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

Saha Malaki[a,*], Navid Izady[a], Lilian M. de Menezes[a]

[a] *Bayes Business School (formerly Cass), City, University of London, London, UK*

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. In this paper, we propose a two-stage stochastic optimization framework to inform recruitment decision making for a period of highly uncertain demand in a setting where all patients must be served. The first stage 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 permanent recruitment process, stochasticity of the service delivery, and asymmetry in demand information at the times of permanent and temporary recruitment. 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 of their monotonicity properties. A case study based on data from a geriatric ward illustrates the application of our framework, and numerical experiments provide further managerial insights.

Keywords: OR in health services, Blended workforce recruitment, Queueing, Stochastic optimization, Demand and supply uncertainty

1. Introduction

In the past few decades, the healthcare sector has witnessed significant changes in the way that jobs are structured. Chronic staff shortages (Bae et al., 2010), long lead times in recruiting permanent staff (Lu & Lu, 2017), predictable and unpredictable variabilities in patient demand (Seo & 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). In the UK, for example, the total hours of temporary nurses

*Corresponding author
Email address: saha.malaki@bayes.city.ac.uk (Saha Malaki)

requested by the hospitals within the National Health Service (NHS) doubled from 2011 to 2015 (National Audit Office, 2016). Temporary workers provide volume flexibility, i.e., the ability to adjust staffing patterns flexibly and quickly in response to variations in patient demand and (un)availability of permanent staff (Kesavan et al., 2014). However, temporary skilled workers are more expensive for the provider than their permanent counterparts. In fact, findings from a recent survey imply 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). It is therefore important for healthcare providers to strike a balance between staffing costs and service quality by recruiting the right mix of permanent and temporary HCWs.

Finding this mix is challenging for the following reasons. First, permanent and temporary recruitment decisions are not contemporaneous; advertising for permanent workers typically starts 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 days/hours 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 during the permanent recruitment period. Second, there is uncertainty in recruitment since there is no guarantee that all required positions (especially permanent ones) 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 (NHS Improvement, 2018). The recent COVID-19 pandemic has also added to demand uncertainty; see, e.g., Thorlby et al. (2020) on fluctuations in emergency care demand during the pandemic. In addition to making the decision on timing permanent advertisement more critical, this calls for applying models other than the usual Poisson process for capturing demand uncertainty.

We focus on recruitment decision making concerning a mix of permanent and temporary HCWs for a provider facing a highly uncertain demand period (HUDP). The main trade-off in this decision making is that permanent HCWs are cheaper for the provider, 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 UK's NHS were not filled in 2020 (NHS Vacancy Statistics, 2021).

We consider a setting in which patients' requests arrive to the system during the HUDP, and 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 when only partial information about the demand for service 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 assumes that advertisement for permanent positions must begin at an exogenously given time. However, we also investigate the benefit/loss of delaying this advertisement, which could lead to more accurate demand information at the expense of a higher risk of not filling the advertised positions.

We model the patient demand as a Poisson mixture process, i.e., a Poisson process with a random rate; see, e.g., Jongbloed & Koole (2001). As we will explain in §2, this provides the desired features for demand modelling in the context of recruitment decision making for service systems. We represent the uncertainty in the permanent recruitment process by 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.

We analytically characterize the optimal first and second-stage recruitment decisions and propose fast numerical algorithms for finding their values. We prove that the optimal first-stage decision is insensitive to the probability distribution of the number of qualified applications, as long as the distribution support remains the same. We also prove that it is never optimal to staff the system with only temporary workers. Using stochastic ordering, we further prove 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 are 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 specific properties.

These properties are intuitive and we prove that they hold for three common queueing models: (i) a single-server approximation model with Exponential inter-arrival and service times; (ii) a single-server approximation model with Exponential inter-arrival and general service times; and (iii) a multi-server model with Exponential inter-arrival and service times. The first two models are approximations since they estimate the behaviour of a multi-server system by inflating the service rate of a single-server queue, but are useful as they provide further analytical tractability. In particular, we obtain a closed-form expression for the optimal second-stage decision and the corresponding cost for model (i), and prove that the optimal first- and second-stage decisions as well as their corresponding costs increase with service time variability for model (ii).

We combine analytical results with numerical experiments to derive managerial insights. This includes assessing the value of recruiting temporary HCWs, incorporating demand rate uncertainty into recruitment decision making (instead of using only its average), and delaying advertisement for permanent positions. We also conduct a case study using real data from a geriatric ward to demonstrate how our framework can be adopted to guide nurse recruitment decision making in a complex environment involving multiple types of resources.

2. Literature Review

Two main streams can be identified in the blended workforce literature. The first stream seeks the optimal mix of workforce using single-stage optimization models; see, e.g., Abraham (1988), Berman & Larson (1994), Jeang (1996), Bhandari et al. (2008), Harper et al. (2010), and Dong & Ibrahim (2020). However, single-stage models assume simultaneous recruitment of permanent and temporary HCWs, thus ignoring the asymmetry in demand information at the times of temporary and permanent recruitment. This shortcoming is addressed in the second stream of research, which focuses on two-stage optimization models.

Kao & Queyranne (1985) propose a two-stage optimization for budgeting workforce requirements. In the first stage of their model, the number of permanent nurses in each skill class for a budget cycle is determined. In the second stage, given the number of permanent nurses and the realized demand in each period, the numbers of overtime and agency nurses to meet shortages are identified. Indeed, similar models are proposed in the literature, as reported in the review by Qin et al. (2015). In particular, Pinker &

Larson (2003) propose a variation in which the second-stage decision is divided into two decisions that are made dynamically over two time intervals, one concerning the number of temporary workers and the other focusing on the amount of overtime work. A more recent study by Lu & Lu (2017) tests the results obtained from a two-stage model using empirical data to investigate the impact of mandatory overtime legislation on staffing ratios in nursing homes.

All two-stage models cited above ignore the dynamics of service delivery and assume a linear relationship between demand and the number of servers required to meet this demand. This amounts to nurse-to-patient ratio models in healthcare, which according to Yankovic & Green (2011) can lead to under- or over-staffing because factors such as the unit size and variability in service durations are not explicitly accounted for. This issue is addressed in the recent work of Hu et al. (2022), which is the closest to our research. Representing the dynamics of service delivery by an abandonment queueing model, they propose a two-stage framework 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). Using a Poisson mixture process for modeling demand, Hu et al. (2022) 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). They also propose near-optimal two-stage staffing rules minimizing the sum of staffing and performance costs. Hu et al. (2022) extend their model to allow for the demand rate prediction error in the second stage, and make several empirical adjustments to their staffing rules to facilitate implementation in an emergency department.

Similar to Hu et al. (2022), we adopt a Poisson mixture model for patient demand. As illustrated in Jongbloed & 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. (2022), and implies that a degree of demand uncertainty remains at the time of temporary recruitment, which matches the reality of service systems. Another similarity is that the dynamics of service delivery and the resulting system stochasticity are modelled explicitly.

Our study, however, differs from that of Hu et al. (2022) in several ways. First, Hu et al. (2022) (and other two-stage studies cited above) assume that the base staffing level can always be achieved, whereas we consider the uncertain nature of permanent recruitment, thus accounting for the possibility that some positions are not filled. In addition to making our models more realistic, this allows investigating the benefit of a lower demand rate uncertainty as a result of a later advertisement for permanent HCWs versus the associated risk of a shorter advertisement window. Second, Hu et al. (2022) focus on abandonment queues, whereas we consider delay queues. Delay queues are more appropriate for representing the services provided in inpatient or residential care settings where, once patients are admitted, their requests rarely leave the system. The same applies to diagnostic services in hospitals. Embedding delay queues in optimization models are, however, more challenging as a stability constraint is required. Third, Hu et al. (2022) derive staffing rules via an asymptotic approach that increases the system scale to infinity. This may lead to significant errors in small systems, as illustrated in Tables 2 and 3 in Hu et al. (2022). Our methodology addresses this problem by taking an exact approach which works equally well with small systems. This is an important feature given that the systems for residential or inpatient care are relatively small, as we illustrate in our case study. Fourth, our models capture the impact of the service time distribution on recruitment decisions, whereas the staffing rules in Hu et al. (2022) are obtained by assuming Exponential service times.

3. The Two-stage Framework

Consider a HUDP preceded by a permanent recruitment period of length t_e as illustrated in Figure 1. 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 §6, we explore the potential benefit/risk of a later advertisement.) The problem addressed here is a two-stage decision problem. The first-stage decision identifies the number of permanent full-time equivalent (FTE) positions to advertise at time t, denoted by $a \in \mathbb{R}^+$, and the second-stage decision determines the number of temporary FTEs to recruit at time t_e, denoted by $g \in \mathbb{R}^+$, where $\mathbb{R}^+$ is the set of non-negative real numbers. Note that the quantity of HCWs is measured in FTEs, which are non-negative real numbers. This is consistent with previous studies on blended workforce (e.g., Kao & Queyranne, 1985; Abraham, 1988), and also facilitates the derivation of analytical results.

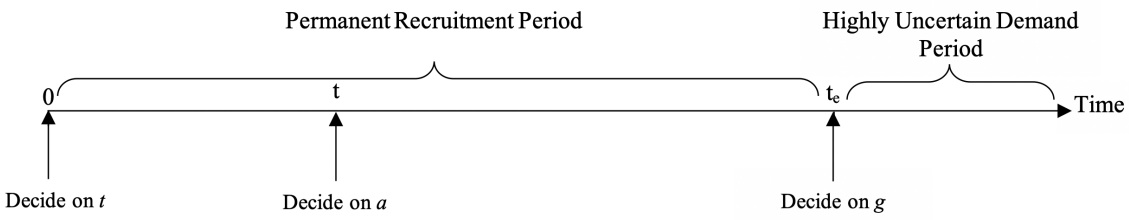


Figure 1: Schematic diagram of permanent and temporary recruitment decision making process

We start with the formulation of the second-stage problem. Suppose patients' requests arrive for HCW services according to a Poisson process with rate λ during the HUDP. The requests wait in a queue until they are served by a member of the pool of HCWs (permanent or 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. We model the service delivery in the HUDP by a generic delay queueing system. Assuming a steady-state is achieved, we denote by $l(\lambda, s)$ the corresponding system-size function evaluating the mean number of requests in the system given the demand rate is $\lambda > 0$ and the size of the HCW pool is $s \in \mathbb{R}^+$ with $s > \lambda$. Clearly, if the traffic intensity exceeds one; i.e., $s < \lambda$, there is no steady-state and this situation will be excluded from the analysis. Following Lu & Lu (2017), we assume that each permanent HCW must provide an additional $r_o \geq 0.0$ percentage of mandatory overtime work. The objective is to minimize costs, with the cost parameters defined as follows.

Let c_p, c_o, and c_t be the cost rates of permanent, mandatory overtime, and temporary work, respectively. Similar to Lu & Lu (2017), we assume that $c_p < c_o < c_t$. 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 $c_p = 1.0$.

We assume the provider has full knowledge of λ at time t_e. This can be achieved using a demand rate forecasting model such as the one proposed in Hu et al. (2021). The total number of permanent HCWs, denoted by p, is also known at time t_e. Given p and λ, the

second-stage problem is formulated as

$$v(\lambda, p) = \min_{g}\{u(\lambda, p, g) : g \in \mathbb{R}^+, g > \lambda - p(1 + r_o)\}, \tag{1}$$

where

$$u(\lambda, p, g) = p(1 + r_o c_o) + g c_t + l(\lambda, p(1 + r_o) + g)c_w, \tag{2}$$

and the second constraint ensures the stability of the system (recall that service rate is set to one). In the objective function $u(\lambda, p, g)$, the first two terms yield the total staffing cost and the last term gives the performance cost. We denote the optimal solution to problem (1) by $g^*(\lambda, p)$.

To formulate the first-stage problem, note that 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 $d \in [0, t_e)$. The distribution of this rate (and thus its mean) can be estimated from historical data as we shall illustrate in §7. 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 & 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 in the system at time t who are expected to be available during the HUDP, there will be $n+\min\{Q_t, a\}$ permanent HCWs at time t_e. The first-stage problem is therefore formulated as

$$m(n) = \min_{a}\{\mathbb{E}[v\left(\Lambda_t, n + \min\{Q_t, a\}\right)] : a \in \mathbb{R}^+\}, \tag{3}$$

where $v(\lambda, p)$ is evaluated through the second-stage problem given in (1). 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) by $a^*(n)$.

In order to characterize the optimal solutions to the first- and second-stage problems, given in Equations (3) and (1), respectively, we need to make the following set of assumptions concerning the generic 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 §4, 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 §4, we formally prove these properties for three common queueing models.

Given Assumption 1, we propose

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

$$g^*(\lambda, p) = \begin{cases} 0 & \text{if } \lambda \leq \tilde{\lambda}(p), \\ \tilde{g}(\lambda, p) & \text{if } \lambda > \tilde{\lambda}(p), \end{cases} \tag{4}$$

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

$$\phi_p(x) \triangleq c_t + c_w \frac{\partial l(x, s)}{\partial s}\Big|_{s=p(1+r_o)}, \tag{5}$$

in the interval $(0, p(1 + r_o))$ when $p > 0$, and $\tilde{\lambda}(0) = 0$. $\tilde{g}(\lambda, p)$ in (4) *is the unique root of function*

$$\theta_{\lambda,p}(g) \triangleq c_t + c_w \frac{\partial l(\lambda, s)}{\partial s}\Big|_{s=p(1+r_o)+g}, \tag{6}$$

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

The formal proof of Proposition 1, which is based on the Karush–Kuhn–Tucker (KKT) approach to nonlinear optimization (Kuhn & Tucker, 1951), is given in the e-companion to this paper (It also includes other proofs and a summary of notations.) Below, we provide the intuition behind this proof.

Function $\theta_{\lambda,p}(g)$ given in (6) is in fact the derivative of the objective function provided in (2) with respect to g. By assumption A(iv), the objective function is convex in g and so its minimum occurs at a point $\tilde{g}(\lambda, p)$ where the derivative is equal to zero. Hence, $g^*(\lambda, p) = 0$ when $\tilde{g}(\lambda, p) \leq 0$, and $g^*(\lambda, p) = \tilde{g}(\lambda, p)$ otherwise (a negative $\tilde{g}(\lambda, p)$ implies that too many HCWs are already in the system.) To identify the values of λ for which $\tilde{g}(\lambda, p) \leq 0$, note that $\theta_{\lambda,p}(g)$ is strictly increasing in g and strictly decreasing in λ by assumptions A(iv) and A(v), respectively. As such, $\tilde{g}(\lambda, p)$ will be negative for sufficiently small (and positive) values of λ. To find these values, we set $g = 0$ in $\theta_{x,p}(g)$ to obtain $\phi_p(x)$ as given in Equation (5). Let $\tilde{\lambda}(p)$ be the solution of $\phi_p(x) = 0$. Since $\theta_{\lambda,p}(g)$ is decreasing in λ, we will then have $\theta_{\lambda,p}(0) \geq \theta_{\tilde{\lambda}(p),p}(0) = \phi_p(\tilde{\lambda}(p)) = 0$ for $\lambda \leq \tilde{\lambda}(p)$. This implies that $\tilde{g}(\lambda, p) \leq 0$ when $\lambda \leq \tilde{\lambda}(p)$, and $\tilde{g}(\lambda, p) > 0$ otherwise.

Algorithm 1 in the e-companion 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 (5)) in the interval $(0, p(1 + r_o))$. As proved in the e-companion, $\phi_p(x)$ 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. Algorithm 1 also needs to evaluate the unique root of function $\theta_{\lambda,p}(g)$ (given in Equation (6)) in the interval $((\lambda - p(1 + r_o))^+, \infty)$. As shown in the e-companion, $\theta_{\lambda,p}(g)$ 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. Before proceeding to the first-stage problem, below 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.*

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) 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

$$\mathbb{E}[v\left(\Lambda_t, n + \min\{Q_t, a\}\right)] = \int_0^{\lambda_u} \int_0^{q_u} v(\lambda, n + \min\{q, a\}) f_t(q) h_t(\lambda) dq\, d\lambda. \tag{7}$$

Evaluating Equation (7) 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, q), which requires evaluating two pdf functions and the optimal second-stage cost $v(\lambda, n + \min\{q, a\})$, which in turn requires evaluating $g^*(\lambda, n + \min\{q, a\})$. The computation time would therefore be significant, making this approach impractical. Instead, we consider the structural properties of the objective function as elaborated in the e-companion. In short, let us denote by $\psi_n(a)$ the derivative of the objective function of the first-stage problem 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

$$\psi_n(a) = 1 + r_o c_o + c_w(1 + r_o) \int_0^{\tilde{\lambda}(n+a)} \frac{\partial l(\lambda, s)}{\partial s}\Big|_{s=(n+a)(1+r_o)} h_t(\lambda) d\lambda \\ - c_t(1 + r_o)\left(1 - H_t\left(\tilde{\lambda}(n + a)\right)\right), \tag{8}$$

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)*,*

$$a^*(n) = \begin{cases} 0.0 & \text{if } \psi_n(0) \geq 0.0, \\ \min\{\tilde{a}(n), q_u\}, & \text{otherwise}, \end{cases} \tag{9}$$

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

Proposition 2 leads to two important corollaries:

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

Corollary 3. *$a^*(n)$ is independent of 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 $a^*(n) = (\tilde{a}(0.0) - n)^+$.*

Corollary 2 implies that it is never cost-effective to serve patients with only temporary HCWs. The first part of Corollary 3 is due to the implicit assumption that f_t is independent of the value of a, i.e., advertising a larger or smaller number of positions does

not change the number of applications. Under this assumption, it is intuitive that the provider must always advertise for as many permanent positions as required (depending on the cost parameters and the distribution of Λ_t) regardless of the potential number of applications. Knowing that, for example, more or fewer applications are likely to arrive in a permanent recruitment period should not influence this decision. The second part of 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 in the e-companion 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 e-companion, $\psi_n(a)$ 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 $\psi_n(a_u) > 0$. 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 intervals, and does not require evaluating the optimal second-stage decision and its cost. The computations 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_t and c_w, and decreases with c_o and n.*

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

4. Special Cases

In this section, we consider three queueing models for the system serving patients' requests, and show that the corresponding system-size functions meet the properties in Assumption 1, and so can be applied with Propositions 1 and 2. We also explain the benefits of using each of the three queueing models, and provide further analytical results for two of them.

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 & Reiman (1998) and Anily & 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,

$$l(\lambda, s) = \frac{\lambda}{(s - \lambda)}, \tag{10}$$

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,*

$$\tilde{\lambda}(p) = p(1 + r_o) + \frac{c_w - \sqrt{4c_t c_w p(1 + r_o) + c_w^2}}{2c_t}, \tag{11}$$

$$\tilde{g}(\lambda, p) = \lambda + \sqrt{\frac{c_w \lambda}{c_t}} - p(1 + r_o), \tag{12}$$

and

$$v(\lambda, p) = \begin{cases} p(1 + r_o c_o) + \dfrac{\lambda c_w}{p(1 + r_o) - \lambda} & \text{if } \lambda \leq \tilde{\lambda}(p)\ , \\ \left(-c_t(1 + r_o) + 1 + r_o c_o\right) p + c_t \lambda + 2\sqrt{c_t c_w \lambda} & \text{if } \lambda > \tilde{\lambda}(p). \end{cases} \tag{13}$$

Equation (12) 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 & 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 $\sqrt{c_w/c_t}$ in our formula. An adjustment is also made to account for the number of permanent workers. The expression given in (12) replaces the numerical procedure for obtaining $\tilde{g}(\lambda, p)$ in Algorithm 1, and the expression given in (11) 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

$$l(\lambda, s) = \frac{1+\tau^2}{2}\frac{\lambda^2}{s(s-\lambda)} + \frac{\lambda}{s}, \tag{14}$$

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 (14) meets the properties given in Assumption 1.*

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.*

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

$$l(\lambda, s) = \frac{\lambda C(\lambda, s)}{s-\lambda} + \lambda, \tag{15}$$

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

$$C(\lambda, s) = \left(\int_0^\infty \lambda e^{-\lambda x}(1+x)^{s-1} x dx\right)^{-1}, \tag{16}$$

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

Proposition 5. *The system-size function $l(\lambda, s)$ given in Equation (15) meets the properties given in Assumption 1.*

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.

5. 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 & Koole (2001), and is verified empirically in our case study in §7. For Q_t, Pinker & 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.

5.1. 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

$$m_{single}(n) = \min_{a} \Big\{ \mathbb{E}\left[(n + \min\{Q_t, a\})(1 + r_o c_o) + l(\Lambda_t, (n + \min\{Q_t, a\})(1 + r_o))c_w \,|\, \mathbb{S}\right] : \mathbb{P}(\mathbb{S}) \geq \gamma;\, a \in \mathbb{R}^+ \Big\}, \quad (17)$$

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 (17), 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_t, 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 (7). The optimal solution to (17) is estimated by complete enumeration over values of $a \in [0.0, 0.1, \ldots, 5\xi]$. The results are plotted in Figure 2 for $\gamma = 0.95$.

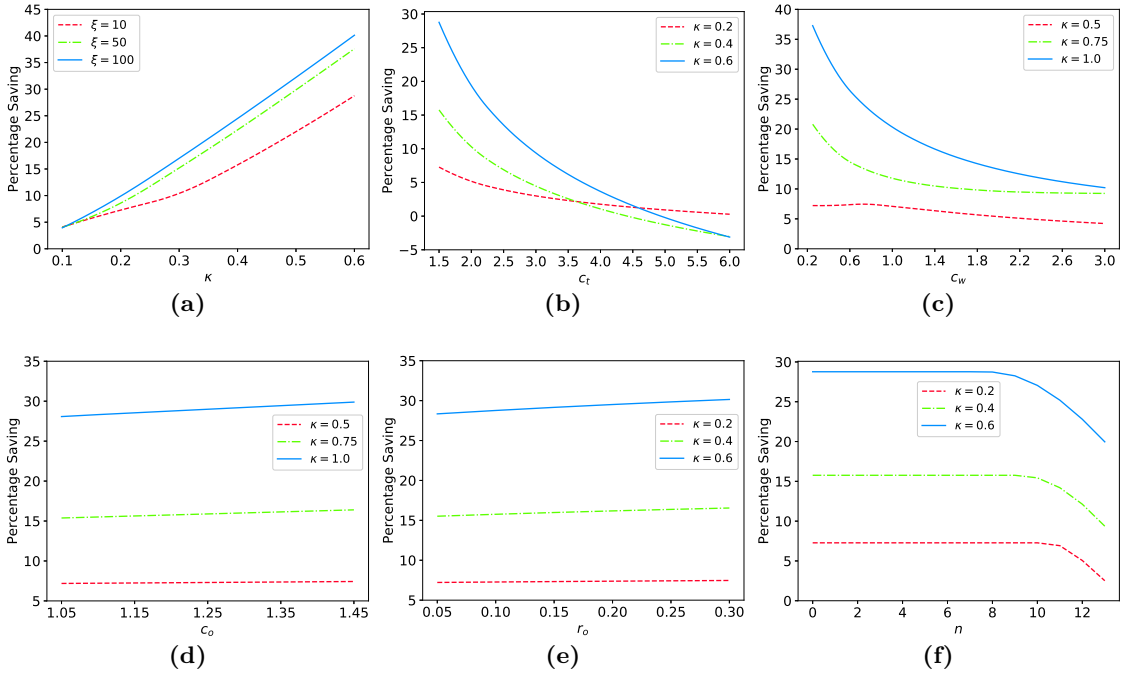


Figure 2: The savings of our model as compared to the single-stage model using an $M/M/s$ queue. The parameters not given in the plots are $\xi = 10.0$, $c_t = 1.5$, $c_w = 0.5$, $c_o = 1.2$, $r_o = 0.1$, $n = 0$, $\mu_r = 10\xi$, and $\kappa_r = 0.5$.

The plots in panel (a) of Figure 2 suggest that savings from our model typically increase with the system scale and the level of uncertainty in demand rate, exceeding 10.0% 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 $c_t \geq 4.5$ and $c_t \geq 5.0$ 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_t is extremely high (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. The plots in panels (b) to (f) are obtained with $\xi = 10.0$, but we observed similar trends for $\xi = 50.0$ and $\xi = 100.0$.

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 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.

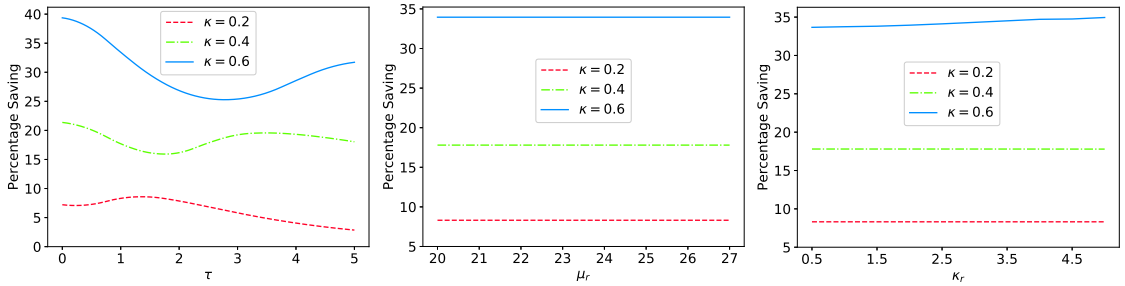


Figure 3: The savings of our model as compared to the single-stage model using an $M/G/1$ queue. The parameters are $\xi = 10.0$, $c_t = 1.5$, $c_w = 0.5$, $c_o = 1.2$, $r_o = 0.1$, $n = 0$, and (a) $\mu_r = 10\xi$, $\kappa_r = 0.5$, (b) $\tau = 1.0$, $\kappa_r = 0.5$, and (c) $\tau = 1.0$ and $\mu_r = 20.0$.

5.2. Comparison with a Two-Stage Model with No Demand Rate Uncertainty

We consider a two-stage optimization framework similar to §3, 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

$$m(n) = \min_a \{\mathbb{E}[v\,(\xi, n + \min\{Q_t, a\})] : a \in \mathbb{R}^+\}, \tag{18}$$

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* (18) *is obtained from Proposition 2 with* $\psi_n(a)$ *simplified as*

$$\psi_n(a) = 1 + r_o c_o + \begin{cases} -c_t(1 + r_o), & \tilde{\lambda}(a + n) < \xi, \\ c_w(1 + r_o)\dfrac{\partial l(\xi, s)}{\partial s}\Big|_{s=(n+a)(1+r_o)}, & \tilde{\lambda}(a + n) \geq \xi. \end{cases} \tag{19}$$

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

$$\tilde{a}(n) = \frac{\sqrt{\frac{\xi c_w(1+r_o)}{1+r_o c_o}} + \xi}{1+r_o} - n. \tag{20}$$

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_t, c_w, c_o, r_o, and n on savings. For the same reason as in §5.1, 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 §5.1. 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 (7). The results are plotted in Figure 4.

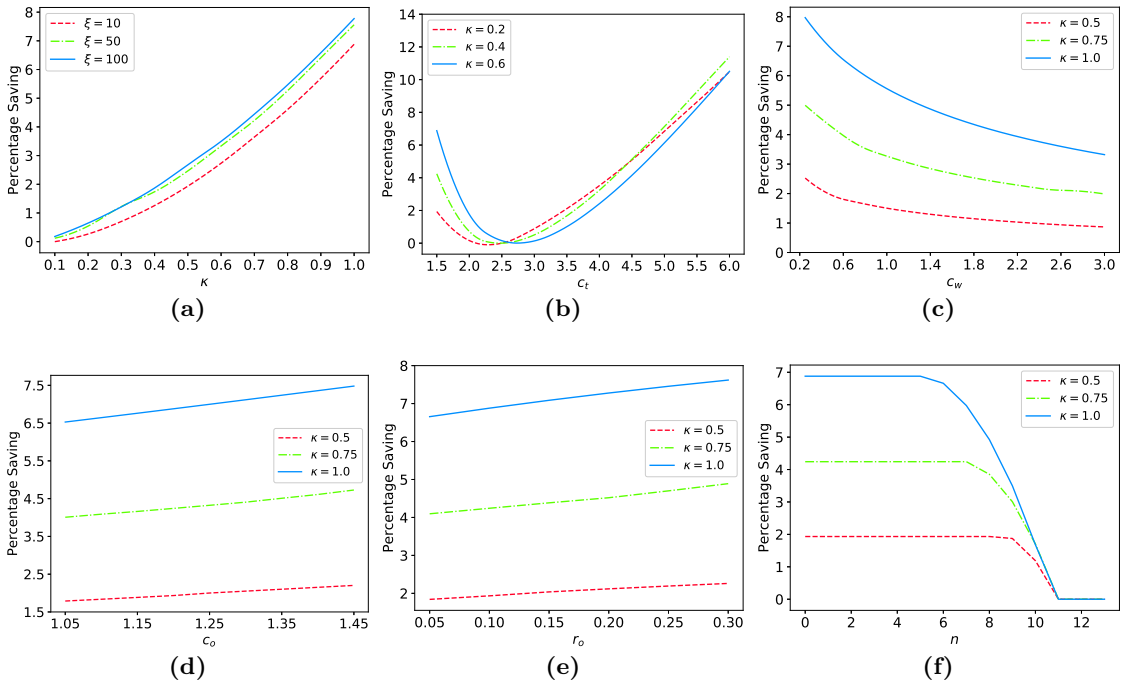


Figure 4: The savings of our model as compared to the model with no demand rate uncertainty using an $M/M/s$ queue. The parameters not given in the plots are $\xi = 10.0$, $c_t = 1.5$, $c_w = 0.5$, $c_o = 1.2$, $r_o = 0.1$, $n = 0$, $\mu_r = 10\xi$, and $\kappa_r = 0.5$.

Panel (a) of Figure 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 uncertainty, i.e., $\kappa \approx 0.5$, and a medium system, i.e., $\xi \approx 50.0$. Panel (b) of Figure 4 suggests that savings show a non-monotone behaviour with respect to c_t, initially decreasing

but then increasing. 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 $n \geq 11.0$. The plots in panels (b) to (f) are obtained with $\xi = 10.0$, but we observed similar trends for $\xi = 50.0$ and $\xi = 100.0$.

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 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.

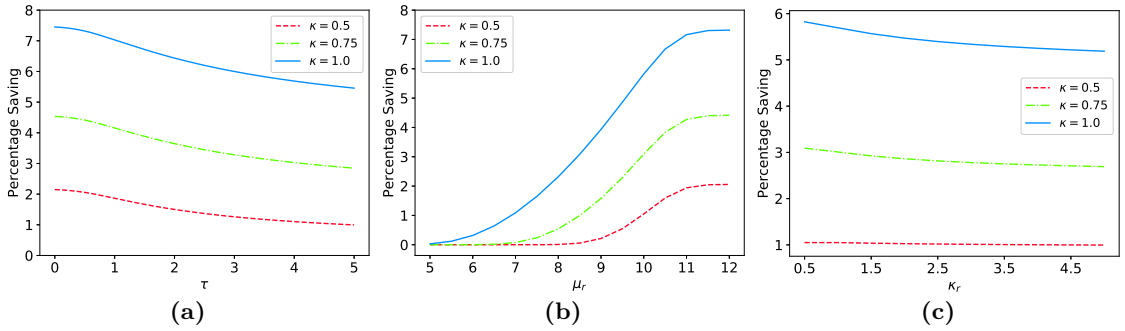


Figure 5: The savings of our model as compared to the model with no demand rate uncertainty using an $M/G/1$ queue. The parameters are $\xi = 10.0$, $c_t = 1.5$, $c_w = 0.5$, $c_o = 1.2$, $r_o = 0.1$, $n = 0$, and (a) $\mu_r = 10\xi$, $\kappa_r = 0.5$, (b) $\tau = 1.0$, $\kappa_r = 0.5$, and (c) $\tau = 1.0$ 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_t 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.

6. Delaying Advertisement

In §3, 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 & 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. 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 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 $m(n, Q_t, \Lambda_t) \leq m(n, Q_{t'}, \Lambda_t)$.*

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

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

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

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

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 (11) indicates that λ_u should be less than 4.45 for condition (a) to apply when $n = 5.0, c_t = 2.0, c_w = 0.5$

and $r_o = 0.1$. For smaller values of n or c_t, 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 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)).

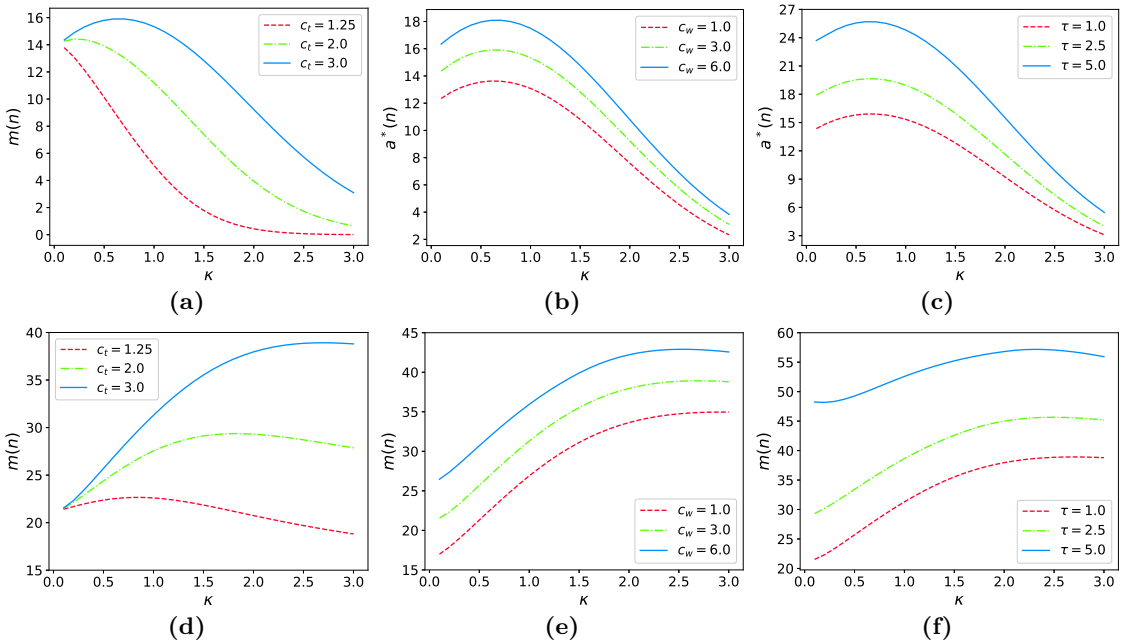
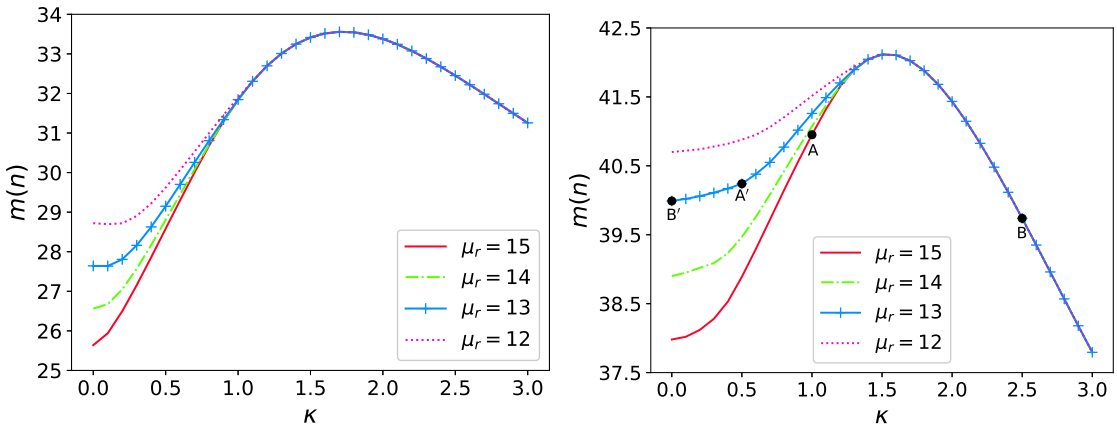


Figure 6: Optimal number of permanent positions (top panel) and the corresponding cost (bottom panel) as a function of demand rate uncertainty. The parameters not given in the plots are $c_t = 3.0$, $c_w = 3.0$, $c_o = 1.2$, $r_o = 0.1$, $n = 0.0$, $\tau = 1.0$, and $\xi = 10.0$. For cost evaluations, Q_t is assumed to follow a LogNormal distribution with $\mu_r = 15.0$ and $\kappa_r = 0.3$.

More specifically, the plots at the top of Figure 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

accurate information on demand rate) so as to avoid over-staffing. The plots also show that the value of κ threshold increases with c_t, but is insensitive to c_w and τ. Similarly, the plots in the bottom panels of Figure 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 6 that the κ threshold for $m(n)$ (above which the decreasing trend occurs) increases with c_t, 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 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 7: Optimal cost as a function of κ for different values of μ_r. The other parameters are $\xi = 10.0$, $c_t = 2.0$, $c_o = 1.2$, $r_o = 0.1$, $n = 0$, and $\kappa_r = 0.3$.

Figure 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 6), which implies that the number of qualified applications will be less relevant. We further observe in Figure 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 (a) in Figure 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.

7. 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 8 depicts the empirical 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 our framework can be used 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. (2022) 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

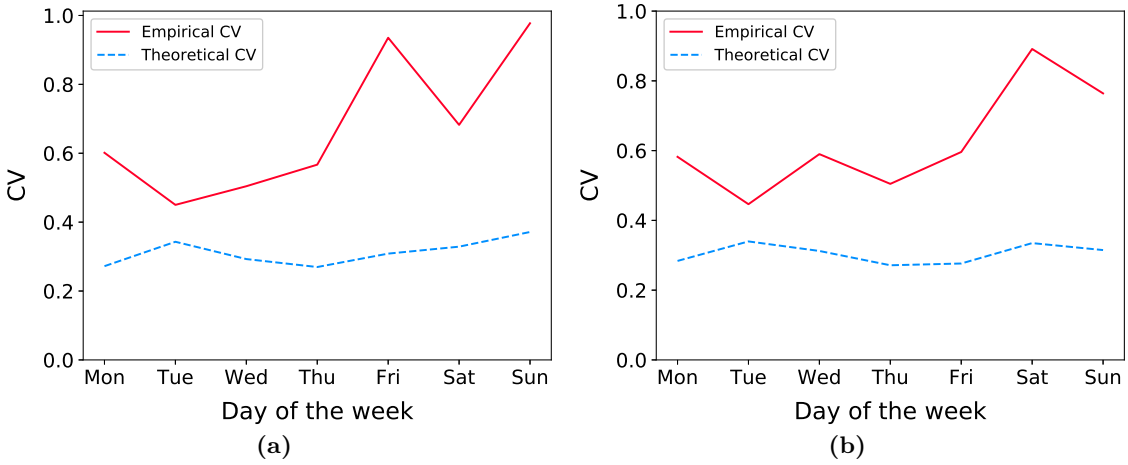


Figure 8: Empirical CV and theoretical CV under Poisson assumption for daily admissions in (**a**) December and (**b**) January.

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 so-called "trolley wait" in emergency departments (Abo-Hamad & 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 §7.1, 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 §7.2, 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 §7.3, 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.

7.1. The MRMS Model

We first develop a detailed discrete-event simulation model which includes all dynamics of bed and nursing queueing systems. Following Yankovic & Green (2011), the model considers two types of resources, beds and nurses, each of which has its own separate

queue. We use the superscripts (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 & 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

$$m_{sim}(n) = \min_a \left\{ \mathbb{E}\left[v_{sim}\left(\Lambda_t^{(b)}, n + \min\{Q_t, a\} \right) \right] : a = 0, \cdots, a_{max} \right\}, \tag{21}$$

and

$$v_{sim}(\lambda, p) = \min_g \left\{ p(1 + r_o c_o) + g c_t + l^{(n)}\left(\lambda^{(b)}, p(1 + r_o) + g \right) c_w : \right.$$
$$\left. g = \lceil \lambda - p(1 + r_o) \rceil, \cdots, g_{max} \right\}, \tag{22}$$

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 (21) 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 & 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 & 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 §5.2, 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 & Segesten (2001) and Dochterman & Bulechek (2004). Following Yankovic & 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 & 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 & 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 $\kappa^{(b)} \in \{0.58, 1.0\}$, $c_t \in \{2, 3\}$, $c_w \in \{1.5, 3.0\}$, $c_o \in \{1.5, 1.7\}$, $n \in \{0, 1\}$, 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_t are consistent with the estimates provided in Lu & Lu (2017), and the values of c_w follow Hu et al. (2022). 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 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.

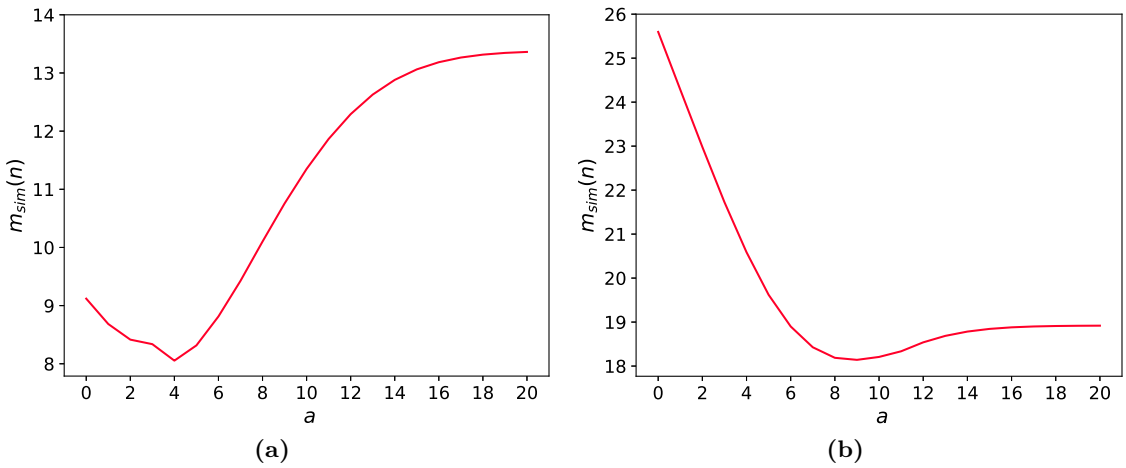


Figure 9: First-stage cost as a function of a for the scenario with (a) $c_t = 2.0, c_w = 1.5, c_o = 1.5, n = 1.0, r_o = 0.1, \mu_r = 12.0, \lambda^{(n)} = 0.4$, and $\kappa^{(b)} = 1.0$, and (b) $c_t = 3.0, c_w = 3.0, c_o = 1.5, n = 0.0, r_o = 0.05, \mu_r = 10, \lambda^{(n)} = 0.5$, and $\kappa^{(b)} = 0.58$.

Since there is no explicit form for the system-size function, $l^{(n)}(\lambda^{(b)}, s)$, it would be difficult to check if it follows the properties in Assumption 1. However, we can confirm from the results that $a^*_{sim}(n)$ obtained from complete enumeration and its corresponding cost $m_{sim}(n)$ satisfy the properties proved in §3. In particular, we observe that $a^*_{sim}(n)$ does not change with μ_r, as proved in Corollary 3, and that it shows an increasing (decreasing) trend with respect to c_t and c_w (c_o and n), as proved in Corollary 4. We also observe that $m_{sim}(n)$ increases with c_t, c_w and c_o, as proved in Corollary 5.

7.2. 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. (2022) 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 §7.1 with both SRSS and SRMS approximations. We then run the simulation model developed in §7.1 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 our experiments with $\xi^{(b)}$ equal to 8.64, 7.41, and 6.17 patients per day, which yield traffic intensities of 70.0%, 60.0%, and 50.0% for the bed queueing system, to assess the SRMS and SRSS approximations in less congested departments. The number of matching scenarios and average difference in cost as compared to the MRMS model are presented in Table 1 for these traffic intensities including the original 80.0%. We observe that for both SRSS and SRMS approximations, the accuracy typically reduces as traffic intensity decreases. The only exception occurs at 70.0% traffic intensity, where the number of matching scenarios of SRMS approximation remains the same as that of 80.0% and the corresponding average difference in cost reduces. We also observe in Table 1 that SRMS is more accurate than SRSS, especially at lower values of traffic intensity. Overall, both SRSS and SRMS approximations are reasonably accurate as long as the traffic intensity is not too low.

Table 1: Summary of the comparison between MRMS model with SRSS and SRMS approximation models.

Traffic intensity (%)	SRSS		SRMS	
	Average difference in cost (%)	# of scenarios with matching a^*	Average difference in cost (%)	# of scenarios with matching a^*
83	0.97	108	0.97	112
70	1.33	80	0.8	112
60	1.92	32	0.95	72
50	2.47	28	1.083	64

7.3. Delaying Advertisement

As discussed in §7.3, 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 §7.2, it is used in the analysis that follows.

We consider the scenario with $\lambda^{(n)} = 0.4, \mu_r = 10.0, c_t = 2.0, c_w = 3.0, c_o = 1.5, n = 0.0$, 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 10 for different levels of temporary cost rate, c_t, and different levels of current demand rate uncertainty, κ.

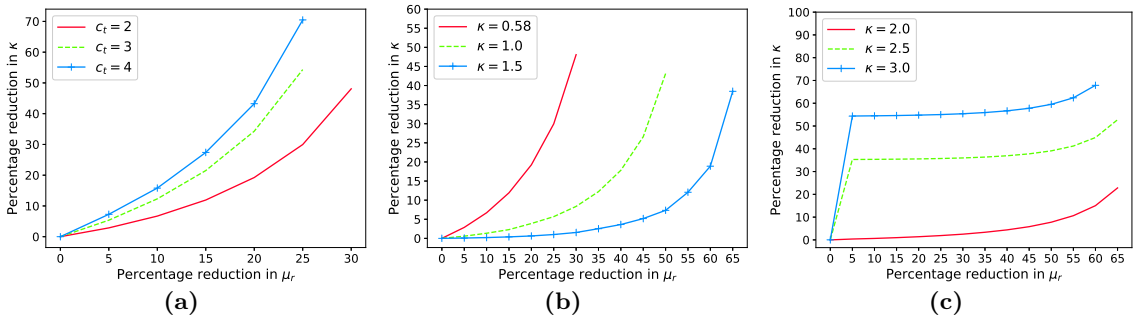


Figure 10: The reduction required in demand rate uncertainty as a function of reduction in mean application numbers. For panel (a), $\kappa = 0.58$, and for panels (b) and (c), $c_t = 2.0$.

The plots in panel (a) of Figure 10 show that a larger reduction in demand rate uncertainty is needed to make a later advertisement beneficial as c_t grows. They also imply that reductions above 30% (25%) in mean application numbers for $c_t = 2.0$ ($c_t = 3.0$ and $c_t = 4.0$) 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 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 10 provides valuable insights on how and when delaying advertisement may create value to the provider.

8. Conclusions & Future Research

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 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.

Our framework is based on the assumption that all patient requests must be served. This makes it suitable for settings such as inpatient or residential care. Our solution approach is exact and works well irrespective of the system scale. This is an important feature since, as illustrated in our case study, the systems representing nursing care are typically small, rendering asymptotic large-scale approximations inaccurate. Another important feature of our framework is that it accounts for the uncertainty in the permanent recruitment process. This enabled us to investigate the timing of permanent advertisement, which, to the best of our knowledge, has not been addressed in the literature.

By combining analytical results with numerical experiments, we derived the following managerial insights. First, temporary staffing alone is never cost-effective. At the same

time, unless the cost of temporary staffing is extremely high, there is value in recruiting temporary workers. This value is likely to increase with the system scale and decrease with the waiting cost. This result complements the findings from Hu et al. (2022) by showing that temporary staffing is valuable even when demand rate uncertainty is very low.

Second, when demand rate uncertainty is moderate to high, there is value in obtaining the demand rate distribution and incorporating it into the recruitment decision making. Otherwise, using only the average demand rate would suffice. We illustrated how this average can be applied to estimate the optimal number of permanent positions. Third, the optimal first- and second-stage decisions and their corresponding costs increase with service time variability. Fourth, the optimal number of permanent positions and the corresponding cost show a non-monotone behaviour with respect to demand rate uncertainty. In particular, both exhibit a decreasing trend when demand rate uncertainty exceeds 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.

Using real data, we illustrated how our framework can be applied for recruitment decision making in inpatient departments. In particular, we showed how a detailed simulation model of the department can be embedded within our two-stage optimization framework for estimating the optimal number of permanent nursing positions. This simulation model 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 highlighted that simple single-resource approximations based on our analytical models provide reliable and sufficiently accurate results with substantially less effort when traffic intensity is not too low. 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 expected reduction in mean application numbers.

Our proposed framework can be extended to capture cost functions involving other performance metrics as long as the specified properties are met. Although we have not considered time-varying demand, this is a future avenue for research that can benefit from our formulation and methodology. The long-term commitment of service providers to permanent workers may also be investigated by considering longer planning horizons.

References

Abo-Hamad, W., & Arisha, A. (2013). Simulation-based framework to improve patient experience in an emergency department. *European Journal of Operational Research*, *224*, 154–166.

Abraham, K. G. (1988). *Flexible Staffing Arrangements and Employers' Short-Term Adjustment Strategies*. Technical Report National Bureau of Economic Research Cambridge, MA.

Anily, S., & Haviv, M. (2010). Cooperation in service systems. *Operations Research*, *58*, 660–673.

Bae, S.-H., Mark, B., & Fried, B. (2010). Use of temporary nurses and nurse and patient safety outcomes in acute care hospital units. *Health Care Management Review*, *35*, 333–344.

Belzunce, F., Martínez-Riquelme, C., & Mulero, J. (2016). Chapter 2 - univariate stochastic orders. In F. Belzunce, C. Martínez-Riquelme, & J. Mulero (Eds.), *An Introduction to Stochastic Orders* (pp. 27–113). Academic Press.

Berman, O., & Larson, R. C. (1994). Determining optimal pool size of a temporary call-in work force. *European Journal of Operational Research*, *73*, 55–64.

Bhandari, A., Scheller-Wolf, A., & Harchol-Balter, M. (2008). An Exact and Efficient Algorithm for the Constrained Dynamic Operator Staffing Problem for Call Centers. *Management Science*, *54*, 339–353.

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

Dochterman, J. M., & Bulechek, G. M. (2004). *Nursing Interventions Classification (NIC)*. (4th ed.). St. Louis: Mosby.

Dong, J., & Ibrahim, R. (2020). Managing Supply in the On-Demand Economy: Flexible Workers, Full-Time Employees, or Both? *Operations Research*, *68*, 1238–1264.

Gross, D., Shortie, J. F., Thompson, J. M., & Harris, C. M. (2008). *Fundamentals of Queueing Theory*. (4th ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.

Halfin, S., & Whitt, W. (1981). Heavy-Traffic Limits for Queues with Many Exponential Servers. *Operations Research*, *29*, 567–588.

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

Hu, Y., Cato, K. D., Chan, C. W., Dong, J., Gavin, N., Rossetti, S. C., & Chang., B. P. (2021). Use of real-time information to predict future arrivals in the emergency department. Working Paper, Columbia Business School.

Hu, Y., Chan, C. W., & Dong, J. (2022). Prediction-Driven Surge Planning with Application in the Emergency Department. URL: `http://www.columbia.edu/~cc3179/SurgeStaffing_2021.pdf` [Accessed on 05-08-2022].

Jagers, A. A., & van Doorn, E. A. (1991). Convexity of functions which are generalizations of the Erlang loss function and the Erlang delay function. *SIAM Rev.*, *33*, 281–282.

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

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

Kao, E. P. C., & Queyranne, M. (1985). Budgeting Costs of Nursing in a Hospital. *Management Science*, *31*, 608–621.

Kesavan, S., Staats, B. R., & Gilland, W. (2014). Volume Flexibility in Services: The Costs and Benefits of Flexible Labor Resources. *Management Science*, *60*, 1884–1906.

Kuhn, H. W., & Tucker, A. W. (1951). Nonlinear Programming. In *Berkeley Symposium on Mathematical Statistics and Probability* (pp. 481–492). University of California Press.

Lu, S. F., & Lu, L. X. (2017). Do Mandatory Overtime Laws Improve Quality? Staffing Decisions and Operational Flexibility of Nursing Homes. *Management Science*, *63*, 3566–3585.

Lundgren, S., & Segesten, K. (2001). Nurses' use of time in a medical-surgical ward with all-RN staffing. *Journal of Nursing Management*, *9*, 13–20.

Maman, S. (2009). *Uncertainty in the Demand for Service: The Case of Call Centers and Emergency Departments*. Master thesis Senate of the Technion - Israel Institute of Technology. URL: `https://iew.technion.ac.il/serveng/References/Thesis_Shimrit.pdf` [Accessed on 05-08-2022].

Mandelbaum, A., & Reiman, M. I. (1998). On pooling in queueing networks. *Management Science*, *44*, 971–981.

National Audit Office (2006). *Improving the use of temporary nursing staff in NHS acute and foundation trusts*. Technical Report.

National Audit Office (2016). *Managing the supply of NHS clinical staff in England*. Technical Report February Department of Health.

NHS Improvement (2018). *NHS review of winter 2017/18*. Technical Re-

port September. URL: http://dickyricky.com/Medicine/Guidelines/Gov/2018NHSreviewofwinter.pdf [Accessed on 05-08-2022].

NHS Vacancy Statistics (2021). NHS Vacancy Statistics England April 2015 – March 2021 Experimental Statistics. URL: https://digital.nhs.uk/data-and-information/publications/statistical/nhs-vacancies-survey/april-2015---march-2021 [Accessed on 05-08-2022].

Pinker, E., & Tilson, V. (2013). The impact of technology on the labor procurement process. In *46th Hawaii International Conference on System Sciences* (pp. 4176–4185). IEEE.

Pinker, E. J., & Larson, R. C. (2003). Optimizing the use of contingent labor when demand is uncertain. *European Journal of Operational Research*, *144*, 39–55.

Qin, R., Nembhard, D. A., & Barnes, W. L. (2015). Workforce flexibility in operations management. *Surveys in Operations Research and Management Science*, *20*, 19–33.

Royal College of Nursing (2021). NHS conditions of employment. URL: https://www.rcn.org.uk/employment-and-pay/nhs-conditions-of-employment [Accessed on 05-08-2022].

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

Shaked, M., & Shanthikumar, J. G. (Eds.) (2007). *Stochastic Orders*. New York, NY: Springer New York.

The Kings Fund (2021). Key facts and figures about the NHS. URL: https://www.kingsfund.org.uk/audio-video/key-facts-figures-nhs [Accessed on 05-08-2022].

The Open University (2018). *Tackling the nursing shortage*. Technical Report May.

Thorlby, R., Fraser, C., & Gardner, T. (2020). Non-COVID-19 NHS care during the pandemic. URL: https://www.health.org.uk/news-and-comment/charts-and-infographics/non-covid-19-nhs-care-during-the-pandemic [Accessed on 05-08-2022].

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

Vovak, R. (2010). *Staffing trends at Maryland hospitals: FY 2000 to FY 2010*. Technical Report Maryland Hospital Association Elkridge.

West, M., Bailey, S., & Williams, E. (2020). *The courage of compassion: Supporting nurses and midwives to deliver high-quality care*. Technical Re-

port The King's Fund. URL: https://www.kingsfund.org.uk/publications/courage-compassion-supporting-nurses-midwives [Accessed on 28-02-2022].

Yankovic, N., & Green, L. V. (2011). Identifying Good Nursing Levels : A Queuing Approach. *Operations Research*, *59*, 942–955.

E-Companion for: A Framework for Optimal Recruitment of Temporary and Permanent Healthcare Workers in Highly Uncertain Environments

A. Proof of Proposition 1

The Lagrangean of the optimization problem in (1) is given by

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

where the constraint $g > \lambda - p(1 + r_o)$ is not included as it is always active, and so its multiplier is equal to zero. By property A(iv), it is easy to verify that the optimization problem is a convex programming, hence its optimal solution is the value g satisfying the Kuhn-Tucker conditions

$$\frac{\partial \mathcal{L}(\lambda, p, \beta; g)}{\partial g} = 0,$$
$$\beta \geq 0, \quad g \geq 0, \quad \beta g = 0.$$

This leads to the following scenarios:

(i) $\beta = 0$: we must have $g \geq 0$, $g > \lambda - p(1 + r_o)$, and

$$\frac{\partial \mathcal{L}(\lambda, p, \beta; g)}{\partial g} = c_t + c_w \frac{\partial l(\lambda, s)}{\partial s}\Big|_{s=p(1+r_o)+g} = 0. \tag{A.2}$$

(ii) $\beta > 0$: we must have $g = 0$, $\lambda < p(1 + r_o)$, and

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

which gives

$$\beta = c_t + c_w \frac{\partial l(\lambda, s)}{\partial s}\Big|_{s=p(1+r_o)}. \tag{A.3}$$

These scenarios imply that the optimal solution is $g^*(\lambda, p) = 0$ for values of λ and p which yield a positive value for β given in (A.3) when $\lambda < p(1 + r_o)$. For other values of λ and p, the value of g satisfying Equation (A.2) will be optimal as long as it is positive and larger

than $\lambda - p(1+r_o)$. To specify the optimal solution for different ranges of λ and p, we first assume that $p > 0$, and define the function

$$\phi_p(x) \triangleq c_t + c_w \frac{\partial l(x,s)}{\partial s}\Big|_{s=p(1+r_o)}, \tag{A.4}$$

for $x \in (0, p(1+r_o))$. We show that $\phi_p(x)$ has a unique root in the interval $(0, p(1+r_o))$: (i) by properties A(i) and A(v), $\dfrac{\partial l(x,s)}{\partial s}$ is continuous and strictly decreasing in x; (ii) by Lemma A.1, we have $\lim_{x\downarrow 0} \dfrac{\partial l(x,s)}{\partial s} = 0$, and so $\lim_{x\downarrow 0} \phi_p(x) = c_t$, which is always positive; and (iii) by Lemma A.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, there exists a unique solution to $\phi_p(x) = 0$, which we denote by $\tilde{\lambda}(p)$. Now, we consider three situations.

1. $\lambda \in (0, \tilde{\lambda}(p))$: we have $\phi_p(\lambda) = \beta > 0$, and also $\lambda < \tilde{\lambda}(p) < p(1+r_o)$. Hence, the conditions of scenario (ii) are met for $g = 0$, and so $g^*(\lambda, p) = 0$.

2. $\lambda = \tilde{\lambda}(p)$: we have $\phi_p(\tilde{\lambda}(p)) = 0$, and so the conditions of scenario (ii) are not met. However, defining the function

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

   on $g \in [0, \infty)$, we obtain $\theta_{\tilde{\lambda}(p),p}(0) = \phi_p(\tilde{\lambda}(p)) = 0$. This, along with the fact that $\lambda = \tilde{\lambda}(p) < p(1+r_o)$, implies that conditions of scenario (i) are met for $g = 0$, and so $g^*(\tilde{\lambda}(p), p) = 0$.

3. $\lambda \in (\tilde{\lambda}(p), \infty)$: we know that $\phi_p(\lambda)$ is negative for $\lambda \in (\tilde{\lambda}(p), p(1+r_o))$, and undefined for $\lambda \in [p(1+r_o), \infty)$, and so the conditions of scenario (ii) are not met. However, we show that a $\tilde{g}(\lambda, p) \in ((\lambda - p(1+r_o))^+, \infty)$ can be found satisfying $\theta_{\lambda,p}(g) = 0$: (i) by properties A(i) and A(iv), $\theta_{\lambda,p}(g)$ is continuous and strictly increasing in g; (ii) for $\lambda \in (\tilde{\lambda}(p), p(1+r_o))$, we have $\lim_{g\downarrow 0} \theta_{\lambda,p}(g) = \phi_p(\lambda)$, which is negative ; (iii) for $\lambda \in [p(1+r_o), \infty)$, $\lim_{s\downarrow\lambda} \dfrac{\partial l(\lambda,s)}{\partial s} = -\infty$ by Lemma A.2, and so $\lim_{g\downarrow\lambda - p(1+r_o)} \theta_{\lambda,p}(g) = -\infty$; (iv) from (ii) and (iii), we conclude that $\lim_{g\downarrow(\lambda-p(1+r_o))^+} \theta_{\lambda,p}(g)$ is always negative; and (iv) by Lemma A.3, $\lim_{s\to\infty} \dfrac{\partial l(\lambda,s)}{\partial s} = 0$, and so $\lim_{g\to\infty} \theta_{\lambda,p}(g) = c_t$, which is always positive. As a result, 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 $p = 0$, the conditions of scenario (ii) cannot be met. However, following a similar

argument to that of the situation 3, we can find a value $\tilde{g}(\lambda, 0) \in (\lambda, \infty)$ satisfying the equation $\theta_{\lambda,0}(\tilde{g}(\lambda, 0)) = 0$, and consequently the conditions of scenario (i). Therefore $g^*(\lambda, 0) = \tilde{g}(\lambda, 0)$, and the proof is complete. □

Lemma A.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

$$\left| \frac{\partial l(\lambda, s)}{\partial s} \right| = -\frac{\partial l(\lambda, s)}{\partial s} < \epsilon, \tag{A.5}$$

when $0 < \lambda < \delta$, where the equality in equation 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

$$\frac{l(\lambda, s + h) - l(\lambda, s)}{h} > -\epsilon. \tag{A.6}$$

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

Lemma A.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. Then, by 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 A.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 $\lim_{s \to \infty} \dfrac{\partial l(\lambda, s)}{\partial s} = 0$. □

B. Algorithm 1

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

Require: $l(x, y)$, λ, p, c_t, 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

C. Proof of Corollary 1

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 C.1, proves that $g^*(\lambda, p)$ is decreasing in p. □

Lemma C.1. *$\tilde{\lambda}(p)$ 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. □

D. Proof of Proposition 2

Denoting the expected value in the first-stage problem given in (3) 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

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

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

$$\begin{aligned}\frac{\partial v(\lambda, n+a)}{\partial a} &= \frac{\partial}{\partial a}\mathcal{L}(\lambda, n+a, \beta; g)\big|_{g=g^*(\lambda,n+a),\beta=\beta^*(\lambda,n+a)} \\ &= 1 + r_o c_o + c_w(1+r_o)\frac{\partial l(\lambda,s)}{\partial s}\Big|_{s=(n+a)(1+r_o)+g^*(\lambda,n+a)}.\end{aligned} \tag{D.2}$$

where $\mathcal{L}(\lambda, n+a, \beta; g)$ is the Lagrangean of the second stage problem given in Equation (A.1) with $p = n+a$, and $g^*(\lambda, n+a)$ and $\beta^*(\lambda, n+a)$ are the corresponding optimal solution and Lagrange multiplier, respectively. Taking the expectation of the above expression, it follows that

$$\mathbb{E}\left[\frac{\partial v(\Lambda_t, n+a)}{\partial a}\right] = 1 + r_o c_o + c_w(1+r_o)\mathbb{E}\left[\frac{\partial l(\Lambda_t, s)}{\partial s}\Big|_{s=(n+a)(1+r_o)+g^*(\Lambda_t,n+a)}\right]. \tag{D.3}$$

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

$$\begin{aligned}\mathbb{E}\left[\frac{\partial v(\Lambda_t, n+a)}{\partial a}\right] &= 1 + r_o c_o + c_w(1+r_o)\left(\mathbb{E}\left[\frac{\partial l(\Lambda_t, s)}{\partial s}\Big|_{s=(n+a)(1+r_o)}, \Lambda_t \leq \tilde{\lambda}(n+a)\right]\right. \\ &\qquad \left. + \mathbb{E}\left[\frac{\partial l(\Lambda_t, s)}{\partial s}\Big|_{s=(n+a)(1+r_o)+\tilde{g}(\Lambda_t, n+a)}, \Lambda_t > \tilde{\lambda}(n+a)\right]\right) \\ &= 1 + r_o c_o + c_w(1+r_o)\left(\mathbb{E}\left[\frac{\partial l(\Lambda_t, s)}{\partial s}\Big|_{s=(n+a)(1+r_o)}, \Lambda_t \leq \tilde{\lambda}(n+a)\right]\right. \\ &\qquad \left. + \mathbb{E}\left[-\frac{c_t}{c_w}, \Lambda_t > \tilde{\lambda}(n+a)\right]\right) \\ = 1 + r_o c_o + c_w(1+r_o) &\int_0^{\tilde{\lambda}(n+a)} \frac{\partial l(\lambda, s)}{\partial s}\Big|_{s=(n+a)(1+r_o)} h_t(\lambda) d\lambda - c_t(1+r_o)\int_{\tilde{\lambda}(n+a)}^{\infty} h_t(\lambda) d\lambda,\end{aligned} \tag{D.4}$$

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 (6) 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

$$\frac{\partial y(n,a)}{\partial a} = \psi_n(a)(1 - F_t(a)). \tag{D.5}$$

Differentiating the equation above with respect to a, we obtain

$$\frac{\partial^2 y(n,a)}{\partial a^2} = \frac{\partial \psi_n(a)}{\partial a}(1 - F_t(a)) - f_t(a)\psi_n(a),$$

which shows that the objective function is not necessarily convex. However, its minimum can still be found as follows.

By Lemma D.2, $\psi_n(a)$ is a continuous and strictly increasing function of a. Further, by Lemma D.1, $\lim_{a\to\infty} \psi_n(a)$ is always positive. Hence, if $\psi_n(0) < 0$, there exists a unique solution to $\psi_n(a) = 0$, which we denote by $\tilde{a}(n)$. $\psi_n(a)$ will then be negative (positive) for $a \in [0, \tilde{a}(n))$ ($a \in (\tilde{a}(n), \infty)$). On the other hand, if $\psi_n(0) \geq 0$, then $\psi_n(a) > 0$ for all $a > 0$.

First, suppose $\psi_n(0) \geq 0$. Then $\psi_n(a)(1 - F_t(a))$ will be non-negative for all a, and so $y(n,a)$ will be an increasing function of a. This implies that the minimum of $y(n,a)$ occurs at $a = 0$.

Next, suppose $\psi_n(0) < 0$. Two situations may arise: $\tilde{a}(n) < q_u$ or $\tilde{a}(n) \geq q_u$. If $\tilde{a}(n) < q_u$, $\psi_n(a)(1 - F_t(a))$ will be non-positive for $a < \tilde{a}(n)$, and non-negative for $a > \tilde{a}(n)$, thus implying that the minimum of $y(n,a)$ occurs at $a = \tilde{a}(n)$. If $\tilde{a}(n) \geq q_u$, $\psi_n(a)(1 - F_t(a))$ will be non-positive for $a < q_u$, and zero for $a \geq q_u$, implying 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 (9).

□

Lemma D.1. $\lim_{a\to\infty} \psi_n(a)$ *is positive.*

Proof. By Lemma A.3,$\lim_{s\to\infty} \dfrac{\partial l(\lambda,s)}{\partial s} = 0$. From Lemma C.1, 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 D.2. $\psi_n(a)$ *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

$$\begin{aligned}
\frac{\partial \psi_n(a)}{\partial a} &= c_w(1+r_o)\frac{\partial \tilde{\lambda}(n+a)}{\partial a}\frac{\partial l(\tilde{\lambda}(n+a),s)}{\partial s}\Big|_{s=(a+n)(1+r_o)} h_t(\tilde{\lambda}(n+a)) \\
&\quad + c_w(1+r_o)^2 \int_0^{\tilde{\lambda}(n+a)} \frac{\partial^2 l(\lambda,s)}{\partial s^2}\Big|_{s=(a+n)(1+r_o)} h_t(\lambda)d\lambda + c_t(1+r_o)\frac{\partial \tilde{\lambda}(n+a)}{\partial a} h_t(\tilde{\lambda}(n+a)) \\
&= (1+r_o)\frac{\partial \tilde{\lambda}(n+a)}{\partial a} h_t(\tilde{\lambda}(n+a)) \left(c_t + c_w \frac{\partial l(\tilde{\lambda}(n+a),s)}{\partial s}\Big|_{s=(a+n)(1+r_o)} \right) \\
&\quad + c_w(1+r_o)^2 \int_0^{\tilde{\lambda}(n+a)} \frac{\partial^2 l(\lambda,s)}{\partial s^2}\Big|_{s=(a+n)(1+r_o)} h_t(\lambda)d\lambda \\
&= c_w(1+r_o)^2 \int_0^{\tilde{\lambda}(n+a)} \frac{\partial^2 l(\lambda,s)}{\partial s^2}\Big|_{s=(a+n)(1+r_o)} h_t(\lambda)d\lambda,
\end{aligned} \tag{D.6}$$

where the last equality is because $\tilde{\lambda}(n+a)$ is the unique root of $\phi_{n+a}(x)$ given in Equation (A.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

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

which is positive by property A(iv). □

E. Proof of Corollary 2

Setting $n = a = 0.0$ in Equation (8), 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 $a^*(0.0) > 0.0$. □

F. Proof of Corollary 3

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 (9). 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 D.2, we have $\psi_{n+x}(0.0) > \psi_n(0.0) \geq 0.0$, implying that

$$a^*(n + x) = a^*(n) = 0.0, \tag{F.1}$$

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

$$a^*(n + x) = \tilde{a}(n) - x, \tag{F.2}$$

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 D.2, $\psi_n(a)$ increases strictly with n, we have $\psi_{n+x}(0) > \psi_{n+\tilde{a}(n)}(0) = 0$, implying that

$$a^*(n + x) = 0.0, \tag{F.3}$$

for $x > \tilde{a}(n)$. Combining Equations (F.1), (F.2), and (F.3), we arrive at

$$a^*(n + x) = \begin{cases} 0.0 & \text{if } \psi_n(0) \geq 0.0, \\ \left(\tilde{a}(n) - x\right)^+, & \text{otherwise,} \end{cases} \tag{F.4}$$

for any $x \geq 0$. Setting $n = 0.0$ in Equation (F.4) and a change of variable yield

$$a^*(n) = \begin{cases} 0.0 & \text{if } \psi_0(0.0) \geq 0.0, \\ (\tilde{a}(0.0) - n)^+, & \text{otherwise.} \end{cases} \tag{F.5}$$

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

G. Algorithm 2

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

Require: $l(x, y)$, h_t, c_t, c_o, c_w, r_o, n, q_u, and function $\tilde{\lambda}(p)$ from Algorithm 1
1: Evaluate $\psi_n(0)$ from Equation (8) using $\tilde{\lambda}(n)$ as input
2: **if** $\psi_n(0) \geq 0$ **then**
3: $\quad a^*(n) \leftarrow 0$
4: **else**
5: $\quad a_u \leftarrow 10$
6: $\quad$ **while** $\psi_n(a_u) < 0$ **do**
7: $\quad\quad a_u \leftarrow a_u \times 10$
8: $\quad$ **end while**
9: $\quad \tilde{a}(n) \leftarrow$ root of $\psi_n(a)$ in the interval $[0, a_u]$
10: $\quad a^*(n) \leftarrow \min\{\tilde{a}(n), q_u\}$
11: **end if**

H. Proof of Corollary 4

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

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

which is negative. This, along with the fact that $\psi_n(a)$ is strictly increasing in a by Lemma D.2, implies that $\tilde{a}(n)$ and thus $a^*(n)$ is increasing in c_t.

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

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

which is negative by property A(iii). This, along with the fact that $\psi_n(a)$ is strictly increasing in a by Lemma D.2, 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 D.2, implies that $\tilde{a}(n)$ and thus $a^*(n)$ is decreasing in c_o.

(iv) By Lemma D.2, $\psi_n(a)$ is increasing in n. This, along with the fact that $\psi_n(a)$ is strictly increasing in a by Lemma D.2, implies that $\tilde{a}(n)$ and thus $a^*(n)$ is decreasing in n. □

I. Proof of Corollary 5

Taking the derivative of $m(n)$ given in Equation (3) with respect to c_t, we arrive at

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

For an arbitrary λ, we then have

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

Similarly, $\partial v(\lambda, n+a)/\partial c_t = g^*(\lambda, n+a)$. Substituting these derivatives into Equation I.1, we obtain

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

which is clearly non-negative.

Similarly, since

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

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. □

J. Proof of Proposition 3

Taking the derivative of $l(\lambda, s)$ given in Equation (10) with respect to s, plugging it into Equation (5), and setting the result equal to zero yields

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

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

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

Inserting the derivative of $l(\lambda, s)$ with respect to s in Equation (6), setting the result equal to zero and solving for g, we obtain the value for $\tilde{g}(\lambda, p)$ given in Equation (12). Equation (13) is then obtained by evaluating the objective function in problem (1) for $g = g^*(\lambda, p)$. □

K. Proof of Proposition 4

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 (14) 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 $s > \lambda$. □

L. Proof of Corollary 6

Taking the derivate of $l(\lambda, s)$ given in Equation (14) with respect to τ, we obtain

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

This gives

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

which, given the stability constraint, is always positive, and so $v(\lambda, p)$ is increasing in τ. Taking the derivate of $l(\lambda, s)$ given in Equation (14) with respect to τ and s, we obtain

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

which is negative. This implies that $\tilde{\lambda}(p)$, the route of function $\phi_p(x)$ given in (5) 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 (6) 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 $a^*(n)$. □

M. Proof of Proposition 5

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

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

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

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

Using the relation between $C(\lambda, s)$ and $B(\lambda, s)$ given in Equation (M.1), we then have

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

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 (M.2) is non-negative. 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. This has already been proved in the extant literature; see, e.g., Karsten et al. (2015). □

N. Proof of Proposition 6

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 (8) 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 (8) simplifies to $-c_t(1+r_o)$ and the third term simplifies to zero. These yield the expression in Equation (19).

For the second part, the expression given for $\tilde{a}(n)$ is obtained by replacing the derivative given in (L.1) in Equation (19) and solving $\psi_n(a) = 0$ for a. □

O. Proof of Proposition 7

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

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

where the second equality is obtained by replacing a with q in Equation (D.4). The expression given in (O.1) 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. □

P. Proof of Proposition 8

$\Lambda_{t'} \leq_{cx} \Lambda_t$ 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 (8) to indicate its dependence to Λ_t. From (D.2), (4), and the fact that $\tilde{g}(\lambda, p)$ is the root of $\theta_{\lambda,p}(g)$, we now have

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

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

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

respectively. By condition (b) in the proposition, $\dfrac{\partial^2 \omega(\lambda)}{\partial \lambda^2}$ is non-positive. By condition (a) and Lemma C.1, $\lambda_u \leq \tilde{\lambda}(n) \leq \tilde{\lambda}(n+a)$, and so $\lambda \leq \tilde{\lambda}(n+a)$ 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 $\psi_n^{\Lambda_t}(0) \geq 0$, we will also have $\psi_n^{\Lambda_{t'}}(0) \geq 0$, 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 $\psi_n^{\Lambda_{t'}}(0) \geq 0$, which implies that $a^*(n, Q_t, \Lambda_{t'}) = 0 < a^*(n, Q_t, \Lambda_t)$, or $\psi_n^{\Lambda_t}(0) \leq \psi_n^{\Lambda_{t'}}(0) < 0$, which implies that $a^*(n, Q_t, \Lambda_{t'}) \leq a^*(n, Q_t, \Lambda_t)$. □

Q. Proof of Proposition 9

First, note that $\Lambda_{t'} \leq_{cx} \Lambda_t$ implies that $\mathbb{E}\left[\omega(\Lambda_{t'})\right] \leq \mathbb{E}\left[\omega(\Lambda_t)\right]$ for any convex function $\omega(\lambda)$. Second, note that $m(n, Q_t, \Lambda_{t'}) \leq \mathbb{E}[v(\Lambda_{t'}, n + \min\{Q_t, a\})]$ 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

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

and

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

It then follows from Proposition 1 that

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

By condition (a) and Lemma C.1, $\lambda_u \leq \tilde{\lambda}(n) \leq \tilde{\lambda}(n + \min\{q, a\})$, 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. □

R. Summary of Notations

Symbol	Description
Parameters	
t	given advertisement epoch
t_e	beginning of the highly uncertain demand period (HUDP)
t'	delayed advertisement epoch
c_t, c_p, c_o	cost rates of temporary, permanent and mandatory overtime work
c_w	waiting cost incurred by 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 (5) in the interval $(0, p(1+r_o))$
$\tilde{g}(\lambda, p)$	unique root of function $\theta_{\lambda,p}(g)$ given in (6) 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 (8) 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
$\lambda^{(b)}$	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
Decision Variables	
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 demand rate λ and p permanent workers
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
Functions	
$u(\lambda, p, g)$	second-stage objective function given arrival rate λ, p permanent workers, and g temporary workers
$v(\lambda, p)$	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 Equation (16)
$l^{(n)}(\lambda^{(b)}, s)$	mean number of requests in the nursing system given patient arrival rate $\lambda^{(b)}$ and s nurse
$\phi_p(x)$	function given in Equation (5)
$\theta_{\lambda,p}(g)$	function given in Equation (6)
$\psi_n(a)$	function given in Equation (8)

References

Karsten, F., Slikker, M., & van Houtum, G.-J. (2015). Resource Pooling and Cost Allocation Among Independent Service Providers. *Operations Research*, *63*, 476–488.

Pacheco, A. (1993). *Some Properties of the Delay Probability in M/M/s/s+c Systems*. Technical Report Cornell University Ithaca, NY.

Takayama, A. (1985). *Mathematical Economics*. (2nd ed.). New York: Cambridge University Press.

Thomas, G. B. (2014). *Thomas' Calculus*. (Thirteenth ed.). Pearson Education.