to the transition probability matrix $\boldsymbol{Q}_h$ is larger or equal to that corresponding to $\boldsymbol{Q}_l$.

4.3.2 Savings Evaluation

We evaluate the potential savings obtained from our dynamic recruitment policy as compared to a myopic recruitment policy, which identifies the number of permanent positions in each interval based only on the cost of that interval and does not take the cost of future intervals into account. More specifically, given state (n_t, i), we define the myopic permanent

recruitment policy as

$$a_t^m(n_t, i) = \arg\min_{a_t}\{\mathbb{E}\left[v(\Lambda^i, n_t + a_t)\right] : a_t \in \mathbb{R}^+\}, \tag{4.11}$$

for each interval $t = 1, \ldots, T$. Note that the myopic policy yields the same permanent recruitment number as the single-interval model proposed in Chapter 3 excluding the impact of Q_t.

We perform two sets of experiments, one with $T = 2$ and the other with $T = 5$. We investigate the impact on savings of transition probability matrices and the observed initial state of the demand. For each set of parameters, $p_{t,i}^*$ and its corresponding cost are evaluated via Algorithm 3. The myopic policy $a_t^m(n_t, i)$ is estimated from (4.11) by complete enumeration over values of $a_t \in [0, 0.1, \cdots, 50]$, and its cost is obtained using the following recursive equation:

$$W_t(n_t, i) = \mathbb{E}\left[v(\Lambda^i, n_t + a_t^m(n_t, i))\right] + \alpha \sum_{j \in \mathcal{S}} q_{i,j} W_{t+1}(n_t + a_t^m(n_t, i), j), \tag{4.12}$$

for $t = 1, \ldots, T$, where

$$W_{T+1}(n, i) = cn, \tag{4.13}$$

for all $i \in \mathcal{S}$. We set $\alpha = 0.8$, $c = r_o = 0.0$, $\xi^0 = 5.0, \xi^1 = 10.0$, $c_g = 1.5$, and $c_w = 0.5$. We consider three variations for $\boldsymbol{Q}_h$, i.e.,

$$\boldsymbol{Q}_{h,1} = \begin{pmatrix} 0.01 & 0.99 \\ 0.01 & 0.99 \end{pmatrix}, \boldsymbol{Q}_{h,2} = \begin{pmatrix} 0.1 & 0.9 \\ 0.1 & 0.9 \end{pmatrix}, \boldsymbol{Q}_{h,3} = \begin{pmatrix} 0.3 & 0.7 \\ 0.3 & 0.7 \end{pmatrix}, \tag{4.14}$$

to represent increasing probabilities of moving to a low state for the situation where transition to a high state is more likely than a low state, and three variations for $\boldsymbol{Q}_l$, i.e.,

$$\boldsymbol{Q}_{l,1} = \begin{pmatrix} 0.99 & 0.01 \\ 0.99 & 0.01 \end{pmatrix}, \boldsymbol{Q}_{l,2} = \begin{pmatrix} 0.9 & 0.1 \\ 0.9 & 0.1 \end{pmatrix}, \boldsymbol{Q}_{l,3} = \begin{pmatrix} 0.7 & 0.3 \\ 0.7 & 0.3 \end{pmatrix}, \tag{4.15}$$

to represent increasing probabilities of moving to a high state for the situation where transition to a low state is more likely than a high state.

Table 4.1 illustrates the results for the dynamic and myopic policies with $T = 2$ for $\boldsymbol{Q} \in \{\boldsymbol{Q}_{h,1}, \boldsymbol{Q}_{h,2}, \boldsymbol{Q}_{h,3}\}$, i.e., the situation with a higher probability of transition to a high demand state than a low demand state. They suggest that, when the initial observed demand state is low, the myopic and dynamic policies are exactly the same, hence no savings. This is because there is no value in earlier recruitment of permanent HCWs, even if the demand is expected to increase in the future, when the provider is able to hire as many permanent HCWs as needed in the next interval. There is a small difference between myopic and dynamic policies, and thus small savings, when the initial observed demand state is high and there is some probability of moving to a low state (see panels (b) and (c)). This is because the possibility of moving to a lower demand state in the next interval is captured in the dynamic policy, hence a smaller number of permanent HCWs are recruited.

Table 4.2 illustrates the results for the dynamic and myopic policies with $T = 2$ for $\boldsymbol{Q} \in \{\boldsymbol{Q}_{l,1}, \boldsymbol{Q}_{l,2}, \boldsymbol{Q}_{l,3}\}$, i.e., the situation with a higher probability of transition to a low demand state than a high demand state. We observe that dynamic and myopic policies

Table 4.1: Comparison between the myopic and MDP policies with $T = 2$ and $\kappa^0 = \kappa^1 = 0.1$ for: (a) $\boldsymbol{Q}_{h,1}$, (b) $\boldsymbol{Q}_{h,2}$, and (c) $\boldsymbol{Q}_{h,3}$.

(a)

i	Myopic		MDP		Saving
	$W_1(0,i)$	$a_1^m(0,i)$	$V_1(0,i)$	$p_{1,i}^*$	(%)
0	20.0953	6.6	20.0953	6.6	0.0
1	26.6190	12.2	26.6190	12.2	0.0

(b)

i	Myopic		MDP		Saving
	$W_1(0,i)$	$a_1^m(0,i)$	$V_1(0,i)$	$p_{1,i}^*$	(%)
0	19.6281	6.6	19.6281	6.6	0.0
1	26.4572	12.2	26.4488	12	0.03

(c)

i	Myopic		MDP		Saving
	$W_1(0,i)$	$a_1^m(0,i)$	$V_1(0,i)$	$p_{1,i}^*$	(%)
0	18.5897	6.6	18.5897	6.6	0.0
1	26.0776	12.2	26.0254	11.6	0.20

produce the same results when the current demand state is low. When the current demand state is high, however, the dynamic policy proposes a substantially smaller number of permanent HCWs. This is to avoid over-staffing in the next interval, and creates savings of 1.8%, 3.89%, and 5.0% for $\boldsymbol{Q}_{l,3}$, $\boldsymbol{Q}_{l,2}$, $\boldsymbol{Q}_{l,1}$, respectively.

Table 4.2: Comparison between the myopic and MDP policies with $T = 2$ and $\kappa^0 = \kappa^1 = 0.1$ for: (a) $\boldsymbol{Q}_{l,1}$, (b) $\boldsymbol{Q}_{l,2}$, (c) $\boldsymbol{Q}_{l,3}$.

(a)

i	Myopic		MDP		Saving
	$W_1(0,i)$	$a_1^m(0,i)$	$V_1(0,i)$	$p_{1,i}^*$	(%)
0	15.0074	6.6	15.0074	6.6	0
1	24.8574	12.2	23.6141	7.8	5.00

(b)

i	Myopic		MDP		Saving
	$W_1(0,i)$	$a_1^m(0,i)$	$V_1(0,i)$	$p_{1,i}^*$	(%)
0	15.4746	6.6	15.4746	6.6	0
1	25.0192	12.2	24.0446	8.1	3.89

(c)

i	Myopic		MDP		Saving
	$W_1(0,i)$	$a_1^m(0,i)$	$V_1(0,i)$	$p_{1,i}^*$	(%)
0	16.513	6.6	16.513	6.6	0
1	25.3787	12.2	24.9115	9.7	1.84

Tables 4.3 and 4.4 present the results of experiments with the same parameters as those presented in Tables 4.1 and 4.2, respectively, but with $T = 5$. We observe in these tables that the savings increase to up to 0.56% (17.03%) for the situation with a higher probability

of moving to a high (low) state. This is because with a longer planning horizon, the system is more likely to move to a high demand state in one of the intermediate intervals, which is the situation where the dynamic policy creates savings.

Table 4.3: Comparison between the myopic and MDP policies with $T = 5$ and $\kappa^0 = \kappa^1 = 0.1$ for: (a) $\boldsymbol{Q}_{h,1}$, (b) $\boldsymbol{Q}_{h,2}$, (c) $\boldsymbol{Q}_{h,3}$.

(a)

i	Myopic		MDP		Saving
	$W_1(0,i)$	$a_1^m(0,i)$	$V_1(0,i)$	$p_{1,i}^*$	(%)
0	43.1691	6.6	43.1691	6.6	0
1	49.693	12.2	49.693	12.2	0

(b)

i	Myopic		MDP		Saving
	$W_1(0,i)$	$a_1^m(0,i)$	$V_1(0,i)$	$p_{1,i}^*$	(%)
0	42.3568	6.6	42.3404	6.6	0.03
1	49.2154	12.2	49.1893	12	0.05

(c)

i	Myopic		MDP		Saving
	$W_1(0,i)$	$a_1^m(0,i)$	$V_1(0,i)$	$p_{1,i}^*$	(%)
0	40.391	6.6	40.1834	6.6	0.51
1	48.1542	12.2	47.882	11.3	0.56

Table 4.4: Comparison between the myopic and MDP policies with $T = 5$ and $\kappa^0 = \kappa^1 = 0.1$ for: (a) $\boldsymbol{Q}_{l,1}$, (b) $\boldsymbol{Q}_{l,2}$, (c) $\boldsymbol{Q}_{l,3}$.

(a)

i	Myopic		MDP		Saving
	$W_1(0,i)$	$a_1^m(0,i)$	$V_1(0,i)$	$p_{1,i}^*$	(%)
0	28.2042	6.6	28.1269	6.6	0.27
1	44.4928	12.2	36.9151	6.9	17.03

(b)

i	Myopic		MDP		Saving
	$W_1(0,i)$	$a_1^m(0,i)$	$V_1(0,i)$	$p_{1,i}^*$	(%)
0	30.5025	6.6	29.8927	6.6	1.99
1	44.9704	12.2	38.6621	7	14.02

(c)

i	Myopic		MDP		Saving
	$W_1(0,i)$	$a_1^m(0,i)$	$V_1(0,i)$	$p_{1,i}^*$	(%)
0	34.673	6.6	33.76	6.6	2.63
1	46.0316	12.2	42.4607	7.3	7.75

Overall, we conclude that the dynamic policy is more likely to create savings as compared to the myopic policy when the current state of the demand rate is high and transition to a low state is more likely than a high state. The savings also increase as the number of intervals within the planing horizon increases.

4.4 Summary

We proposed a multi-interval blended recruitment framework for a provider facing periods of highly uncertain demand. This framework captures the trade-offs between staffing costs, recruitment lead times, and placement durations of temporary and permanent HCWs. To the best of our knowledge, this is the first study that captures these trade-offs explicitly in a blended workforce environment.

In our framework, a two-stage decision making process is repeated in each interval given the existing number of permanent workers and the observed state of demand at the beginning of the interval. The number of permanent workers is decided in the first stage, and the number of temporary workers in the second stage. The state of demand identifies the distribution of demand rate out of a finite set of distributions, and evolves according to a discrete-time MC. This aims to capture the variations in the mean and/or uncertainty of demand rate over time. Formulating the problem as an MDP, we proved that the optimal permanent recruitment policy is a state-dependent hire-up-to type. This reduces the dimension of the search space for each interval to one, enabling us to estimate the optimal permanent policy for a reasonable-size problem in a short time.

Assuming a high and a low state for the demand rate distribution, we investigated the sensitivity of the hire-up-to value with respect to various parameters numerically. In particular, we illustrated that, while this value increases with the cost rate of temporary workers and waiting cost of patients, it decreases with the dismissal cost of permanent

workers at the end of the horizon. We also showed that the behaviour of the hire-up-to threshold is non-monotone with respect to demand rate uncertainty and time interval. Our results further suggest that the threshold value is typically higher when the current state of the demand rate distribution is high and subsequent intervals are more likely to be a high state.

We finally evaluated the savings obtained from our dynamic policy as compared to a myopic policy in which the permanent recruitment decision is based only on the cost of the current interval. For a two-state demand rate distribution, we observed that the savings are likely to be significant when the probability of transition to a low state is higher than a high state. This is because the dynamic policy captures this potential transition to a low demand state in the future and thus recruits a smaller number of permanent workers to avoid over-staffing. The amount of savings are also likely to increase with the length of the planning horizon.

Chapter 5

Summary, Conclusions and Future Research

5.1 Summary

The incentives for and challenges of using temporary workforce in the healthcare sector were reviewed in Chapter 1. This review concluded the necessity of deploying temporary HCWs, alongside permanent HCWs, for delivering an efficient and quality service. It was highlighted, however, that finding the right mix of these two types of workforce is difficult, in particular for periods of highly uncertain demand. This is mainly due to different timings of permanent and temporary recruitment and the random nature of the recruitment process. The former implies an asymmetry in demand information at the times of temporary and permanent recruitment, and the latter results in some positions not being filled. This led

to the general aim of the thesis, which was outlined as developing optimization frameworks to inform blended recruitment decision making for periods of highly uncertain demand. It was argued that such frameworks should capture the trade-off between shorter recruitment lead time and placement duration of temporary workers versus the lower staffing cost of permanent workers. We explained that the shorter recruitment lead time of temporary workers implies a more accurate demand information at the time of their recruitment as well as a higher likelihood of success in recruiting for the desired number of positions. Their shorter placement duration, on the other hand, implies that future demand information need not be taken into account for their recruitment. Lower staffing cost of permanent workers, however, makes them attractive for the provider. We set out to investigate these in two main chapters of the thesis; Chapter 3 was dedicated to the trade off between staffing costs and recruitment lead times, and Chapter 4 to the trade off between staffing costs, recruitment lead times, and placement durations.

The literature of recruitment models was reviewed in Chapter 2. This literature was divided into two main categories, the mid-term recruitment literature and the long-term recruitment literature. In the former category, a single-interval planning horizon is considered with a single opportunity for permanent recruitment, while in the latter category, a multi-interval planning horizon is considered with multiple opportunities for permanent recruitment. It was noted that some studies in the mid-term recruitment category consider a single opportunity for temporary recruitment while some others allow multiple opportunities. From a methodological perspective, the mid-term literature was divided into single- and

two-stage streams. In the single-stage stream, recruitment decisions of permanent and temporary workers are made at the same time, while in the two-stage stream these decisions are made at different times. We decided to follow a two-stage approach in our research as it would allow the asymmetry in demand information. Having reviewed the mid-term two-stage recruitment literature, a gap was identified for a framework which captures the dynamics of the service delivery in delay systems, accounts for the uncertain nature of permanent recruitment as well as the asymmetry of demand information, and produces exact results irrespective of the system size.

Two studies were identified in our review of the long-term recruitment literature in Chapter 2. We highlighted that neither of these two studies consider the co-existence of temporary and permanent workers explicitly. A gap was therefore identified for a framework that considers blended recruitment of temporary and permanent workforce, taking into account the differences in their recruitment lead times, staffing costs, and placement durations.

In Chapter 3, we proposed a two-stage stochastic optimization framework to inform recruitment decision making for a highly uncertain demand period. We provided analytical characterizations for the optimal first- and second- stage decisions, proposed fast numerical algorithms for evaluating their values, and proved some insensitivity and monotonicity properties for optimal recruitment policies and their corresponding cost. We further investigated the potential benefit/risk of delaying permanent advertisement, and derived some managerial insight using numerical experiments. A case study was provided to

illustrate how our approach can be applied in conjunction with a simulation model to inform nurse recruitment in inpatient departments. An extension of the main model, which considers multiple opportunities for temporary recruitment, was also proposed.

In Chapter 4, we proposed a multi-interval two-stage optimization framework to capture the trade-off between staffing costs, recruitment lead-times, and placement durations. We ignored the uncertainty in permanent recruitment by assuming that the required number of permanent positions can always be filled. Formulating the problem as an MDP, we proved that the optimal policy is a hire-up-to policy. Using numerical experiments, we investigated the impact of various model parameters on the hire-up-to value. We also assessed the savings obtained from our dynamic policy as compared to a myopic policy.

We summarize our major methodological and managerial contributions in the two main chapters of the thesis in §5.2. We then identify some areas for future research in §5.3.

5.2 Contributions

Our methodological contributions in Chapter 3 are as follows. First, in contrast with the studies in the blended workforce literature (e.g., Lu and Lu, 2017; Hu et al., 2021b), our framework accounts for the uncertainty associated with permanent recruitment by incorporating a probability distribution for the number of qualified applicants. This makes our framework more realistic. It also allows investigating the potential benefit of obtaining a more accurate demand information by delaying permanent advertisement versus the

associated risk of a shorter advertisement window. To the best of our knowledge, such investigation has not been conducted in the literature.

Second, our framework captures the dynamics of service delivery in delay systems, i.e., systems where all requests must be served. The other two studies which capture the dynamics of service delivery explicitly in the blended recruitment literature are those of Dong and Ibrahim (2020) and Hu et al. (2021b). Both of these studies, however, are focused on abandonment systems, i.e., systems where some requests may leave the system before their service begins. This implies that our framework is more appropriate for settings such as inpatient departments or care homes as in these settings all requests (of admitted patients) must be served.

Third, we derived our results following an exact approach without making any assumption on the type or scale of the delay queueing model. This implies that our approach can be applied for any delay queueing model as long as a set of intuitive assumptions are met. Further, the results obtained from our model are valid irrespective of the system size. This is an important feature since the systems representing nursing care in inpatient settings are typically small, rendering asymptotic large-scale approximations, as followed in Hu et al. (2021b), inaccurate.

Fourth, we illustrated how a simulation model can be incorporated into our framework to provide recruitment decision support for multi-resource environments such as inpatient departments. This complements the study of Yankovic and Green (2011) by distinguishing between temporary and permanent nurses and capturing the two-stage nature of their

recruitment. We further showed that single-resource approximations based on our analytical models provide reliable and robust results with substantially less effort.

Fifth, we developed an extension of our main framework which allows multiple opportunities for temporary recruitment by dividing the service delivery period into smaller segments. Our framework accounts for the possibility of correlated demand in different segments, and our characterization of the optimal permanent recruitment shows that this correlation influences the optimal permanent recruitment decision only when the multi-variate demand distribution is not closed under marginalization. To the best of our knowledge, the impact of demand correlation on the optimal permanent recruitment decision has not been investigated in the studies with a similar decision making process (e.g., Kao and Queyranne, 1985; Pinker and Larson, 2003).

The managerial contributions in Chapter 3 are a follows. First, we assessed the value of temporary staffing by comparing the expected overall cost obtained under our two-stage framework with the cost of a single-stage model where only permanent recruitment is permitted. We observed that, except in cases where the temporary cost rate is extremely high, there is value in recruiting temporary HCWs. We also showed that this value is likely to increase with the system scale but decrease with the cost of waiting. This result, however, contradicts the findings from Harper et al. (2010), which suggests permanent staffing as a more cost-effective approach when dealing with fluctuations in demand. This contradictory result can be due to the fact that Harper et al. (2010) uses either nurse-to-patient ratio method or the dependency-activity-quality method for identifying staffing levels. Our

contribution further complements the results from Dong and Ibrahim (2020) and Hu et al. (2021b) by showing that temporary staffing is valuable even when demand rate uncertainty is very low. In particular, the findings from Dong and Ibrahim (2020) suggest a blended staffing approach as the optimal strategy in high-demand periods, and the analyses in Hu et al. (2021b) illustrate that temporary staffing is most beneficial when demand rate uncertainty dominates the system stochasticity.

Second, we assessed the value of obtaining demand rate distribution by comparing the expected overall cost obtained under our two-stage framework with the cost of a two-stage model which uses only the average demand rate. We observe that when demand rate uncertainty is moderate to high, there is value in obtaining the demand rate distribution and incorporating it into recruitment decision making. Otherwise, using only the average demand rate would suffice. We illustrated how this average can be applied in our framework to evaluate the optimal number of permanent positions. Such investigation has not been conducted in the only other study that captures demand rate uncertainty (i.e., the study of Hu et al., 2021b).

Third, we proved that the optimal first- and second-stage decisions and their corresponding costs increase with service time variability. This result has not been reported in the blended workforce literature. Finally, we illustrated via numerical experiments that the optimal number of permanent positions and the corresponding system cost show a non-monotone behaviour with respect to demand rate uncertainty. In particular, both of these values exhibit a decreasing trend when demand rate uncertainty exceeds a threshold.

The main implication is that delaying advertisement is less likely to be beneficial when demand rate uncertainty is already high and the cost of temporary recruitment is small relative to the cost of patients waiting. Our study is in fact the first that explores the impact of demand rate uncertainty on the optimal permanent recruitment decision, and investigates its implications for delaying advertisement.

Our main methodological contribution in Chapter 4 is capturing the joint impact of different recruitment lead times, staffing costs, and placement durations of temporary and permanent workers. In particular, the difference in placement durations is captured by assuming that temporary workers are released at the end of each interval, whereas permanent workers stay in the system until the end of the planning horizon. The difference in recruitment lead time is captured by considering a two-stage decision making in each interval, and modelling the demand as a Poisson mixture process. The difference in staffing cost are captured by different cost rates for temporary and permanent workers. Similar to the framework in Chapter 3, our framework in Chapter 4 is flexible with respect to the delay queueing model used for representing service delivery. This framework also uses a Markov chain to represent the variations of mean demand rate or the uncertainty around it over time. In contrast to the studies of Gans and Zhou (2002) and Ahn et al. (2005), we explicitly consider temporary recruitment and capture the higher variability of demand relative to the standard Poisson process. The main managerial implication in Chapter 4 was that the savings created by a dynamic policy is likely to increase when there is a higher probability of transition to a low demand state, and a longer planning horizon. Note

that these findings are restricted to situations where there is no limitation for permanent recruitment.

5.3 Future Research

We propose future areas of research for the mid-term and long-term recruitment problems below.

The methodology we proposed for the mid-term recruitment problem in Chapter 3 is based on a cost function defined in terms of the number of requests waiting in the queue or being served. It would be interesting to investigate if cost functions based on other performance metrics, e.g., the mean number of requests in the queue or the percentage of requests waiting above a certain limit, meet the properties in Assumption 1, and so can be applied with our framework. Also, our methodology is restricted to delay systems. The application area would expand substantially if it is generalized to cover abandonment systems. This would be an important generalization as the only study that captures the dynamics of service delivery in abandonment queues, i.e., the study of Hu et al. (2021b), applies only to large-scale systems. Further, our methodology works only with stationary demand rates. The extension proposed in Chapter 3 does consider time-variation of demand over different segments, but it assumes that the system achieves a steady-state within each segment of the service delivery. This does not necessarily happen in reality and so another area worth exploring is expanding our methodology to capture time-varying

rates during different segments of service delivery. The main difficulty of such extension is that, with time-varying demand rates, exact performance evaluation would be challenging. There exist numerical methods and approximations (see, e.g., Defraeye and Nieuwenhuyse, 2016), however, which can be incorporated into our framework. Accounting for staff absenteeism would also make our models more realistic. This can be achieved by assuming a random percentage of recruited permanent and temporary workers will not show up for work during the HUDP. In §3.5, we introduced delayed advertisement as an approach to reduce the demand rate uncertainty. Using machine learning algorithms to estimate the probability distribution of demand rate with more accuracy (instead of maximum likelihood and Kolmogorov-Smirnov goodness of-fit test used in §3.6) or using regression models with realistic features to predict the exact value of demand rate during HUDP are other approaches to further reduce the demand rate uncertainty.

In the framework we proposed in Chapter 4, we did not capture the randomness of the recruitment process to simplify the analysis. An area worth exploring would therefore be including the random number of qualified applications received during the permanent recruitment period in the formulation and investigating if the hire-up-to structure still holds. This would change the recursive function proposed in Equation (4.3) to

$$V_t(n_t, i) = \min_{a_t} \Big\{ \mathbb{E}\left[v(\Lambda^i, n_t + \min\{X_t, a_t\})\right] + \alpha \sum_{j \in \mathcal{S}} q_{i,j} \mathbb{E}\left[V_{t+1}(n_t + \min\{X_t, a_t\}, j)\right] : a_t \in \mathbb{R}^+ \Big\}, \quad (5.1)$$

for $t = 1, \ldots, T$, where X_t is the random number of qualified applications received during the permanent recruitment period of interval t, and

$$V_{T+1}(n, i) = cn,$$

for all $n \in \mathbb{R}^+$ and $i \in \mathcal{S}$. This addition may lead to new insights. For example, we observed in Chapter 4 that the myopic policy matches the dynamic policy in a two-interval planning horizon where the current demand state is low. Inclusion of recruitment randomness may change this as a dynamic policy may advise advertising a larger number of positions in the current interval when there is a high probability of transition to a high demand state. This is to increase the possibility of recruiting the required number of permanent staff for the next interval.

Another avenue for future research is considering learning and turn over of permanent workforce over time in the spirit of Gans and Zhou (2002). Learning can be captured by defining different skill levels for permanent workers, and modeling their evolution as a MC. Turn over, on the other hand, should be modeled as a function of workload as reported in the empirical research, see, e.g., Holland et al. (2019). Predicting the performance of recruited temporary and permanent employees may well be another area worth exploring in future research. In particular, one can use historical data on the work experience, education, and training hours of the hired workers, and then apply machine learning algorithms such as classification to predict how well or poor they will be doing at their jobs.

References

Waleed Abo-Hamad and Amr Arisha. Simulation-based framework to improve patient experience in an emergency department. *European Journal of Operational Research*, 224 (1):154–166, 2013.

Katharine G. Abraham. Flexible staffing arrangements and employers' short-term adjustment strategies. Technical report, National Bureau of Economic Research, Cambridge, MA, 1988.

Hyun-Soo Ahn, Rhonda Righter, and J. George Shanthikumar. Staffing decisions for heterogeneous workers with turnover. *Mathematical Methods of Operations Research*, 62 (3):499–514, 2005.

Linda H. Aiken, Douglas M. Sloane, and Jennifer L. Klocinski. Hospital nurses' occupational exposure to blood: Prospective, retrospective, and institutional reports. *American Journal of Public Health*, 87(1):103–107, 1997.

Linda H. Aiken, Ying. Xue, Sean P. Clarke, and Douglas M. Sloane. Supplemental nurse

staffing in hospitals and quality of care. *Journal of Nursing Administration*, 37(7-8): 335–342, 2007.

Linda H. Aiken, Jingjing. Shang, Ying. Xue, and Douglas M. Sloane. Hospital use of agency-employed supplemental nurses and patient mortality and failure to rescue. *Health Services Research*, 48(3):931–948, 2013.

Shoshana Anily and Moshe Haviv. Cooperation in service systems. *Operations Research*, 58(3):660–673, 2010.

Sung-Heui Bae, Barbara Mark, and Bruce Fried. Use of temporary nurses and nurse and patient safety outcomes in acute care hospital units. *Health Care Management Review*, 35(4):333–344, 2010.

Sung-Heui Bae, Maureen Kelly, Carol S. Brewer, and Alexandra Spencer. Analysis of nurse staffing and patient outcomes using comprehensive nurse staffing characteristics in acute care nursing units. *Journal of Nursing Care Quality*, 29(4):318–326, 2014.

Félix Belzunce, Carolina Martínez-Riquelme, and Julio Mulero. Chapter 2 - univariate stochastic orders. In Félix Belzunce, Carolina Martínez-Riquelme, and Julio Mulero, editors, *An Introduction to Stochastic Orders*, pages 27–113. Academic Press, 2016.

Anna Berg Jansson and Åsa Engström. Working together: critical care nurses experiences of temporary staffing within Swedish health care: A qualitative study. *Intensive and Critical Care Nursing*, 41:3–10, 2017.

Oded Berman and Richard C. Larson. Determining optimal pool size of a temporary call-in work force. *European Journal of Operational Research*, 73(1):55–64, 1994.

Richard P. Brent. An Algorithm with Guaranteed Convergence for Finding a Zero of a Function. In *Algorithms for Minimization without Derivatives*, chapter 4. Prentice-Hall, Englewood Cliffs, NJ, 1973.

Michael D. Cabana and Sandra H. Jee. Does continuity of care improve patient outcomes? *Journal of Family Practice*, 53(12):974–980, 2004.

Francis de Véricourt and Otis B. Jennings. Nurse Staffing in Medical Units: A Queueing Perspective. *Operations Research*, 59(6):1320–1331, 2011.

Mieke Defraeye and Inneke Van Nieuwenhuyse. Staffing and scheduling under nonstationary demand for service: A literature review. *Omega*, 58:4–25, 2016.

Joanne M. Dochterman and Gloria M. Bulechek. *Nursing Interventions Classification (NIC)*. Mosby, St. Louis, 4 edition, 2004.

Jing Dong and Rouba Ibrahim. Managing supply in the on-demand economy: Flexible workers, full-time employees, or both? *Operations Research*, 68(4):1238–1264, 2020.

Carole A. Estabrooks, William K. Midodzi, Greta G. Cummings, Kathryn L. Ricker, and Phyllis Giovannetti. The impact of hospital nursing characteristics on 30-day mortality. *Nursing research*, 54(2):74–84, 2005.

Noah Gans and Yong Pin Zhou. Managing learning and turnover in employee staffing. *Operations Research*, 50(6):991–1006, 2002.

Donald Gross, John F. Shortie, James M. Thompson, and Carl M. Harris. *Fundamentals of Queueing Theory*. John Wiley & Sons, Inc., Hoboken, New Jersey, 4 edition, 2008.

Haresh Gurnani and Christopher S. Tang. Note: Optimal ordering decisions with uncertain cost and demand forecast updating. *Management Science*, 45(10):1456–1462, 1999.

Shlomo Halfin and Ward Whitt. Heavy-traffic limits for queues with many Exponential servers. *Operations Research*, 29(3):567–588, 1981.

Paul R. Harper, N H. Powell, and Janet E. Williams. Modelling the size and skill-mix of hospital nursing teams. *Journal of the Operational Research Society*, 61(5):768–779, 2010.

Peter Holland, Tse Leng Tham, Cathy Sheehan, and Brian Cooper. The impact of perceived workload on nurse satisfaction with work-life balance and intention to leave the occupation. *Applied Nursing Research*, 49(March):70–76, 2019.

William C. Horrace. Some results on the multivariate truncated normal distribution. *Journal of Multivariate Analysis*, 94(1):209–221, 2005.

Yue. Hu, Kenrick D. Cato, Carrie W. Chan, Jing. Dong, Nicholas. Gavin, Sarah C. Rossetti, and Bernard P. Chang. Use of real-time information to predict future arrivals in the emergency department. Working Paper, Columbia Business School, 2021a.

Yue Hu, Carri W Chan, and Jing Dong. Prediction-driven surge planning with application in the emergency department. 2021b. URL `http://www.columbia.edu/~cc3179/SurgeStaffing_2021.pdf`.

Katbarine K. Hughes and Richard J. Marcantonio. Recruitment, Retention, and Compensation of Agency and Hospital Nurses. *Journal of Nursing Administration*, 21(10):46–52, 1991.

Keith Hurst. Selecting and applying methods for estimating the size and mix of nursing teams: a systematic review of the literature commissioned by the Department of Health. Technical report, 2002. URL `https://web2.aabu.edu.jo/tool/course_file/1001463_calculating%20staffing.pdf`.

A A. Jagers and Erik A. van Doorn. Convexity of functions which are generalizations of the Erlang loss function and the Erlang delay function. *SIAM Review*, 33(2):281–282, 1991.

Angus Jeang. Flexible nursing staff planning when Patient Demands are Uncertain. *Journal of Medical Systems*, 20(4):173–182, 1996.

Geurt Jongbloed and Ger Koole. Managing uncertainty in call centres using poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17(4):307–318, 2001.

Edward P. C. Kao and Maurice Queyranne. Budgeting Costs of Nursing in a Hospital. *Management Science*, 31(5):608–621, 1985.

Frank Karsten, Marco Slikker, and Geert-Jan van Houtum. Resource pooling and cost

allocation among independent service providers. *Operations Research*, 63(2):476–488, 2015.

Saravanan Kesavan, Bradley R. Staats, and Wendell Gilland. Volume flexibility in services: The costs and benefits of flexible labor resources. *Management Science*, 60(8):1884–1906, 2014.

Susan Feng Lu and Lauren Xiaoyuan Lu. Do mandatory overtime laws improve quality? Staffing decisions and operational flexibility of nursing homes. *Management Science*, 63 (11):3566–3585, 2017.

Solveig Lundgren and Kerstin Segesten. Nurses' use of time in a medical-surgical ward with all-RN staffing. *Journal of Nursing Management*, 9(1):13–20, 2001.

Shimrit Maman. *Uncertainty in the Demand for Service: The Case of Call Centers and Emergency Departments*. Master thesis, Senate of the Technion - Israel Institute of Technology, 2009.

Avishai Mandelbaum and Martin I. Reiman. On pooling in queueing networks. *Management Science*, 44(7):971–981, 1998.

National Audit Office. Improving the use of temporary nursing staff in nhs acute and foundation trusts. Technical Report July, Department of Health, 2006. URL `https://webarchive.nationalarchives.gov.uk/ukgwa/20170207052351/https://www.nao.org.uk/wp-content/uploads/2006/07/05061176.pdf`.

National Audit Office. Managing the supply of nhs clinical staff in england. Technical Report February, Department of Health, 2016. URL https://www.nao.org.uk/wp-content/uploads/2016/02/Managing-the-supply-of-NHS-clinical-staff-in-England.pdf.

NHS Vacancy Statistics. NHS Vacancy Statistics England April 2015 – March 2021 Experimental Statistics, 2021. URL https://digital.nhs.uk/data-and-information/publications/statistical/nhs-vacancies-survey/april-2015---march-2021.

Antonio Pacheco. Some Properties of the Delay Probability in M/M/s/s+c Systems. Technical report, Cornell University, Ithaca, NY, 1993.

Edieal Pinker and Vera Tilson. The impact of technology on the labor procurement process. In *46th Hawaii International Conference on System Sciences*, pages 4176–4185. IEEE, 2013.

Edieal J. Pinker and Richard C. Larson. Optimizing the use of contingent labor when demand is uncertain. *European Journal of Operational Research*, 144(1):39–55, 2003.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, New York, 1994.

Ruwen Qin, David A. Nembhard, and Walter L. Barnes. Workforce flexibility in operations management. *Surveys in Operations Research and Management Science*, 20(1):19–33, 2015.

Christine Roche, Michael O Brien-pallas, and Linda Catling-paull. The implications of staff "churn" for nurse managers, staff, and patients. *Nursing Economics*, 27(2):103–110, 2009.

Connie Roseman and John M. Booker. Workload and environmental factors in hospital medication errors. *Nursing research*, 44(4):226–230, 1995.

Royal College of Nursing. NHS conditions of employment, 2021. URL `https://www.rcn.org.uk/employment-and-pay/nhs-conditions-of-employment`.

Sukyong Seo and Joanne Spetz. Demand for temporary agency nurses and nursing shortages. *INQUIRY: The Journal of Healthcare Organization, Provision, and Financing*, 50(3): 216–228, 2013.

Moshe Shaked and J. George Shanthikumar, editors. *Stochastic Orders*. Springer New York, New York, NY, 2007.

Karen M. Stratton. Pediatric Nurse Staffing and Quality of Care in the Hospital Setting. *Journal of Nursing Care Quality*, 23(2):105–14, 2008.

Josef Svoboda, Stefan Minner, and Man Yao. Typology and literature review on multiple supplier inventory control models. *European Journal of Operational Research*, 293(1): 1–23, 2021.

Akira Takayama. *Mathematical Economics*. Cambridge University Press., New York, second edition, 1985.

The Kings Fund. Key facts and figures about the NHS, 2021. URL https://www.kingsfund.org.uk/audio-video/key-facts-figures-nhs.

The Open University. Tackling the nursing shortage. Technical Report May, 2018. URL https://cdn.ps.emap.com/wp-content/uploads/sites/3/2018/05/The_Open_University_Tackling_the_nursing_shortage_Report_2018.pdf.

George B. Thomas. *Thomas' Calculus*. Pearson Education, thirteenth edition, 2014.

Ruth Thorlby, Caroline Fraser, and Tim Gardner. Non-COVID-19 NHS care during the pandemic, dec 2020. URL https://www.health.org.uk/news-and-comment/charts-and-infographics/non-covid-19-nhs-care-during-the-pandemic.

Jos Van Ommeren and Giovanni Russo. Firm recruitment behaviour: Sequential or non-sequential search? *Oxford Bulletin of Economics and Statistics*, 76(3):432–455, 2014.

R Vovak. Staffing trends at Maryland hospitals: FY 2000 to FY 2010. Technical report, Maryland Hospital Association, Elkridge, 2010.

Tong Wang, Atalay Atasu, and Mümin Kurtuluş. A multiordering newsvendor model with dynamic forecast evolution. *Manufacturing & Service Operations Management*, 14(3): 472–484, 2012.

Yimin Wang and Brian Tomlin. To Wait or Not to Wait: Optimal Ordering Under Lead Time Uncertainty and Forecast Updating. *Naval Research Logistics*, 56:766–779, 2009.

Michael West, Suzie Bailey, and Ethan Williams. The courage of compassion: Supporting nurses and midwives to deliver high-quality care. Technical report, The King's Fund, 2020. URL `https://www.kingsfund.org.uk/publications/courage-compassion-supporting-nurses-midwives`.

Ying Xue, Joyce Smith, Deborah A. Freund, and Linda H. Aiken. Supplemental nurses are just as educated, slightly less experienced, and more diverse compared to permanent nurses. *Health Affairs*, 31(11):2510–2516, 2012.

Houmin Yan, Ke Liu, and Arthur Hsu. Optimal ordering in a dual-supplier system with demand forecast. *Production and Operations Management*, 12(1):30–45, 2003.

Natalia Yankovic and Linda V. Green. Identifying good nursing levels: A queuing approach. *Operations Research*, 59(4):942–955, 2011.