MANAGEMENT SCIENCE
Vol. 57, No. 11, November 2011, pp. 1979–1995
ISSN 0025-1909 | EISSN 1526-5501 | 11 | 5711 | 1979

http://dx.doi.org/10.1287/mnsc.1110.1399

Modeling Security-Check Queues

Zhe George Zhang
School of Management, Lanzhou University, Lanzhou, 730000 Gansu, People's Republic of China;
Beedie School of Business, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada; and
College of Business and Economics, Western Washington University, Bellingham, Washington 98225,
george.zhang@wwu.edu

Hsing Paul Luh, Chia-Hung Wang
Department of Mathematical Sciences, National Chengchi University, Taipei, 11605 Taiwan, Republic of China
{slu@nccu.edu.tw, jhwang728@hotmail.com}

Motivated by the waiting lines between the U.S.–Canadian border crossings, we investigate a security-check system with both security and customer service goals. In such a system, every customer has to be inspected by the first-stage inspector, but only a proportion of customers need to go through the second stage for further inspection. This "further inspection proportion," affecting both security screening and the system congestion, becomes a key decision variable for the security-check system. Using a stylized two-stage queueing model, we established the convexity of the expected waiting cost function. With such a property, the optimal further inspection proportion can be determined to achieve the balance of the two goals and the service capacities can be classified into "security-favorable," "security-unfavorable," or "security-infeasible" categories. A specific capacity category implies if the security and customer service goals are consistent or in conflict. In addition, we have verified that the properties discovered in the stylized model also hold approximately in a more general multiserver setting. Numerical results are presented to demonstrate the accuracy and robustness of the approximations and the practical value of the model.

Key words: security inspection level; service capacity; two-stage queue; renewal process approximation; Coxian distribution; quasi-birth-and-death process
History: Received September 13, 2010; accepted May 16, 2011, by Assaf Zeevi, stochastic models and simulation. Published online in *Articles in Advance* September 2, 2011.

1. Introduction

When a country senses nothing but a single threat to national security, travellers are unfortunately stricken with the poor and inconvenient effects of congested security checkpoints at international border crossing stations. Is there an effective security-check system that not only caters to the needs of nonrisk individuals, but also sustains the efficiency of screening high-risk individuals? The aim of this paper is to answer this important question.

A security-check system at a border crossing station between the United States and Canada can be modeled as a two-stage queueing system. Customers arriving at the security-check system join the primary queue, called stage 1 queue, for a standard inspection. During the standard inspection, the border inspector performs preliminary examination by checking the traveler's documents, verifying the vehicle's license plate, and asking some routine questions. If any suspicion is raised or a random number for further inspection is generated from the computer, the vehicle is pulled aside to the secondary queue, called stage 2, for further detailed inspection. Otherwise, the inspector will continue to ask further questions and/or visually check the car trunks or back seats. The vehicle is either allowed or denied entry at the end of primary (standard) or secondary (detailed) inspection. Clearly, the customer may be selected for further inspection in the middle of stage 1 inspection (or between the two phases of stage 1 inspection). Therefore, after the initial screening at stage 1, a certain proportion (p) of customers (vehicles) are selected for further inspection and the rest proportion of $(1-p)$ are inspected briefly without going through the second stage. Such a system is called a two-stage security-check system (TSCS). The security-check level is reflected by two parameters: p at stage 1; and the average time of further inspection at stage 2, denoted by S. We use the lower bounds of these two parameters as the required security-check level and denote them by (p_0, S_0). Although p_0 is adjusted from time to time based on the security level change, S_0 is relatively stable and depends mainly on the inspection procedure and the facility capacity. For example, a security-check level $(0.2, 5)$ for a border crossing station represents that at least 20% of the vehicles are selected for further inspection and the average further inspection time is at least five minutes. If S_0 is the required

expected service time for one inspection station, the overall potential maximum service rate (when all stations are busy) for the second stage with k stations is $\nu = k/S_0$. For a given S_0, the service capacity for further inspection can be represented by the rate ν. In this paper, we examine the security screening and congestion effects of changing p and ν in a security-check system. In many practical situations, a TSCS is required because inspection procedures/equipment for the second-stage checks are different from those of the first-stage checks. Besides the procedural requirements, a TSCS also offers the advantage of distinguishing between nonselected and selected customers to improve customer service. A unique feature of TSCS is that customers are not classified into different screening levels until after part of stage 1 inspection is performed. This feature makes the security-check system different from other service systems with multiple classes of customers, which are usually identified by either service providers or customers themselves before services start. Another difference from other systems (such as call centers) is that the number of servers in a security-check system is usually not large (fewer than 20 for most practical systems). Thus, the solutions based on stochastic process limits (Whitt 2002) do not apply to the system considered here.

Motivated by a repair facility, El-Taha and Maddah (2006) present a study on a multiserver system with similar setting to the TSCS. In their model, the service requirements of arriving customers are not known in advance. Only the customers whose service times in the first stage exceed a threshold will continue to the second-stage service. There are two major differences between our model and El-Taha and Maddah's: (1) The customer class in our model is determined by the inspection policy, which can be adjusted by the server according to the security level required. Thus, the proportion of customers going to the second-stage service is a decision variable to maximize the probability of "True Alarm," defined as an event where the system gives an alarm and a threat exists. In contrast, the customer class in El-Taha and Maddah (2006) is determined by the overall service time distribution, which is beyond the server's control. (2) In our model, if the customer is classified as a "selected" customer who must go through further inspection in the second stage, the service time is a completely new random variable, which is independent of the first-stage service time. However, in El-Taha and Maddah's model, the service time in the second stage is the continuation of the first-stage service. Moreover, our analysis is different from El-Taha and Maddah's.

Besides the relevant queueing analysis literature, there is a growing body of research on the security screening of air passengers after the 9/11 terrorist attacks. Most of these works focus on how to classify passengers into different classes and how to allocate them into different security-check lanes to maximize the probability of the "true alarm" without considering the minimization of the average passenger waiting time. Kobza and Jacobson (1997) develop probability models based on Type I (false alarm) and Type II (false clear) errors to determine different paths for customers to go through the security screening system. Jacobson et al. (2001) present a model for minimizing the false alarm probability in an aviation security-check system by solving a knapsack problem. McLay et al. (2006, 2010) study the security-check system with customers of multiple inspection classes and propose a multilevel allocation problem to allocate multiclass customers into different security-check channels to maximize the true alarm rate, subject to some budget constraints. Nikolaev et al. (2007) consider a sequential stochastic security design problem where both design and operating issues in security-check system are addressed. Other discrete optimization models for sequentially assigning passengers to a set of security classes after initial screening upon check-in include Lee et al. (2009) and Babu et al. (2006). Our work complements these studies by focusing on the performance analysis and trade-off between security and customer service goals of a security-check system. There are relatively fewer studies on security-check waiting lines using analytical models. Wilson et al. (2006) present the security checkpoint optimizer, a two-dimensional spatially aware discrete event simulation tool, for building simulation models to evaluate how changes or additions to the security-check facilities or procedures impact security effectiveness, operational costs, and passenger throughput. Zhang (2009) provides a detailed analysis about the first-stage inspection queue motivated by border crossing stations. Lee et al. (2009) study the impact of aviation checkpoint queues on optimizing security screening effectiveness by assuming that the first-stage primary screening time is exponentially distributed. However, using the exponential inspection time may not realistically model the detailed congestion dynamics in a security-check system.

In this paper, we address the important trade-off issue between the security screening effectiveness and the customer service quality, the two main goals of a security-check system. Our approach is to develop a stylized queueing model with the novel features pertaining to the real system. To overcome the shortcomings of exponential random variables, we model the first-stage service time by using the Coxian-2 distributed random variable, which captures the main characteristics of the inspection process in a TSCS. This model reveals the fundamental properties of the inspection policy, such as the convexity of

the expected waiting cost function. The convexity property has an important practical implication that suggests in some situations it is better to select more customers for further inspection beyond the minimum required p_0, because it not only increases the True Alarm but also reduces the average delay of customers. To establish the convexity of the cost function, we develop a new and effective approximation technique to derive the major performance measures. Moreover, using simulations, we confirm that these properties also hold in a more complex and realistic security-check system. Another benefit of our study is that the model developed can also be applied to other situations such as "help-desk call centers" or "walk-in clinics" where the server has the choice of serving or referring the customer to another service system (see Shumsky and Pinker 2003).

The rest of this paper is structured as follows: §2 provides the model formulation and analysis. A two-stage, single-server model is used to investigate the performance characteristics of the security-check system. Both exact analysis and approximation approach are presented. In §3, the managerial insights obtained from the stylized model are discussed. Section 4 presents numerical examples to show the accuracy and robustness of the approximations and the practical value of the model. Finally, §5 concludes. All proofs of propositions are relegated to the appendix. Additional numerical results and some computational algorithms appear in the e-companion (available at http://www.cbe.wwu.edu/zhang/ecompanion.htm).

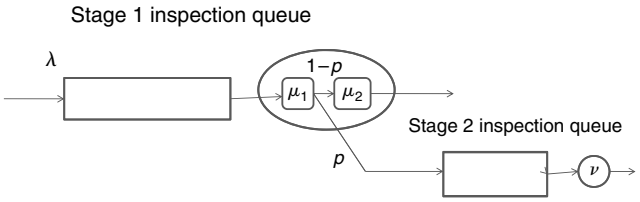
2. Model Formulation and Analysis

2.1. A Stylized Inspection Queue Model

To simplify the analysis, we focus on a two-stage queueing model with a single server for each stage, as shown in Figure 1, which is an approximation of the real system with multiple servers at either one or both stages. With this stylized model, we reveal the fundamental properties of the system performance. In particular, the expected waiting cost function is shown to be convex in the proportion for further inspection or p, the key decision variable for practitioners. Furthermore, simulation results indicate that such a stylized model can be used to estimate the optimal p for a practical multiserver TSCS with generally distributed service times. The applicability of results from the stylized model to a more general system justifies the use of the stylized model. It is worth noting that even with the approximation approach, treating a multiserver model with generally distributed service times prevents us from establishing the convexity of the cost function.

Assume that customers arrive at the system according to a Poisson process with rate λ and join a queue

Figure 1 A Single-Server TSCS Model



in front of the stage 1 server. The Poisson arrival process is reasonable as cars arrive from different sources independently. Because of the nature of the inspection process at stage 1, the service time is assumed to follow a Coxian-2 distribution with rate μ_1 for phase 1 and rate μ_2 for phase 2. Thus, the queueing performance measures for stage 1 can be obtained by solving an M/Cox-2/1 system. The service time in the second-stage queue is assumed to be exponentially distributed with rate ν. Because of the Coxian-2 service, there are two suboutflows from stage 1, as shown in Figure 1. Therefore, after the first phase of the service (initial screening), the customer will be inspected in one of the two ways. With probability $1-p$, this customer continues to complete the second phase of the service at stage 1, called primary inspection, and then leaves the system. The primary inspection usually takes less than one minute and may consist of asking more questions and a visual inspection of the vehicle (inside and trunk) on the spot. With probability p, a customer is selected for further/detailed stage 2 inspection, called secondary or further inspection. Further inspection is much more detailed. The inspectors will check the traveller's documents and search a computer database to check if the customer has inconsistent information or criminal records. At the same time, the vehicle is completely inspected. Using the Coxian-2 distribution is a unique feature that captures the main characteristics of the first-stage inspection. However, this feature makes the analysis more complex, as we cannot use a Jackson network here. Because both "primary inspection" at stage 1 and "further inspection" at stage 2 are the standard procedures, p is independent of service time parameters of μ_1, μ_2, and ν. More general and realistic systems with nonexponential phases 1, 2, and stage 2 services are examined by using the simulations in §4.

2.2. Determination of Minimum Proportion for Further Inspection

In reality, the actual proportion of customers selected for further inspection is determined by (a) the screening standard, and (b) the random number generation at phase 1 of stage 1. Here are a few possible cases of (a): (i) Customers carry inconsistent documents. For example, a person is holding a TN visa (temporary

stay) but intends to stay longer (or permanently) in the United States. (ii) Travel purpose is suspicious. (iii) Customers need visa processing. (iv) A vehicle is suspected to carry a terrorist threat or illegal items. The proportion of these customers, denoted by p_c, is usually small (<5%) and is not controllable for a given screening standard, because it depends on the proportion of the travellers who fall in the specific categories. Based on the historical data, this proportion can be estimated. However, for (b), the proportion from the random number generation, denoted by p_d, is completely controllable and is determined by the security level. Thus, the actual proportion of further inspections p is the proportion of the union of these two nonoverlapping subsets and can be controlled via p_d. Letting p_0 be the minimum proportion for further inspection, we need $p = p_c + p_d - p_c p_d \geq p_0$ or $p_d \geq (p_0 - p_c)/(1 - p_c)$. Any customer crossing the border (with or without further inspection) will be given either a "clear" signal (the entry is allowed) or an "alarm" (the entry is denied). There are four possible outcomes: (i) True Alarm (TA)—this system gives an alarm and a threat exists; (ii) False Alarm (FA)—this system gives an alarm but a threat does not exist; (iii) True Clear (TC)—this system gives a clear signal and a threat does not exist; and (iv) False Clear (FC)—this system gives a clear signal but a threat exists. The security goal is to maximize the probability of TA, which is equivalent to minimizing the probability of FC for a given threat rate of the customer population. Define the following events: $A = \{$*inspection system gives an alarm*$\}$; $T = \{$*customer carries a threat*$\}$; $FI = \{$*customer is selected for further inspection*$\}$ (FI^c is the complement of FI); $SIS = \{$*customer is selected by the initial screening procedure*$\}$; and $SRN = \{$*customer is selected by the random number generation*$\}$. Then $FI = SIS \cup SRN$. For a given $P(FI) = p$, we also define the following probabilities: TA rate for selected customers $\theta_{FI} = P(A \mid T \cap FI)$, TA rate for nonselected customers $\theta_{FI^c} = P(A \mid T \cap FI^c)$, threat rate of the population $\tau = P(T)$, threat rate of selected customers $\alpha(p) = P(T \mid FI)$, and threat rate of nonselected customers $\beta(p) = P(T \mid FI^c)$. Note that the last two threat rates have been denoted as functions of p, as shown later. Because the second-stage inspection procedure is more strict than the primary inspection of the first stage, we assume $\theta_{FI} > \theta_{FI^c}$. Using the law of total probability, for given $\theta_{FI}, \theta_{FI^c}$, and τ, we write the probability of true alarm $P(TA)$ as a function of p:

$$P(TA) = f(p) = \theta_{FI}\alpha(p)p + \theta_{FI^c}\beta(p)(1-p). \qquad (1)$$

Define $\gamma = P(T \mid$ *selected for further inspection by initial screening standard*$) = P(T \mid SIS)$. Obviously $\tau = P(T \mid$ *selected for further inspection by random number generation*$) = P(T \mid SRN)$, which should be the same as unconditional $P(T)$. For an effective initial screening performed in stage 1, it is reasonable to assume that $\gamma > \tau$. Obviously, we have $\alpha(p) = [(p - p_c)/p]\tau + (p_c/p)\gamma$ and the following proposition.

Proposition 1. (i) *The conditional threat rates* $\alpha(p)$ *and* $\beta(p)$ *are decreasing functions of* p *and* $\beta(p) < \tau < \alpha(p)$, *and* (ii) $P(TA)$ *is an increasing function of* p.

See the appendix for the proof.

Based on this proposition, a sufficiently high $P(TA)$ (<1) can be achieved with a sufficiently high p (achieved by choosing a sufficiently high p_d, because p_d is controllable). For a given threat level measured by $\tau = P(T) = P(TA) + P(FC)$, maximizing $P(TA)$ is equivalent to minimizing probability of false clear or $P(FC)$, which is usually the ultimate goal of a security-check system. In practice, the minimum $p = p_0$ is determined by ensuring a sufficiently high $P(TA)$ (or a sufficiently low $P(FC)$) is achieved. For example, suppose that $\tau = P(T) = 0.013$, $\theta_{FI} = 0.99$, $\theta_{FI^c} = 0.89$, $\gamma = 0.048$, and $p_c = 0.05$ (these values are in the same magnitudes as the data provided by the Bureau of Transportation Statistics 2006). If we want the probability of the false clear (also called Type II error) to be no more than 0.001, the required security level becomes $P(FC) \leq 0.001$. Substituting this data set into $\tau = \alpha(p)p + \beta(p)(1-p)$, $\alpha(p) = [(p - p_c)/p]\tau + (p_c/p)\gamma$, $P(TA) = \tau - P(FC)$, and (1), we can find $p_0 = 0.1961$. This means that to ensure that the security level is achieved, at least 19.61% of customers must be selected further inspection. Because $p_d \geq (p_0 - p_c)/(1 - p_c)$, and using $p_0 = 0.1961$ and $p_c = 0.05$, we have $p_d \geq 0.1538$. In other words, as long as we use random number generation to select at least 15.38% of customers for further inspection, the security level can be reached. Thus, after p_0 is determined, it becomes a constraint for p in the security-check system. In the following discussion, we treat p as the decision variable, as its value can be completely determined by the controllable p_d and the preestimated p_c.

2.3. System Stability and Exact Performance Analysis

From the workload management perspective, we identify the stability condition for implementing a two-stage inspection policy. Note that because of the nature of the customer flows, p affects the service time of the stage 1 queue and the interarrival time of the stage 2 queue.

Proposition 2. *If* (i) $\lambda/\mu_1 < 1$ *and* (ii) $\nu > \lambda - \mu_2(1 - \lambda/\mu_1)$, *for any* $p \in (p_{\min}, p_{\max})$, *the two-stage inspection policy makes both stage queues stable, where* $p_{\min} = \max(1 - \mu_2(1/\lambda - 1/\mu_1), 0)$ *and* $p_{\max} = \min(\nu/\lambda, 1)$.

See the appendix for the proof.

Under a feasible p, the system is stable and the steady state can be reached. Based on the $M/G/1$ formula (see Gross and Harris 1996), we can compute the expected queue length and expected waiting time at the first stage as

$$L_1^q = \frac{\lambda^2 E[S_1^2]}{2(1-\lambda E[S_1])}, \quad E(W_1^q) = \frac{\lambda E[S_1^2]}{2(1-\lambda E[S_1])}, \qquad (2)$$

where $E[S_1]$ and $E[S_1^2]$ are the first two moments of the Coxian-2 distributed service time of stage 1. For the second stage, we cannot treat it as an independent $GI/M/1$ queue because the arrival process as a suboutflow *between* the two phases of the first-stage service is a nonrenewal process (Ross 1997). To compute the exact performance measures, we can use a computational approach by modeling the entire system as a quasi-birth-and-death process. Such a model has been presented in the e-companion. The expected queue length and waiting time can be numerically computed via the stationary distribution. However, there are three disadvantages for this avenue that prevent us from achieving the modeling goals: (1) for a heavy traffic system ($\rho_2 = \lambda p/\nu \to 1$), we must choose a very large N for state space truncation, which significantly increases the computational complexity; (2) with the computational approach, it is not easy to reveal the fundamental properties of the security-check policy; and (3) because of the "curse of dimensionality," it is very hard to extend the model to multiserver queues in either (or both) stage(s) or nonexponentially distributed inspection durations (see Latouche and Ramaswami 1999). To avoid these disadvantages, we seek an appropriate approximation method.

2.4. Approximation Approach to Performance Evaluation

Now we develop some simple and accurate closed-form approximations to the performance measures of the second-stage queue. We adopt the renewal process approximation approach to obtain the Laplace-Stieltjes transform (LST) of the interarrival time for the second-stage queue (ITSS). It is worth noting that our approach is different from classical approaches, which are mainly based on the first two moments instead of the LSTs (see Whitt 1982, El-Taha and Maddah 2006). Specifically, our approximation is based on the combination of the two renewal processes. The first renewal process is from the following approximation regarding the two conditional interarrival times for the second-stage queue.

Approximation 1. (a) *If the first-stage queue server is busy at an arrival instant to the second stage, the time to the next arrival instant is either the phase* 1 *service time of the next customer in stage* 1 *with probability* p *or the duration of completing both phases of the next customer in stage* 1 *plus an unconditional approximate inter-arrival time of a renewal process for stage* 2 *queue with probability* $1-p$. (b) *If the first-stage queue server is idle at an arrival instant to the second stage, the time to the next arrival instant is a single idle period of the first-stage queue plus a conditional interarrival time of the second-stage queue defined in* (*a*).

Note that the stage 1 service time follows a Coxian-2 distribution, with μ_1 and μ_2 as the parameters of the exponentially distributed phase 1 and phase 2. Denote by $\tilde{A}$ and $A(s)$ the interarrival time and its LST, respectively, of the renewal process approximation for the second-stage queue. Let $X(s) = \mu_1/(\mu_1+s)$, $Y(s) = [\mu_1/(\mu_1+s)][\mu_2/(\mu_2+s)]$, and $I(s) = \lambda/(\lambda+s)$ be the LSTs of the phase 1 duration, the sum of the two phase durations, and the idle period, respectively. Based on part (a) of Approximation 1, the conditional LST of interarrival time for the second-stage queue (ITSS), given that the stage 1 server is busy at an arrival instant is

$$a_b(s) = pX(s) + (1-p)Y(s)A(s). \qquad (3)$$

Using part (b) of Approximation 1, another conditional LST of ITSS, given that the stage 1 server is idle at an arrival instant, is

$$a_I(s) = I(s)a_b(s) = I(s)[pX(s) + (1-p)Y(s)A(s)]. \qquad (4)$$

Letting $\rho_1 = \lambda E(S_1) = \lambda\{1/\mu_1 + (1-p)(1/\mu_2)\}$, as the first stage is an $M/G/1$ queue, we also have

$$A(s) = \rho_1 a_b(s) + (1-\rho_1)a_I(s). \qquad (5)$$

Substituting (3) and (4) into (5) and solving for $A(s)$, we obtain the following proposition.

Proposition 3. *Based on Approximation* 1, *a renewal process approximation for the arrival process to the second-stage queue has the LST of the interarrival time as*

$$A(s) = \frac{pX(s)[\rho_1 + (1-\rho_1)I(s)]}{1-(1-p)[\rho_1 + (1-\rho_1)I(s)]Y(s)}. \qquad (6)$$

Remarks. It is easy to check that (6) has three desirable properties: (i) the mean interarrival time is $1/(p\lambda)$, which is consistent with the flow conservation law (as shown in the proof of Proposition 4 in the appendix); (ii) when $p = 1$ or 0, $A(s)$ becomes exact results of $\lambda/(\lambda+s)$ or 0; and (iii) when the traffic to stage 1 becomes heavier and the chances of an idle period become smaller (in most practical situations), $A(s)$ is approaching the correct LST of the ITSS. For a limiting case, if the first-stage server is always busy, the arrival process to the second-stage queue becomes

a renewal process with the LST of the interarrival time as

$$A(s) = pX(s) + (1-p)pY(s)X(s) + (1-p)^2pY^2(s)X(s) + \cdots (1-p)^npY^n(s)X(s) + \cdots = \frac{pX(s)}{1-(1-p)Y(s)}, \quad (7)$$

which is the same as the limiting case of $A(s)$ in (6) by letting $\rho_1 \to 1$.

Note that part (b) of Approximation 1 implies that an idle period in stage 1 is followed by a conditional stage 1 busy interarrival time to the second-stage queue. This means that we ignore the case where a single customer is served between two consecutive idle periods at the first stage in Approximation 1. Thus, using $A(s)$ as a renewal process approximation to the true LST of the interarrival time should give an upper bound on the expected waiting time for the second-stage queue. (Unfortunately, we do not have a rigorous proof for this argument at this time. But extensive numerical results verified this claim; see the e-companion for some examples.) To improve the approximation, we introduce the second approximation, which usually yields a lower estimate.

Approximation 2. *The arrival process to the second-stage queue can be approximated by a Poisson process with arrival rate λp.*

The basis of proposing Approximation 2 is the fact that (i) under the steady-state the arrival rate into the second-stage queue is λp; and (ii) if the service time of the first stage is a single exponentially distributed (rather than a Coxian) random variable, then the arrival process to the second-stage queue will be exactly the Poisson process with arrival rate λp (the model becomes a Jackson network). In our model, the service time in the first stage follows a Coxian-2 distribution. Thus, the true arrival process as a subout-flow from between the two phases is a nonrenewal process (certainly non-Poisson process). However, we still propose Approximation 2 as the second renewal process, as the Poisson process approximation gives a lower estimate of the expected waiting time than using Approximation 1. Hence, it can be combined with the overestimate from $A(s)$ of (6) to produce a better approximation. Intuitively, this is because going through part of the Coxian-2 distributed stage 1 service process will increase the variability of the interarrival time for the second-stage queue (this is formally proved in the following proposition). Denoting by $\tilde{A}^{\exp}$ the interarrival time of the Poisson process with rate λp, its LST is given by

$$A^{\exp}(s) = \frac{\lambda p}{\lambda p + s}. \quad (8)$$

Like $A(s)$ in (6), for the extreme case of $p = 1$ or 0, $A^{\exp}(s)$ also becomes exact. Note that both renewal processes give the same mean but different variances, denoted by $Var(A)$ and $Var(A^{\exp})$, respectively, of the interarrival times.

Proposition 4. *The expected interarrival time, $E(A) = E(A^{\exp}) = 1/(\lambda p)$, and the variance of interarrival time, $Var(A) > Var(A^{\exp})$.*

See the appendix for the proof.

Proposition 4 supports that the expected waiting time for the second-stage queue based on $A(s)$ is greater than that based on $A^{\exp}(s)$. With the two $A(s)$s, the performance measures can be obtained by treating the second stage as $GI/M/1$ and $M/M/1$ queues, respectively. First, from $A(s)$ of (6), we solve the functional equation of $A(\nu(1-z)) = z$ for the root, denoted by r_0, which is strictly less than 1 (Gross and Harris 1996). Although the numerical solution can be obtained easily, we cannot get a closed-form formula for this root. Based on Approximation 2, we can use the $M/M/1$ formula. Extensive numerical analysis indicated that the simulated performance measures are usually bracketed by the two approximations, with the first approximation as an upper bound and the second approximation as a lower bound (see the e-companion). It is worth noting that the difference between the simulated result and either of the two approximations is a complex and nonmonotonic function of the decision variable. Thus, it is not possible to use a simple adjustment from either approximation alone. Therefore, our proposed approximation to the expected waiting time for the second-stage queue should be a linear combination of the two approximated expected waiting times, as $E(W_2^q)_{\text{approx}} = wr_0/[\nu(1-r_0)] + (1-w)\lambda p/[\nu(\nu-\lambda p)]$, where w is the weight. Based on the numerical tests (see the e-companion), we observed that using a simple average of the two approximations generates very good approximations. Thus, we propose to use $w = 1/2$ or

$$E(W_2^q)_{\text{approx}} = \frac{1}{2}\left(\frac{r_0}{\nu(1-r_0)} + \frac{\lambda p}{\nu(\nu-\lambda p)}\right). \quad (9)$$

Because of the Coxian-2 distributed service times at stage 1, when p is getting larger, the arrival stream into the second stage is getting closer to an output stream from an $M/M/1$ type queue; thus, a simple Poisson process (using Approximation 2 alone) should perform as well as (9). This has been verified by the numerical study for higher $p \geq 0.6$. Comparing the approximations with the simulations indicates that the simple Poisson process produces almost equally good results as proposed approximations. Note that as both component approximations in (9) approach to the Poisson process, hence (9) also

approaches to the Poisson process. However, for the practical security-check system, p is usually in the lower value range (no more than 35% of customers are selected for further inspection). As indicated by the numerical results in the e-companion, for lower p value cases, our proposed approximation generates much more accurate results than the simple Poisson approximation. With the proposed approximation, other performance measures can be obtained easily. Clearly, $E(W_2^q)_{\text{approx}}$ is increasing in p, as an increase in p results in an increase in the arrival rate to the second-stage queue. Denote the expected system time for those nonselected customers by $E(T_1)$. Based on (2), we have the following:

Proposition 5. *The expected values $E(W_1^q)$ and $E(T_1)$ are given by*

$$E(W_1^q) = \frac{\lambda/\mu_1^2 + (1-p)(\lambda/\mu_2^2 + \lambda/(\mu_1\mu_2))}{1-\lambda/\mu_1-(1-p)\lambda/\mu_2},$$
$$E(T_1) = E(W_1^q) + \frac{1}{\mu_1} + \frac{1}{\mu_2}, \tag{10}$$

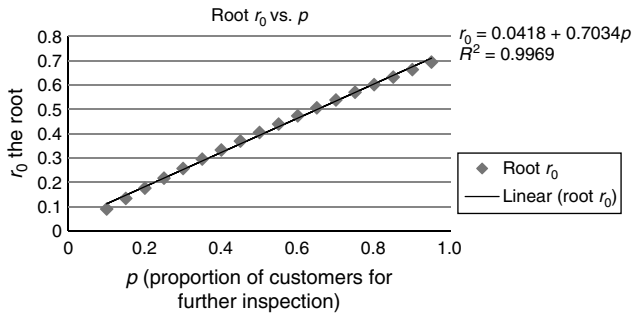
which are decreasing and convex in p.

See the appendix for the proof.

For $E(W_2^q)_{\text{approx}}$, we have studied the relationship between r_0 and p by numerically solving $A(\nu(1-z)) = z$ for the roots over a wide feasible range of p and have found almost perfect linear relationship, as shown in Figure 2, for all cases (more than 100 cases tested). This observation may be because the arrival process to stage 2 is a suboutflow of the Coxian-2 service time of stage 1. The coefficients of determination for all cases tested are greater than 99%, which indicates that the linear function is almost a perfect estimation of the relationship between r_0 and p for our model. Therefore, we make the following approximation.

Approximation 3. *The relation between the root $r_0 < 1$ of $A(\nu(1-z)) = z$ of* (6) *and p is approximated by a linear function of $r_0 = a + bp$, where a and b are the constants determined by system parameters λ, μ_1, μ_2, and ν via a linear regression of r_0 on p.*

Figure 2 **Relationship Between r_0 and p for a TSCS with $\lambda = 8.5$, $\mu_1 = 20$, $\mu_2 = 15$, and $\nu = 8.7$**



Under Approximation 3, a more specific relationship between the approximated expected system time at stage 2, denoted by $E(T_2)_{\text{approx}}$ (waiting time plus inspection time), and p can be established.

Proposition 6. *Using Approximation* 3 *and* (9), *the second-stage expected system time, $E(T_2)_{\text{approx}}$, can be written as a function of p:*

$$E(T_2)_{\text{approx}} = \frac{1}{2}\left(\frac{1}{\nu(1-a-bp)} + \frac{1}{\nu-\lambda p}\right), \tag{11}$$

where a and b are the regression constants for a given parameter set of λ, μ_1, μ_2, and ν, and is increasing and convex in p.

See the appendix for the proof.

Note that $E(W_q^2)_{\text{approx}} = E(T_2)_{\text{approx}} - 1/\nu$, which is also increasing and convex in p. With (10) and (9), we can study the performance characteristics of the selective security-check policy. However, the stylized queueing model treated here is a simplified version of the real system. In most practical security-check systems, either one or both stages have multiple servers, and service times may not be exponentially distributed. However, directly analyzing such a complex system exactly is virtually impossible, and using the approximation approach prevents us from discovering the fundamental relationship between the major performance measures and the decision variable. Therefore, we are motivated to utilize the stylized model to reveal the fundamental performance properties in §3 and verify them in a more realistic system via simulations in §4.

3. Managerial Decisions

3.1. Security-Check Level p Decision

We first address the problem of determining p, which affects the security screening measure $P(TA)$ and the expected customer waiting cost $E(WC)$. These two measures can be balanced by either minimizing $E(WC)$ subject to a minimum $P(TA)$ or maximizing $P(TA)$ subject to a maximum $E(WC)$. We focus on the first option and discuss the second option briefly. Let h_i be the unit waiting cost of class i customer, where $i = 1$ (nonselected), 2 (selected). Given a feasible service capacity of satisfying conditions (i) and (ii) of Proposition 2, our problem of finding the optimal p can be written as

$$\begin{aligned}&\min_p E(WC)\\ &= (1-p)E(T_1)h_1 + p[E(W_1^q) + 1/\mu_1 + E(T_2)]h_2\\ &\approx E(WC)_{\text{approx}}\\ &= (1-p)E(T_1)h_1 + p[E(W_1^q) + 1/\mu_1 + E(T_2)_{\text{approx}}]h_2 \quad (12)\end{aligned}$$

subject to

$$\max\{p_0, p_{\min}\} < p < p_{\max},$$

where p_0 is determined by setting a sufficiently high $P(TA)$, as discussed in §2.2, $p_{\max} = \min(\nu/\lambda, 1)$, and $p_{\min} = \max(1 - \mu_2(1/\lambda - 1/\mu_1), 0)$ (see the proof of Proposition 2 in the appendix). When $h_1 = h_2 = 1$, $E(WC)$ becomes the expected total system time (waiting and inspection) for a customer. Note that the system expected waiting cost per time unit is $\lambda E(WC)$, which can be used as an equivalent objective function. From both numerical and analytical results, we find that (12) is a convex function of p for practical value ranges of h_1 and h_2. Specifically, for $h_1 \geq h_2$ cases, the convexity can be proved, and for $h_1 < h_2$ cases, numerical tests indicate unless h_1/h_2 is extremely small, (12) remains to be convex in p for all practical systems. A discussion on the conditional convexity for the $h_1 < h_2$ case is given in the appendix. Another possible cost structure is to let h_i' be the unit waiting cost of stage i. Then the approximate expected waiting cost per customer becomes

$$E(WC)^{\text{stage}}_{\text{approx}} = [E(W_1^q) + 1/\mu_1 + (1-p)(1/\mu_2)]h_1' + p[E(W_2^q)_{\text{approx}} + 1/\nu]h_2'.$$

The advantage of using this cost structure is that we can establish the convexity of $E(WC)^{\text{stage}}_{\text{approx}}$ easily by using Propositions 5 and 6.

Proposition 7. (i) *If $h_1 \geq h_2$, the approximate expected waiting cost $E(WC)_{\text{approx}}$ per customer is a convex function of p.* (ii) *For any stage cost parameters $h_i' \geq 0$ $(i = 1, 2)$, the approximate expected waiting cost $E(WC)^{\text{stage}}_{\text{approx}}$ per customer is a convex function of p.*

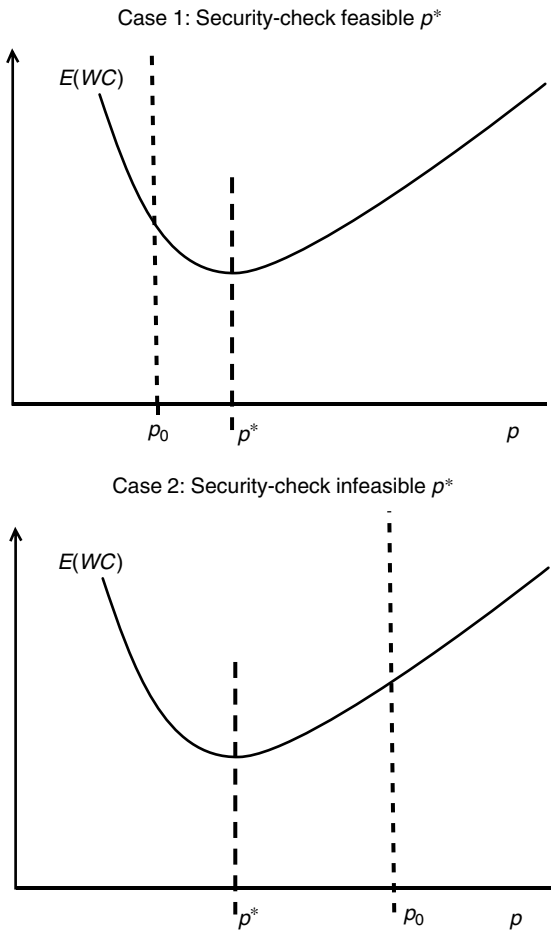
See the appendix for the proof.

The convexity of $E(WC)$ indicates that p determines the workload allocation between the two stages and the optimal p^* balances the two stages' expected waiting costs. Letting $f(p) = (1-p)E(T_1)h_1 + p[E(W_1^q) + 1/\mu_1 + E(T_2)_{\text{approx}}]h_2$ be the approximation function to $E(WC)$ and using Proposition 7, one can determine the optimal p^* to minimize $E(WC)$ with the first-order derivative condition of $f(p)$:

$$p^* = \{p\colon df(p)/dp = 0\}. \tag{13}$$

Although there is no closed-form expression for p^*, its value can be obtained easily from (13) by standard PC programs such as Excel's solver. For a predetermined p_0, there are two possible cases for the optimal p^*: (1) a security-check feasible case (or $p^* \geq p_0$) and (2) a security-check infeasible case (or $p^* < p_0$), as shown in Figure 3. In Case 1, it is interesting to see that the security and customer service goals are consistent, as when p increases in the range of $[p_0, p^*]$, $E(WC)$ decreases. Clearly, with a given p_c (selected by initial screening), to achieve p^*, we should set the random number selected proportion $p_d = (p^* - p_c)/(1 - p_c)$. In contrast, in Case 2, these two objectives are in conflict and any further increase in p beyond p_0 will increase the $E(WC)$. Thus, the actual proportion for further inspection should be set to equal p_0 or $p_d = (p_0 - p_c)/(1 - p_c)$. Whether Case 1 or Case 2 occurs depends on p_0 and the shape of the $E(WC)$ function, which in turn is mainly determined by the service capacity status of the two stages. Because of the security-check requirement, we cannot simply consider the second-stage capacity as a "back-up" capacity for the first-stage service. There are two reasons for this: (a) the secondary inspection procedure is different from the primary inspection procedure, as the use of stage 2 is only due to the security screening necessity rather than expanding the "primary inspection" operation. This service capacity is reserved for further inspection and cannot be used for performing the primary inspection when it becomes idle or lightly loaded. (b) The performance behavior of changing p depends on the shape of the $E(WC)$. When the second-stage service rate is low enough (very high security level required), it is possible that any positive p will increase the overall expected waiting cost (or the minimum $E(WC)$ occurs at $p = 0$). There exists

Figure 3 **Two Possible Cases for p^***



an extreme case of $p_{\min} > p_0$ where the second-stage capacity does play a "back-up capacity" role. However, this case is very unlikely, as the staffing level at stage 1 is high enough to make $p_{\min}$ zero or very small. More implications of these two cases are discussed in the section on service capacity decision.

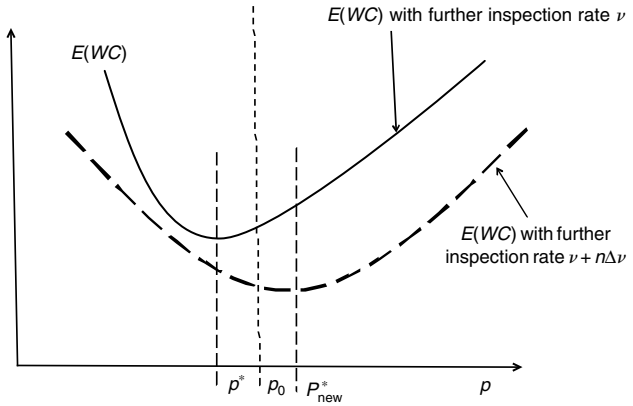
It also follows from the convexity of $f(p)$ that any $E(WC)$ above the minimum point of $f(p)$ corresponds to two p values, namely, low and high inspection proportions. If the system manager wants to maximize $P(TA)$, subject to a maximum $E(WC)$, he or she should use the higher p.

3.2. Service Capacity Decision

Whether p^* is feasible for a required p_0 also provides useful information for a security-check system manager in evaluating the service capacity of the second stage (primary inspection capacity is not considered as the maximum number of inspection booths of the first stage is fixed). Based on Proposition 7, assuming $h_1 = h_2$ ensures that the expected waiting cost (or waiting time) function is convex in p. Whenever p_0 exceeds $p_{\min}$, the service capacity can be classified into three categories: (i) Security-favorable capacity, when $p_{\min} < p_0 \le p^*$ (the feasible p^* case). In this case, the service capacity of the system (represented by rate ν) compared to the inspection demand (λ) is plenty. A more than required proportion ($p^* > p_0$) of customers should be selected for further inspection to improve both the security level and customer service and no capacity expansion is necessary. (ii) Security-unfavorable capacity, when $p^* < p_0 < p_{\max}$ (the infeasible p^* case). In this case, the required p_0 can be implemented. But the minimum system delay (or the best customer service) of the TSCS is not achieved because of the required further security-check level and any increase in p for improving the security level will increase the expected waiting cost of all customers (or hurt the service quality). Capacity expansion is recommended in this case. (iii) Security-infeasible capacity, when $p_0 \ge p_{\max}$. In this case, implementing p_0 will lead to an unstable queue. Thus, a capacity expansion must be made immediately. Note that for a given p_0, an increase in service capacity (an increase in ν) may change the capacity status from (iii) to (ii) or from (ii) to (i). For a security-check system offering quality customer service, the manager should consider capacity expansion in case (ii) rather than wait until case (iii).

The performance properties discovered in the TSCS model help system managers to evaluate capacity expansion proposals. For example, in a security-unfavorable case ($p_0 > p^*$), the manager may consider if it is worthwhile to increase the service rate ν to change the system to a security-favorable case under a cost structure. This kind of decision is usually made based on multiple factors, such as raising the security level, improving customer service, meeting budget constraints, and acquiring facility and training personnel, etc. However, from the cost and performance perspective, we can use the results obtained to evaluate this decision. Let $\Delta\nu$ be the basic unit of service rate increase (e.g., adding an additional inspection station) at a cost denoted by κ. Suppose that a feasible increase in service rate $n\Delta\nu$ (n basic units increase) for the second-stage inspection has been proposed for a security-unfavorable case. This capacity increase can make a new waiting cost minimization $p^*_{\text{new}} > p_0$, where p^*_{new} is determined by (13), with the service rate of $\nu + n\Delta\nu$ as shown in Figure 4. Obviously, given that all other parameters are fixed, the second-stage service rate increase will lower the entire $E(WC)$ curve and move the optimal p^* to the right (a new higher workload-balance proportion occurs to shift more customers to the second stage). Thus, the system with the increased service capacity becomes the security favorable one. If the unit cost is smaller than the unit benefit (average waiting cost reduction) of this capacity expansion, or $\kappa < \lambda\{[E(WC)_{\text{old}} \text{ at } p_0] - [E(WC)_{\text{new}} \text{ at } p^*_{\text{new}}]\}/n$, where $[E(WC)_{\text{old}} \text{ at } p_0]$ and $[E(WC)_{\text{new}} \text{ at } p^*_{\text{new}}]$ are expected waiting costs per customer computed at service rates of ν and $\nu + n\Delta\nu$, respectively, based on (12), this proposed service capacity expansion is not only cost justifiable but also improves the security level from p_0 to p^*_{new}, with happier customers.

Figure 4 Increasing Further Inspection Service Rate to Change the Service Capacity Status



Note. From security-check infeasible p^* to security-check feasible p^*_{new}.

4. Numerical Illustrations

In this section, we show the accuracy of the proposed approximations, the comparison between different configurations of the security-check systems, the performance characteristics of more realistic TSCSs, and the practical value of the TSCS model.

Table 1 **Performance of a Single-Server TSCS with $\lambda = 8.5$, $\mu_1 = 20$, $\mu_2 = 15$, and $\nu = 8.7$**

p	$E(W_1^q)$	$E(W_2^q)_{\text{approx}}$	$E(W_2^q)_{\text{simu}}$	95% CI of $E(W_2^q)_{\text{approx}}$	Error of $E(W_2^q)_{\text{approx}}$ (%)	$E(W)$	$E(T)$
0.20	0.6094	0.0315	0.0311	(0.0303, 0.0320)	1.34	0.6157	0.7420
0.25	0.4722	0.0419	0.0410	(0.0382, 0.0439)	2.09	0.4827	0.6114
0.30	0.3787	0.0536	0.0525	(0.0493, 0.0558)	2.03	0.3947	0.5259
0.35	0.3108	0.0669	0.0632	(0.0598, 0.0670)	5.89	0.3342	0.4677
0.40	0.2592	0.0823	0.0783	(0.0709, 0.0856)	5.12	0.2921	0.4281
0.45	0.2188	0.1002	0.0949	(0.0862, 0.1036)	5.62	0.2639	0.4023
0.50	0.1862	0.1214	0.1135	(0.1051, 0.1220)	6.96	0.2469	0.3877
0.55	0.1594	0.1468	0.1407	(0.1303, 0.1510)	4.33	0.2401	0.3833
0.60	0.1369	0.1779	0.1685	(0.1566, 0.1800)	6.89	0.2436	0.3893
0.65	0.1178	0.2169	0.2031	(0.1874, 0.2189)	6.79	0.2588	0.4069
0.70	0.1014	0.2674	0.2544	(0.2352, 0.2736)	5.13	0.2887	0.4391
0.75	0.0872	0.3358	0.3190	(0.2876, 0.3504)	5.27	0.3390	0.4919
0.80	0.0747	0.4339	0.4218	(0.3769, 0.4667)	2.87	0.4218	0.5771

4.1. Accuracy of Approximations

The computational results are based on the two examples. Table 1 of the first example shows that the error of approximation to $E(W_2^q)$, defined as $\{E(W_2^q)_{\text{approx}} - E(W_2^q)_{\text{simu}}\}/E(W_2^q)_{\text{simu}}$, is no more than 6.89% for a wide range of p by comparing with simulation results generated from Arena models (95% confidence intervals are included based on 20 replications, with a run length of 15 hours and a time unit equal to a minute). Also listed are the overall expected customer waiting times $E(W) = E(W_1^q) + pE(W_2^q)_{\text{approx}}$ (excluding inspection times) and expected system times $E(T) = (1-p)E(T_1) + p\{E(W_1^q) + 1/\mu_1 + E(T_2)_{\text{approx}}\}$, which are convex functions of p. In Table 1, the optimal proportion for minimizing $E(W)$ (or $E(T)$) is $p^* = 0.55$, which can be called the economic inspection proportion. If the required security-check level p_0 is below that proportion, the system has a security-favorable capacity. Otherwise, the system has a security-unfavorable capacity, and a capacity expansion should be considered to improve both the security level and customer service. We have tested the accuracy of the approximations extensively and found very consistent results that confirm the robustness and effectiveness of the approximation approach.

4.2. Two-Stage vs. One-Stage Inspection

A question raised by the inspectors during this study is whether performing both primary and secondary checks at stage 1 improves the system performance. We believe that this is a general question about the configuration of the security-check system. To answer this question, we need to compare the TSCS to a one-stage security-check system (OSCS) with the same service capacity based on the nature of actual operations at a border crossing station. The OSCS shown in Figure 5 has been treated as an $M/PH/2$ queue whose performance measures can be obtained by modeling the system as a quasi-birth-and-death process (see the e-companion). The numerical example shows the TSCS's advantages of (i) improving overall performance when the security level is high and (ii) distinguishing between nonselected and selected customers in terms of offering differentiated customer services. Table 2 shows that at the optimal $p^* = 0.21$ for minimizing the expected system delay $E(T)$ (or $p^* = 0.22$ for minimizing $E(WC)$), the $E(T)$ (or $E(WC)$) of the TSCS is 57% (or 70%) better than that of the OSCS.

Another important finding is that, with the TSCS, the system remains stable for some higher p values that make the OSCS unstable. This observation implies that at the same service capacity, the TSCS configuration is more efficient than the OSCS in terms of service resource utilization when security levels become higher. Unlike TSCS, Table 2 shows that in the OSCS, increasing the security level p and reducing the expected system delay time $E(T)$ (or $E(WC)$) are always two conflicting goals, as $E(T)$ (or $E(WC)$) is increasing in p. These relationships are clearly illustrated in Figure 6, where $E(D_l)$ and $E(D_h)$ are the system times for nonselected and selected customers, respectively. Because of the separation of the two-stage inspections in a TSCS, managers can offer

Figure 5 **One-Stage Security-Check System Modeled as an $M/PH/2$ Type Queue with $\lambda = 52.8571$, $\mu_1 = 300$, $\mu_2 = 60$, $\nu = 11.25$, $h_1 = 3$, and $h_2 = 2$**

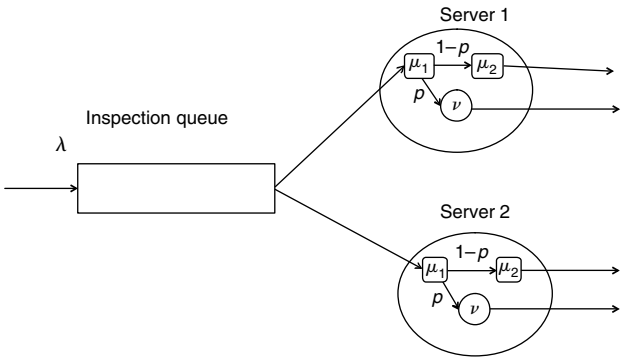


Table 2 Performance Comparison Between a TSCS and an OSCS with $\lambda = 52.8571$, $\mu_1 = 300$, $\mu_2 = 60$, $\nu = 15$, $h_1 = 3$, and $h_2 = 2$

p	$E(W_1^q)$	$E(W_2^q)_{\text{approx}}$	$E(W_2^q)_{\text{simu}}$	Error of $E(W_2^q)_{\text{approx}}$ (%)	$E(T)_{\text{TSCS}}$	$E(T)_{\text{OSCS}}$	$E(WC)_{\text{TSCS}}$	$E(WC)_{\text{OSCS}}$
0.12	0.3313	0.0537	0.0536	0.30	0.3638	0.0871	1.0768	0.2613
0.13	0.2774	0.0620	0.0640	−3.18	0.3119	0.0979	0.9191	0.2937
0.14	0.2378	0.0714	0.0766	−6.86	0.2748	0.1107	0.8050	0.3322
0.15	0.2075	0.0822	0.0871	−5.68	0.2473	0.1262	0.7197	0.3786
0.16	0.1836	0.0947	0.0994	−4.69	0.2268	0.1452	0.6544	0.4356
0.17	0.1642	0.1095	0.1139	−3.87	0.2113	0.1691	0.6041	0.5072
0.18	0.1482	0.1271	0.1314	−3.28	0.2001	0.2000	0.5654	0.6001
0.19	0.1348	0.1484	0.1516	−2.09	0.1925	0.2418	0.5366	0.7253
0.20	0.1233	0.1748	0.1786	−2.10	0.1883	0.3012	0.5166	0.9036
0.21	0.1135	0.2084	0.2069	0.74	**0.1877**	0.3926	0.5054	1.1778
0.22	0.1049	0.2525	0.2509	0.65	0.1914	0.5515	**0.5040**	1.6545
0.23	0.0973	0.3130	0.3055	2.46	0.2008	0.8950	0.5150	2.6849
0.24	0.0906	0.4011	0.3845	4.33	0.2189	1.8935	0.5443	5.6804

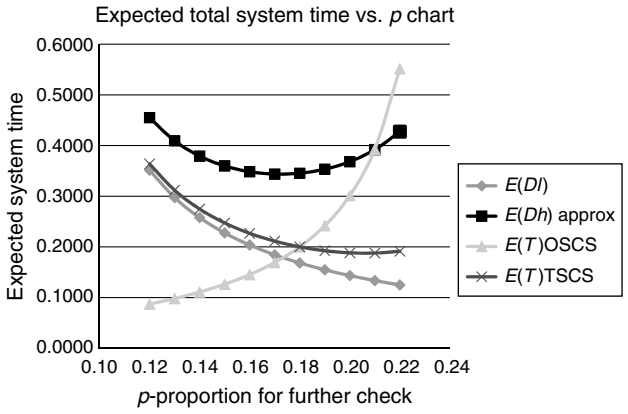
Note. The bold numbers represent the minimums of the performance measures.

different customer service levels for different stages or customer classes by assigning different waiting cost parameters h_i' or h_i values. In contrast, in an OSCS system, all customers are treated with the same waiting cost parameter h. In Table 2, $E(WC)$ for the TSCS is computed by using $h_1 = 3$ and $h_2 = 2$ and $E(WC)$ for the OSCS is computed by using $h = 3$.

4.3. Performance of General TSCS

Because most security-check systems in practice are multiserver waiting lines with generally distributed inspection durations, we need to explore whether the proposed approximations are robust enough in a more general setting and whether the performance properties discovered in the stylized model still hold for these more complex practical situations. Specifically, we check (1) if the approximation still works when the inspection durations are not exponentially distributed or when both stages becomes multiserver queues; and (2) if the security-check feasible optimal p^* obtained from the stylized single-server model also gives a "close-to-optimal" p for a multiserver border crossing inspection setting with equivalent service rates (i.e., using the single-server queue to approximate the multiserver queue by adjusting the service rates).

Figure 6 Comparison Between TSCS Performance and OSCS Performance for a System with the Same Service Capacity

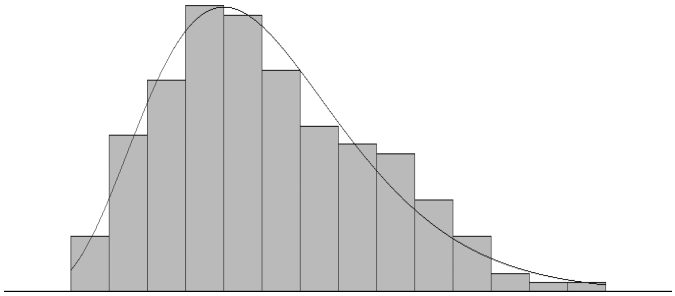


In our stylized model, the inspection process is modeled as the two-phase Coxian distribution. However, because of the customer selection procedure, the real distribution for the first phase of stage 1 is close to a combination of a relatively short and less variable duration (selected by random number generation or nonselected for low-risk customers) and a relatively long exponentially distributed duration (selected or nonselected by a series of screening questions). This fact implies that the actual distribution for the first phase is not exactly exponential and thus has a coefficient of variation smaller than 1 (CV < 1). Based on some real data sets collected at the border crossing stations and summarized in Figure 7, we find that the coefficient of variation is usually between 0.3 and 0.6, and a Erlang-k distribution with k between 3 and 7 is a good fit.

Note that the $A(s)$ approximation in (6) does not require phase 1 and 2 durations to be exponentially distributed. We tested the key approximation to $E(W_2^q)$ for the case with the same parameters as in Table 1 except the Erlang-6 distributed phase 1 duration with a rate of 1/120 and $k = 6$ (the mean of the phase 1 remains to be 1/20). Note that the coefficient of variation for phase 1 duration of Erlang-6 is 0.41, which is very close to the value computed based on the real data sets. The comparison between the approximation and simulated results for this case is presented in Table 3, with 95% confidence interval for $E(W_2^q)$. Surprisingly, our approximation produces the results with extremely good accuracy (even better than the exponentially distributed phase 1 case), which indicates the robustness of the TSCS model.

We have also tested the cases with multiple servers at stage 1 and found that the approximation with equivalent service rates works well for almost all

Figure 7 Fit All Options to a Data Set of Phase 1 (Initial Screening) Durations at a Border Crossing Station



Data summary

Number of data points	= 200
Min data value	= 6.32
Max data value	= 55.7
Sample mean	= 24.3
Sample std. dev.	= 9.81

Distribution summary

Distribution:	Erlang
Expression:	ERLA (4.06, 6)
Square error:	0.001284

Chi-square test

Number of intervals	= 9
Degrees of freedom	= 6
Text statistic	= 4.26
Corresponding *p*-value	= 0.644

Function	Square error
Erlang	0.00128
Beta	0.0016
Lognormal	0.0021
Triangular	0.00341
Normal	0.00499
Uniform	0.0345
Gamma	0.103
Exponential	0.115

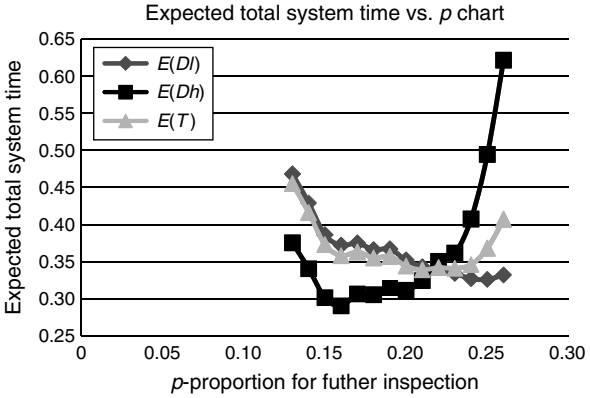
cases, implying that the proposed approximation (9) is robust with respect to the number of servers at stage 1. Larger errors occur only for cases with a large number of servers $c \geq 20$ and higher p values. However, for these cases, the Poisson process approximation with rate λp gives excellent accuracy. As the number of servers increases and p becomes higher, the arrival process can be considered as the sum of many independent arrival processes, which approaches the Poisson process. A strategy for evaluating the performance of a multiserver, two-stage security-check system is proposed as follows: For small to moderate c cases (border crossing stations), the first-stage inspection can be analyzed by using an M/Cox2/c via the matrix analytic method, and the second-stage queue can be treated as a $GI/M/k$ model with the LST of the interarrival time $A(s)$ obtained from the stylized single-server TSCS model under the service rate equivalency condition. For large c cases, the first stage can be analyzed by using either a heavy-traffic limit approach (Whitt 2002) or a hybrid approximation approach, which is a combination of the $M/G/1$ and $M/G/\infty$ systems based on Tijms' (1994) assumptions. The second-stage queue can be treated as an $M/M/k$ queue.

Now we check if the optimal p^* determined by (13) can be used for a general multiserver inspection system with the equivalent service rates at both stages. It is worth noting that unless the traffic is very heavy and the number of the servers is small, the single-server system may not produce very good approximations to the performance measures, such

Table 3 Approximation to $E(W_2^q)$ in a TSCS with $\lambda = 8.5$, Erlang-6 Phase 1 Inspection of Rate 20/6, and Exponential Phase 2 Inspection at Stage 1 and Exponential Further Inspection at Stage 2 of Rates $\mu_2 = 15$ and $\nu = 8.7$

p	$E(W_2^q)_{\text{approx}}$	$E(W_2^q)_{\text{simu}}$	95% CI $E(W_2^q)_{\text{simu}}$	Error of $E(W_2^q)_{\text{approx}}$
0.20	0.0299	0.0292	(0.0265, 0.0319)	2.28
0.25	0.0397	0.0410	(0.0382, 0.0439)	−3.18
0.30	0.0508	0.0525	(0.0493, 0.0558)	−3.17
0.35	0.0636	0.0632	(0.0599, 0.0664)	0.58
0.40	0.0783	0.0783	(0.0710, 0.0856)	−0.06
0.45	0.0954	0.0949	(0.0862, 0.1036)	0.51
0.50	0.1156	0.1135	(0.1051, 0.1220)	1.87
0.55	0.1399	0.1407	(0.1303, 0.1510)	−0.54
0.60	0.1697	0.1664	(0.1566, 0.1763)	2.01
0.65	0.2072	0.2031	(0.1874, 0.2189)	2.00
0.70	0.2557	0.2544	(0.2352, 0.2736)	0.52
0.75	0.3214	0.3190	(0.2879, 0.3504)	0.75
0.80	0.4156	0.4218	(0.3769, 0.4667)	−1.47

Figure 8 **Fifteen Servers at Stage 1 and Two Servers at Stage 2, Poisson Arrivals with Rate $\lambda = 52.8571$, Erlang-6 Phase 1 Service with Rate 20/6, Erlang-4 Phase 2 Service with Rate 4/6 at Stage 1, and Erlang-8 Further Inspection with Rate 7.5/8 at Stage 2**



as $E(W)$ or $E(T)$ for a multiserver system under the equivalent service rates. Thus, we do not expect to use our stylized model to compute good approximations to the performance measures for a general multiserver TSCS here. Instead, we investigate if the performance characteristics, such as the convexity of the $E(T)$ function and the optimal p^*, discovered from our stylized model still exist in general security-check systems. We consider a TSCS with multiple servers at both stages and nonexponentially distributed service durations. The performance measures generated from simulations are shown in Figure 8. It is interesting to see that the $E(T)$ has the convex shape, and the simulated optimal $p^* = 0.22$ is very close to the optimal $p^* = 0.21$, which is determined by (13) with the service rate equivalence of $\mu_1^{\text{single-server}} = c\mu_1^{\text{multiserver}}$ and $\mu_2^{\text{single-server}} = c\mu_2^{\text{multiserver}}$ at stage 1 and $\nu^{\text{single-server}} = k\nu^{\text{multiserver}}$ at stage 2. This implies that the optimal p^* is insensitive to the number of servers and the distributions of the inspection durations under the equivalent rate condition. We have tested various complex examples extensively and found the consistent results. Therefore, the stylized TSCS model with the equivalent arrival and service rates can be used to estimate the "close-to-optimal" proportion for further inspection in practical security-check systems.

4.4. Performance Improvement in Border Inspections

To demonstrate the practical value of the TSCS model, we examine the positive effects of applying the optimal p^* suggested by the model in a real system setting. We take advantage of the availability of border crossing traffic data, as the CascadeGatewayData.com website provides a highly functional and flexible interface to an archive of vehicle volume, wait time, and queue length data from the Cascade Gateway land port of entry between Whatcom County, Washington State, and the Lower Mainland of British Columbia (http://www.wcog.org/Data.aspx). Our example is based on the real traffic data of a typical day from this website. Figure 9 shows the time-varying arrival pattern and the average waiting time on a typical Friday going southbound. Such a pattern can be effectively treated by our TSCS model via the stationary independent period by period (SIPP) approximation (see Green et al. 1991). We are only concerned with the two heavy traffic periods—12:00 to 15:00 and 17:00 to 20:00—because congestion is not an issue for the light traffic periods.

Figure 10 presents the performance curves for these two periods computed from the TSCS model. With the security data set used in §2.2, we obtain the minimum $p_0 = 0.109$ for achieving $P(FC) \le 0.001$. In practice, a $0.11 \le p_0 \le 0.12$ was used and our TSCS model in Figure 10 shows the average waiting times at these practical p_0s are between 88 and 50 minutes for the period of 12:00 to 15:00 and between 48 and 33 minutes for the period of 17:00 to 20:00, respectively. These average waiting times are consistent with the actual recorded average waiting times of 54 minutes for

Figure 9 **Pacific Highway Crossing Station—Southbound Traffic Data on July 9, 2010**

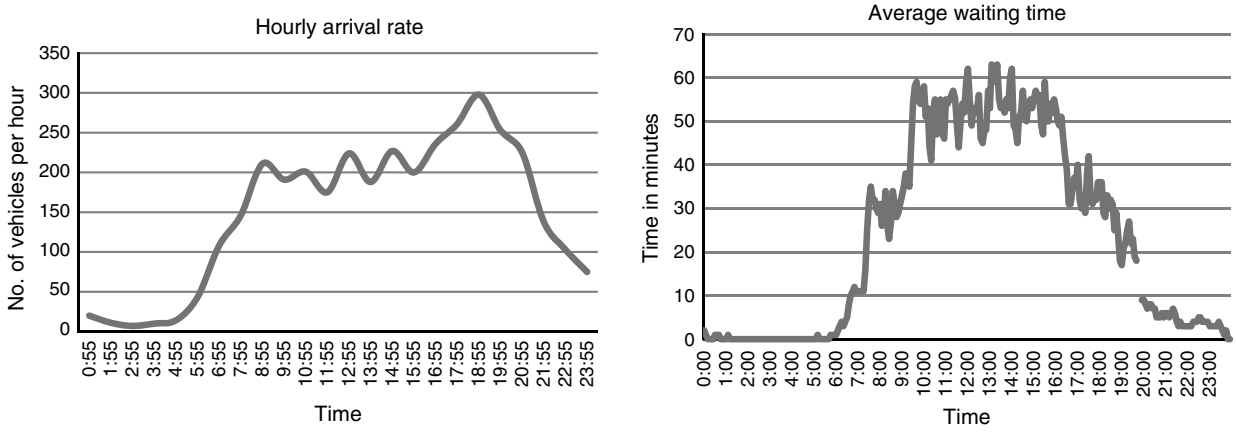
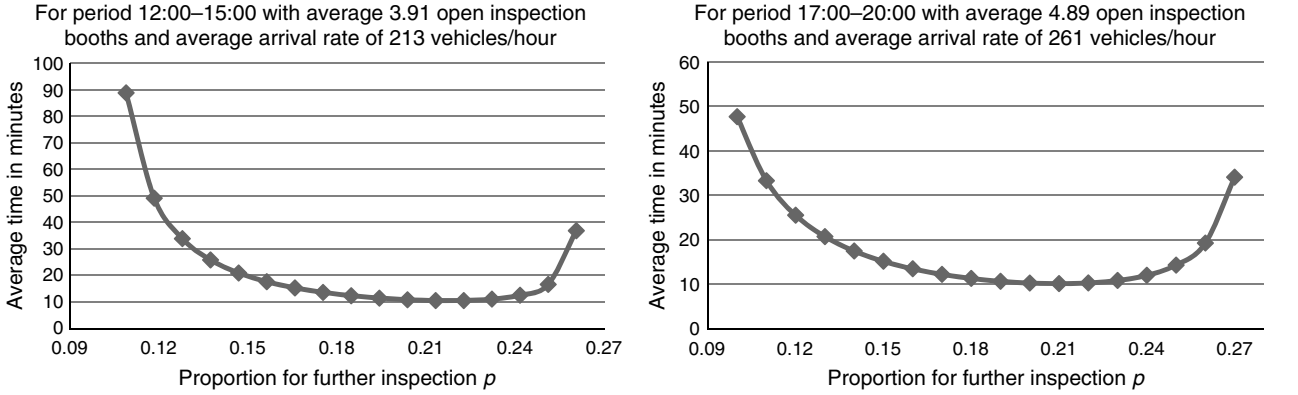


Figure 10 Average Waiting Time as a Function of p for Two Heavy Traffic Periods on July 9, 2010, at Pacific Highway South Crossings with Service Rates of 50 Vehicles per Hour for Stage 1 and 15 Vehicles per Hour for Stage 2



12:00 to 15:00 and 34 minutes for 17:00 to 20:00 in the right graph of Figure 9. This observation indicates that using the SIPP gives acceptable accuracy for the nonstationary arrival process at a border crossing system. The main reason for that is because the relative amplitude (the ratio of the maximum to the average) of the arrival rate, a measure of the degree of nonstationarity, is small over the planning period, and the average service time is very small relative to the length of the planning period (see Green et al. 1991). Of course, SIPP may not work well in other time varying situations; see Green et al. (2001) for some representative examples.

The optimal p^*s obtained from the TSCS model are 0.21 and 0.22 for the two periods of interest. Both optimal proportions for further inspection significantly reduce the average waiting times to about 10 minutes for these two periods, as shown in Figure 10. Note that the performance characteristics of the TSCS model have been verified by extensive simulations. In addition, using the optimal p^* can improve the security screening level, as $P(FC)$ is reduced from 0.000998 at $p_0 = 0.11$ to 0.000867 at $p^* = 0.21$. Another implication is that the optimal p^* is relatively insensitive to the arrival rate change (almost the same p^* for the two periods) because of the congestion-based staffing (adjusting the number of open booths according to the queue length; see Zhang 2009). Thus, both the customer service and the security effectiveness can be significantly improved with the application of our TSCS model in the time-varying arrival situation. We have carried out similar tests on data for the congested hours of all weekends from June to December 2010, and the results are all consistent with this specific example. The main reason these results are consistent is that the traffic patterns are very similar over these time periods.

5. Summary and Discussion

In this paper, we have studied a two-stage security-check system and examined the trade-off between maximizing the security screening level and minimizing the expected customer delay, the two goals of most security-check systems. A stylized two-stage single-server model is analyzed to discover some insightful performance properties of the inspection policy implemented in practice. We utilized the Coxian-2 distribution to capture the main characteristics of the first-stage inspection and the second-stage arrival process. Some simple, accurate, and robust approximations were developed and verified by simulations for a wide range of parameters and more generally distributed inspection durations.

We have obtained several important managerial insights: (1) To achieve a sufficient security screening level, the minimum proportion of selected customers for further inspection (p_0) can be determined by setting a guaranteed high probability of "true alarm" or low probability of "false clear." (2) The expected waiting cost function is shown to be convex in p. Based on this convexity property, TSCS service capacities can be classified into security-favorable, security-unfavorable, or security-infeasible categories. For a security-favorable capacity, the security and customer service goals are consistent and customers are happier with a higher inspection level than p_0. For a security-unfavorable capacity, the two goals are in conflict and the manager should consider a service capacity expansion if he or she wants to achieve a higher security level without annoying customers. For the security-infeasible capacity, a service capacity expansion must be made immediately for meeting the required security level. (3) With the traffic statistics and service rates, we can use the stylized model to estimate the "close-to-optimal" p for a practical and more complex security-check system. (4) By comparing the two-stage system with the one-stage system, we confirmed the advantages of using the

two-stage selective inspection system. These advantages include distinguishing customers with different inspection levels in offering inspection services and more efficiently utilizing the service capacity to support higher security-check levels. Clearly, these managerial insights help inspectors and administrators in planning the service capacity and designing the inspection policy to achieve both the effective security screening and good customer service in any security-check systems with similar settings. We have also shown the predicted improvement in the system performance of a practical system by using the recommended policy with real traffic data sets.

We tested the robustness of the renewal arrival process approximation for the second stage and found that this approximation provides more accurate results in an Erlang-k first phase than in an exponential first-phase case. Unfortunately, we did not come up with a good explanation for this phenomenon. Therefore, further investigation is needed and can be a direction of future research. We have focused on only two classes of customers (selected and nonselected) in this paper. However, our model and analysis can be utilized in the case with multiple classes of customers. If customers can be classified into $k > 2$ categories and customers of class i $(2 \le i \le k)$ are required to go through a type i further inspection, the first stage and a second-stage type i inspection can be analyzed by our model. A detailed investigation on the multiserver TSCS with multiple classes of customers is left as another future research topic.

Acknowledgments

The authors are grateful to the department editor, Assaf Zeevi, the anonymous associate editor, and four referees for their constructive comments, which led to the significant improvement of this paper in numerous ways. The first author is grateful for the support received from Natural Sciences and Engineering Research Council of Canada Grant RGPIN197319. The second and third authors are thankful for partial support from the National Science Council of Taiwan under Grant NSC 98-2221-E-004-001-MY2.

Appendix. Proofs of Propositions and Remarks

Proof of Proposition 1

Rewrite $\alpha(p) = [(p - p_c)/p]\tau + (p_c/p)\gamma$ as

$$\alpha(p) = \tau + \frac{(\gamma - \tau)p_c}{p}.$$

Using $\tau = P(T) = \alpha(p)p + \beta(p)(1 - p)$ and the expression of $\alpha(p)$ above, we obtain

$$\beta(p) = \tau - \frac{(\gamma - \tau)p_c}{1 - p}.$$

Note that p_c is a constant and $\gamma > \tau$. Thus part (i) follows immediately.

Using the expressions of $\alpha(p)$ and $\beta(p)$ above and the definition of $P(TA)$, we have

$$P(TA) = \tau\theta_{FI} + (\theta_{FI} - \theta_{FI^c})(\gamma - \tau)p_c + (\theta_{FI} - \theta_{FI^c})\tau p.$$

From this expression of $P(TA)$, part (ii) of the proposition follows from $\theta_{FI} > \theta_{FI^c}$ and $\gamma > \tau$. □

Proof of Proposition 2

For a two-stage inspection policy, the maximum proportion of customers for further inspection, denoted by $p_{\max}$, can be determined by the stability condition of the second-stage queue. That is $\lambda p/\nu < 1$. Thus, $p_{\max} = \nu/\lambda$. The minimum proportion, denoted by $p_{\min}$, can be determined by the stability condition of the first-stage queue; that is, $\lambda E(S_1) = \lambda[1/\mu_1 + (1 - p)(1/\mu_2)] < 1$. From this condition, we obtain $p_{\min} = 1 - \mu_2(1/\lambda - 1/\mu_1)$. Under conditions (i) and (ii), we can show that $p_{\max} > p_{\min}$; therefore, there exits a feasible p that makes both stage queues stable. □

Proof of Proposition 4

Letting $M(s) = X(s)[\rho_1 + (1 - \rho_1)I(s)]$, $X_2(s) = \mu_2/(\mu_2 + s)$, and $N(s) = 1/M(s)$, $A(s)$ of (6) can be written as $A(s) = p/\{N(s) - (1 - p)X_2(s)\}$. Taking the first-order derivative of $A(s)$ with respect to s, we have $A'(s) = -p\{N'(s) - (1 - p)X_2'(s)\}/\{N(s) - (1 - p)X_2(s)\}^2$. Evaluating $-A'(0)$ by using $N'(0) = 1/\lambda - (1 - p)(1/\mu_2)$ and $-X_2'(0) = 1/\mu_2$, we obtain $E(A) = -A'(0) = 1/(\lambda p) = E(A^{\exp})$. Now taking the second derivative of $A(s)$, we get

$$\begin{aligned} A''(s) = {} & \frac{2pN'(s)\{N'(s) - (1-p)X_2'(s)\}}{\{N(s) - (1-p)X_2(s)\}^3} \\ & - \frac{pN''(s)}{\{N(s) - (1-p)X_2(s)\}^2} + \frac{p(1-p)X_2''(s)}{\{N(s) - (1-p)X_2(s)\}^2} \\ & - \frac{2p(1-p)X_2'(s)\{N'(s) - (1-p)X_2'(s)\}}{\{N(s) - (1-p)X_2(s)\}^3}. \end{aligned}$$

Evaluating $A''(0)$ by using $N'(0), X_2'(0)$ as stated above, $N''(0) = 2(1 - p)^2(1/\mu_2^2) - 2(1 - p)(1/(\lambda\mu_2)) + 2(1 - p)(1/(\mu_1\mu_2))$, and $X_2''(0) = 2/\mu_2^2$, and after some algebraic manipulation, we have

$$\begin{aligned} E(A^2) = {} & A''(0) \\ = {} & \frac{2}{(p\lambda)^2} + \frac{2(1-p)}{\mu_2^2} + \frac{2(1-p)}{p\mu_2}\left(\frac{1}{\lambda} - \frac{1}{\mu_1}\right) + \frac{4(1-p)}{p^2\lambda\mu_2} \\ > {} & \frac{2}{(p\lambda)^2}, \end{aligned}$$

which implies that $Var(A) > Var(A^{\exp})$. □

Proof of Proposition 5

Denote $E(W_1^q)$ by $f(p)$ a function of p small and write it as $f(p) = (\alpha - \beta p)/(\varphi + \psi p)$, where $\alpha = \lambda/\mu_1^2 + \lambda/\mu_2^2 + \lambda/(\mu_1\mu_2)$, $\beta = \lambda/\mu_2^2 + \lambda/(\mu_1\mu_2)$, $\varphi = 1 - \lambda/\mu_1 - \lambda/\mu_2$, and $\psi = \lambda/\mu_2$. Taking the first-order derivative of $f(p)$ with respect to p, we get $df(p)/dp = f'(p) = -(\varphi\beta + \alpha\psi)/(\varphi + \psi p)^2 < 0$. Thus, $E(W_1^q)$is a decreasing function in p. Taking the second-order derivative of $f(p)$ with respect to p, we obtain $d^2f(p)/dp^2 = f''(p) = 2(\varphi\beta\psi + \alpha\psi^2)/(\varphi + \psi p)^3 > 0$. Therefore, $E(W_1^q)$ is a convex function in p. Furthermore, $E(T_1)$, as a constant plus a convex decreasing functions, is also decreasing and convex in p. □

Proof of Proposition 6

Let $E(T_2)_{\text{approx}} = (1/2)g(p)$, where $g(p) = 1/(k - lp) + 1/(\nu - \lambda p)$, where $k = \nu(1-a)$, $l = b\nu$. Taking the first-order derivative of $g(p)$ with respect to p, we have $dg(p)/dp = l/(k - lp)^2 + \lambda/(\nu - \lambda p)^2 > 0$. Thus, $g(p)$ is an increasing function of p and so is $E(T_2)_{\text{approx}}$. Now taking the second-order derivative of $g(p)$ with respect to p, we get $d^2g(p)/dp^2 = 2l^2/(k - lp)^3 + 2\lambda^2/(\nu - \lambda p)^3 > 0$. Therefore, $E(T_2)_{\text{approx}}$ is convex in p. Furthermore, $E(W_2^q) = E(T_2) - 1/\nu$ is also increasing and convex in p. □

Proof of Proposition 7

Part (ii) of the proposition follows easily by taking the second-order derivative of $E(WC)^{\text{stage}}_{\text{approx}}$ with respect p and using Propositions 5 and 6.

For part (i), based on $E(WC)_{\text{approx}}$ in (12),

$$E(WC)_{\text{approx}} = (1-p)E(T_1)h_1 + p[E(W_1^q) + \frac{1}{\mu_1} + E(T_2)_{\text{approx}}]h_2. \tag{14}$$

For simplicity, we rewrite $E(WC)_{\text{approx}}$ as $EWC(p)$ here. From (14), it can be derived that

$$\begin{aligned}\frac{\partial EWC(p)}{\partial p} &= -E(T_1)h_1 + (1-p)h_1\frac{\partial E(T_1)}{\partial p} \\ &\quad + h_2\left[E(W_1^q) + \frac{1}{\mu_1} + E(T_2)_{\text{approx}}\right] \\ &\quad + p\left(\frac{\partial E(W_1^q)}{\partial p} + \frac{\partial E(T_2)_{\text{approx}}}{\partial p}\right)h_2 \end{aligned} \tag{15}$$

and

$$\begin{aligned}&\frac{\partial^2 EWC(p)}{\partial p^2} \\ &= \left(2(h_2 - h_1)\frac{\partial E(W_1^q)}{\partial p}\right) + (1-p)h_1\frac{\partial^2 E(T_1)}{\partial p^2} \\ &\quad + 2h_2\frac{\partial E(T_2)_{\text{approx}}}{\partial p} + p\left(\frac{\partial^2 E(W_1^q)}{\partial p^2} + \frac{\partial^2 E(T_2)_{\text{approx}}}{\partial p^2}\right)h_2. \end{aligned} \tag{16}$$

The first term of (16) is nonnegative because it holds that $h_1 \geq h_2$ and $\partial E(W_1^q)/\partial p < 0$ (from Proposition 5). Also based on Propositions 5 and 6, the remaining three terms of (16) are all nonnegative. Therefore, we prove that $\partial^2 EWC(p)/\partial p^2 \geq 0$ if $h_1 \geq h_2$. □

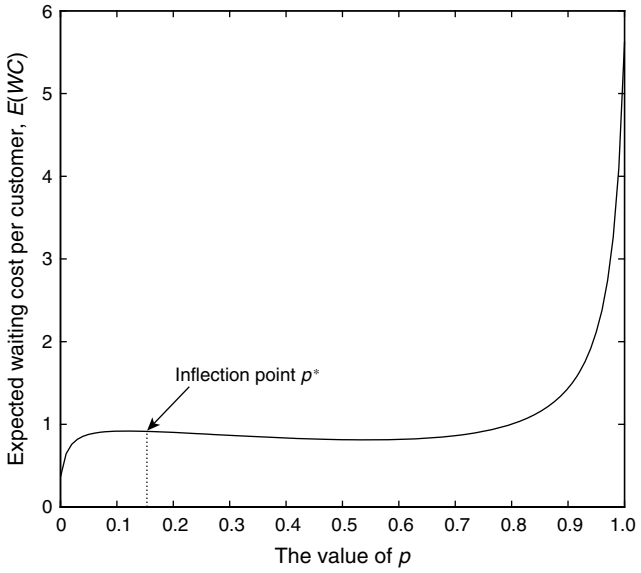
Remarks on the Convexity of $E(WC)_{\text{approx}}$ for $h_1 < h_2$ Cases

Rewrite (16) as

$$\begin{aligned}\frac{\partial^2 EWC(p)}{\partial p^2} &= 2h_2\left(\frac{\partial E(W_1^q)}{\partial p} + \frac{\partial E(T_2)_{\text{approx}}}{\partial p}\right) \\ &\quad - 2h_1\frac{\partial E(W_1^q)}{\partial p} + (1-p)h_1\frac{\partial^2 E(T_1)}{\partial p^2} \\ &\quad + p\left(\frac{\partial^2 E(W_1^q)}{\partial p^2} + \frac{\partial^2 E(T_2)_{\text{approx}}}{\partial p^2}\right)h_2. \end{aligned}$$

Note that except for the first term in the expression above, all terms are nonnegative. To make $\partial^2 EWC(p)/\partial p^2$ negative, the first term must be a very large negative term to dominate all remaining three positive terms. This is possible only when $h_2 \gg h_1$ and $-\partial E(W_1^q)/\partial p \gg \partial E(T_2)_{\text{approx}}/\partial p$, which means that the marginal reduction in expected waiting time at stage 1 must be much greater than the marginal increase in the expected system time at stage 2 because of a unit increase in p. Numerical tests indicate that the nonconvexity exists in some extreme cases where h_2 must be much larger than h_1. Here is an example with the parameters from the example of Figure 2: the arrival rate $\lambda = 8.5$, the phase 1 service rate of stage 1 server $\mu_1 = 20$, the phase 2 service rate of stage 1 server $\mu_2 = 15$, the service rate of stage 2 server $\nu = 8.7$, the constant $a = 0.0418$ and $b = 0.7034$ (for estimating root r_0). We choose $h_1 = 0.01$ and $h_2 = 3$ (300 times of h_1). The function of approximate expected waiting cost $E(WC)_{\text{approx}}$ is shown in Figure A.1. It can be observed that there exists an inflection point p^* on the function $E(WC)_{\text{approx}}$, where p^* is close to 0.15. The function $E(WC)_{\text{approx}}$ is concave in the interval $(0, p^*)$ and convex in the interval $(p^*, 1)$. We also tested other examples, and the nonconvexity may exist only when h_2 is more than 300 times larger than h_1. Therefore, for the case of $h_1 < h_2$, the ratio h_1/h_2 (and its relation to other parameters) affects the convexity of $E(WC)_{\text{approx}}$. It has been observed that $E(WC)_{\text{approx}}$ may remain convex in p for the entire feasible value range, even for $h_1 < h_2$ case, as long as the ratio h_1/h_2 is not too small (almost all practical situations). For example, $E(WC)_{\text{approx}}$ is still convex for all $0 < p < 1$ if we set $h_1 = 1$ and $h_2 = 1.5$ for all examples tested. Therefore, we can safely assume that the $E(WC)_{\text{approx}}$ is convex function for all practical situations.

Figure A.1 **Approximate Expected Waiting Cost $E(WC)_{\text{approx}}$ as Waiting Cost $h_1 < h_2$**



Note. The parameters are taken from Figure 2.

References

Babu, V. L. L, R. Batta, L. Lin. 2006. Passenger grouping under constant threat probability in an airport security system. *Eur. J. Oper. Res.* **168**(2) 633–644.

Bureau of Transportation Statistics. 2006. Prohibited items intercepted at airport screening checkpoints. http://www.bts.gov/publications/national_transportation_statistics/html/table_02_16b.html.

El-Taha, M., B. Maddah. 2006. Allocation of service time in a multiserver system. *Management Sci.* **52**(4) 623–637.

Green, L. V., P. J. Kolesar, A. Svoronos. 1991. Some effects of nonstationary on multiserver Markovian queueing systems. *Oper. Res.* **39**(3) 502–511.

Green, L. V., P. J. Kolesar, J. Soares. 2001. Improving the SIPP approach for staffing service systems that have cyclic demands. *Oper. Res.* **49**(4) 549–564.

Gross, D., C. Harris. 1996. *Fundamentals of Queueing Theory*, 2nd ed. Wiley, New York.

Jacobson, S. H., J. E. Kobza, A. S. Easterling. 2001. A detection theoretic approach to modeling aviation security problems using knapsack problem. *IIE Trans.* **33**(9) 747–759.

Kobza, J. E., S. H. Jacobson. 1997. Probability models for access security system architectures. *J. Oper. Res. Soc.* **48**(3) 255–263.

Latouche, G., V. Ramaswami. 1999. *Introduction to Matrix Analytic Methods in Stochastic Modelling*. ASA/SIAM, Philadelphia.

Lee, A. J., L. A. McLay, S. H. Jacobson. 2009. Designing aviation security passenger screening systems using nonlinear control. *SIAM Control Optim.* **48**(4) 2085–2105.

McLay, L. A., S. H. Jacobson, J. E. Kobza. 2006. A multilevel passenger screening problem for aviation security. *Naval Res. Logist.* **53**(3) 183–197.

McLay, L. A., A. J. Lee, S. H. Jacobson. 2010. Risk-based policies for airport security checkpoint screening. *Transportation Sci.* **44**(3) 333–349.

Nikolaev, A. G., S. H. Kacobson, L. A. McLay. 2007. A sequential stochastic security system design problem for aviation security. *Transportation Sci.* **41**(2) 182–194.

Ross, S. M. 1997. *Introduction to Probability Models*, 6th ed. Academic Press, San Diego.

Shumsky, R., E. Pinker. 2003. Gatekeepers and referrals in services. *Management Sci.* **49**(7) 839–856.

Tijms, H. C. 1994. *Stochastic Models: An Algorithmic Approach*. John Wiley & Sons, New York.

Whitt, W. 1982. Approximating a point process by a renewal process, I: Two basic methods. *Oper. Res.* **30**(1) 125–147.

Whitt, W. 2002. *Stochastic-Process Limits*. Springer, New York.

Wilson, D., E. K. Roe, S. A. So. 2006. Security checkpoint optimizer (SCO): An application for simulating the operations of airport security checkpoints. *Proc. 2006 Winter Simulation Conf., Monterey*, CA. 529–535.

Zhang, Z. G. 2009. Performance analysis of a queue with congestion staffing policy. *Management Sci.* **55**(2) 240–251.