



## Operations Research

Publication details, including instructions for authors and subscription information:  
<http://pubsonline.informs.org>

### When Service Times Depend on Customers' Delays: A Relationship Between Two Models of Dependence

Chenguang (Allen) Wu, Achal Bassamboo, Ohad Perry

To cite this article:

Chenguang (Allen) Wu, Achal Bassamboo, Ohad Perry (2021) When Service Times Depend on Customers' Delays: A Relationship Between Two Models of Dependence. *Operations Research*

Published online in *Articles in Advance* 01 Dec 2021

. <https://doi.org/10.1287/opre.2021.2179>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact [permissions@informs.org](mailto:permissions@informs.org).

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2021, INFORMS

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.


For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

## Methods

# When Service Times Depend on Customers' Delays: A Relationship Between Two Models of Dependence

Chenguang (Allen) Wu,<sup>a</sup> Achal Bassamboo,<sup>b</sup> Ohad Perry<sup>c</sup>

<sup>a</sup>Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong; <sup>b</sup>Kellogg School of Management, Northwestern University, Evanston, Illinois 60208; <sup>c</sup>Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208

Contact: allenwu@ust.hk,  <https://orcid.org/0000-0002-2528-0286> (C(A)W); a-bassamboo@kellogg.northwestern.edu,  <https://orcid.org/0000-0001-7758-4751> (AB); ohad.perry@northwestern.edu,  <https://orcid.org/0000-0002-4584-3015> (OP)

Received: March 7, 2020

Revised: January 10, 2021; April 21, 2021

Accepted: July 25, 2021

Published Online in Articles in Advance:  
December 1, 2021

Subject Classifications: queues: applications;  
balking and reneging

Area of Review: Stochastic Models

<https://doi.org/10.1287/opre.2021.2179>

Copyright: © 2021 INFORMS

**Abstract.** As empirically observed in restaurants, call centers, and intensive care units, service times needed by customers are often related to the delay they experience in queue. Two forms of dependence mechanisms in service systems with customer abandonment immediately come to mind: First, the service requirement of a customer may evolve while waiting in queue, in which case the service time of each customer is *endogenously* determined by the system's dynamics. Second, customers may arrive (*exogenously*) to the system with a service and patience time that are stochastically dependent, so that the service-time distribution of the customers that end up in service is different than that of the entire customer population. We refer to the former type of dependence as *endogenous* and to the latter as *exogenous*. Because either dependence mechanism can have significant impacts on a system's performance, it should be identified and taken into consideration for performance-evaluation and decision-making purposes. However, identifying the source of dependence from observed data is hard because both the service times and patience times are censored due to customer abandonment. Further, even if the dependence is known to be exogenous, there remains the difficult problem of fitting a joint service-patience times distribution to the censored data. We address these two problems and provide a solution to the corresponding statistical challenges by proving that both problems can be avoided. We show that, for any exogenous dependence, there exists a corresponding endogenous dependence, such that the queuing dynamics under either dependence have the same law. We also prove that there exist endogenous dependencies for which no equivalent exogenous dependence exists. Therefore, the endogenous dependence can be considered as a generalization of the exogenous dependence. As a result, if dependence is observed in data, one can always consider the system as having an endogenous dependence, regardless of the true underlying dependence mechanism. Because estimating the structure of an endogenous dependence is substantially easier than estimating a joint service-patience distribution from censored data, our approach facilitates statistical estimations considerably.

**Funding:** C. A. Wu received financial support from the Hong Kong Research Grant Council [Early Career Scheme, Project 26206419]. A. Bassamboo and O. Perry received partial financial support from the National Science Foundation [Grant CMMI 2006350].

**Keywords:** service systems • dependence of service times on delay • censored data

## 1. Introduction

Human behavior has significant impacts on the queuing dynamics in service systems. For example, in many service systems, customers abandon the queue if they deem their waiting time to be too long. Another important phenomenon, which has only started to receive attention recently, is that service times of customers may depend on their delay in queue. This phenomenon is well known to hold in certain healthcare settings (each minute of delay can be detrimental for patients having a stroke or a heart attack, thus

substantially affecting treatment times; see Chan et al. 2017) and has also been empirically observed in other contexts, such as call centers (see Reich et al. 2010) and restaurants (see De Vries et al. 2018).

We consider two different underlying mechanisms that lead to such dependence. Under the first mechanism, the service requirement of a customer evolves while waiting in queue to be served. In this case, customers do not have a specific service-time distribution, but, rather, a *conditional service-time distribution*, which depends on the delay they experience before

their service begins. Thus, the actual service time of a customer who ends up receiving service (and does not abandon the queue) is endogenized by the system's dynamics; we therefore refer to this type of dependence as *endogenous*. In the second mechanism, the dependence of the service times on the delays is induced by a joint service-patience distribution that can be thought of as given exogenously to the system (customers "arrive exogenously to the system" with their bivariate service and patience times). Specifically, the patience and service requirements of each customer are dependent (e.g., in grocery stores, customers with many items tend to be more patient in the checkout line compared with those with few items), so that the service-time distribution of customers who do not abandon the queue is different than that of the entire customer population; we refer to this second type of dependence as *exogenous*.

It is significant that either type of dependence can have substantial impacts on the system's dynamics, and therefore on its performance, and on related operational decisions, such as staffing. This is clear for systems with endogenous dependence (e.g., consider the case in which service times of delayed customers are substantially longer, or shorter, than those of customers who are not delayed) and was demonstrated via a fluid model and simulation experiments for systems with exogenous dependence in Wu et al. (2019). Further, Wu et al. (2019) show that the performance of a system with exogenous dependence depends heavily on the full joint distribution of the service and patience times, and not only on the marginal distributions and their correlation. In turn, optimal staffing decisions depend on this information as well; see Wu et al. (2019, section 6).

Unfortunately, data of service and patience times are necessarily censored due to customer abandonment because observations exist only for the service times of customers who *did not* abandon and for the patience of customers who *did* abandon. This censoring leads to two statistical challenges: First, it requires efficient econometric methods to identify whether the observed dependence is exogenous or endogenous. Second, even if the dependence can be identified, or is believed to be exogenous, there remains the difficult task of fitting a joint service-patience times distribution to the censored data.

In general, estimating bivariate distributions under censoring is a hard problem; see, for example, Lopez and Saint-Pierre (2012). Reich et al. (2010) suggest nonparametric methods to estimate the exogenous dependence, which do not work well for customers with short patience times, for which unrealistic negative service times are predicted. Other estimation procedures were proposed in the literature of survival analysis. Unfortunately, the dependent random variables

observed in this setting are typically censored *simultaneously* (e.g., times in which a couple withdrew from a study), so that the proposed estimation methods are not appropriate for our needs. Dabrowska et al. (1988) and Akritas and Keilegom (2003) develop nonparametric methods to handle bivariate censoring, but, as mentioned in Lopez and Saint-Pierre (2012), those methods have significant drawbacks: They either do not define a true joint distribution, require a careful choice of smoothing parameters, or make additional assumptions on the censoring conditions, so that the proposed methods are again not useful for our needs. Parametric procedures for estimating censored joint distributions have also been considered. Unfortunately, such methods impose stringent assumptions on the bivariate distribution one wishes to estimate and may not perform well when prior knowledge regarding that distribution is unavailable.

In this paper, we provide a solution to the aforementioned statistical challenges by proving that both can be effectively avoided. Specifically, we show that for any exogenous dependence (with joint service-patience distribution), there exists a unique endogenous dependence, such that the queuing dynamics under either dependence mechanism are the same (have the same law), for all arrival rates and staffing levels. We also prove that the class of systems with endogenous dependencies is larger than that with exogenous dependencies, in the sense that there exist endogenous dependence mechanisms for which no equivalent exogenous dependence exists. Therefore, from the point of view of the queuing dynamics (transient and stationary), the endogenous dependence can be considered as a generalization of the exogenous dependence.

Our results demonstrate that, regardless of the true dependence mechanism, and regardless of whether data are available for a desired arrival process or staffing level, both the transient and the stationary behavior of the system can be treated as if the dependence is endogenous. That endogenous dependence can be estimated from available data, even if it was collected for different arrival processes and staffing levels than those we are interested in. Hence, the difficult dependence-identification problem can be avoided, because a system can always be modeled as having an endogenous dependence. Furthermore, the problem of fitting a joint distribution for the service and patience times is avoided as well and is replaced by the substantially easier task of estimating the structure of endogenous dependence (which is either the true dependence in the system or is equivalent to the exogenous dependence in the system). Thus, for distribution-fitting purposes, we advocate that *the* system should be considered as having an endogenous dependence, even if it is somehow known to possess an exogenous dependence.

We remark that our model of endogenous dependence is related to the literature on deteriorating jobs, which studies jobs that deteriorate while waiting, leading to longer processing times of those jobs; see Sugawa and Takahashi (1965), Glazebrook (1992), Browne and Yechiali (1990), and Mosheiov (1991). Motivated mostly by applications in manufacturing, the goal is to develop scheduling policies to process a fixed number of jobs. However, the models considered in this literature have no randomness in the arrival process and no abandonment of jobs from the queue, and are therefore not appropriate for service systems. Other related works on dependencies in queues include Whitt (1990) and Boxma and Vlasiov (2007), both deriving the waiting-time distribution in single-server queues when service and interarrival times depend linearly on the delay.

## 2. The Setting

We consider a service system with  $n \geq 1$  statistically identical agents that are dedicated to serving statistically homogeneous customers that arrive to the system in accordance with a simple counting stochastic process (namely, customers arrive one at a time). We let  $A(t)$  denote the number of customers that arrive by time  $t$ . A customer begins service with an agent immediately upon arrival, if an agent is available, and otherwise waits in queue. We assume customers are served in the order of arrival. Each customer has a finite patience for waiting; if the waiting time in queue exceeds that patience time, the customer abandons the queue. The key feature of the systems we consider is that the service time of each customer depends on the delay that customer experiences. This dependence is either endogenous, in the sense that the service-time distribution is a function of the delay in queue, or is exogenous and induced by a self-selection mechanism of customers under a common joint service-patience times distribution.

### 2.1. The Two Dependence Mechanisms

We consider two types of systems, one with exogenous dependence and the other with endogenous dependence, and refer to each type (with a slight abuse of language) simply as “the system with exogenous (endogenous) dependence.” Throughout the paper, we use superscripts “*ex*” and “*en*” to distinguish between entities (random variables, stochastic processes, etc.) corresponding to systems with exogenous and endogenous dependencies, respectively.

In the system with exogenous dependence, customer  $i$  arrives with a service time  $S_i^{ex}$  and a patience time  $T_i^{ex}$ , which are dependent random variables. The bivariate random variables  $\{(S_i^{ex}, T_i^{ex}) : i \geq 1\}$  are jointly continuous with a joint probability density function

(pdf)  $f^{ex}$ , independent across customers and independent of the system’s state; see Bassamboo and Randhawa (2015) and Wu et al. (2019). We let  $f_T^{ex}$  denote the marginal pdf of  $T_i^{ex}$ .

In the system with endogenous dependence, we use  $T_i^{en}$  to denote the patience time of customer  $i$ , and we assume that  $\{T_i^{en} : i \geq 1\}$  are independent and identically distributed (i.i.d.) continuous random variables that are independent from all other random variables in the queuing system. We denote the cumulative distribution function (cdf) and pdf of  $T_i^{en}$  by  $F_T^{en}$  and  $f_T^{en}$ , respectively, with  $\bar{F}_T^{en} \triangleq 1 - F_T^{en}$  denoting the corresponding complementary cdf (ccdf). The service-time distribution of each customer in this system depends on the delay that customer experiences in queue. Specifically, let  $Z_i^{en}$  denote the *offered wait* of customer  $i$ , representing the virtual waiting time of that customer, namely, the time he would wait if his patience was infinite.

The service times of arriving customers are described by a stochastic process,  $\{S_i^{en}(Z_i^{en}) : i \geq 1\}$ , where  $S_i^{en}(Z_i^{en})$  denotes a random variable representing customer  $i$ ’s “virtual” service time, given that his offered wait is  $Z_i^{en}$ . (We write virtual service time because the customer may abandon and not receive service.) This family of service-time distributions captures the evolution of customers’ service times as they wait in queue. We assume that  $\{S_i^{en}(z) : i \geq 1\}$  are independent across customers, are identically distributed for each value  $z$  of  $Z_i^{en}$ , and are also independent of all other random variables comprising the system. Let  $\Psi^{en}$  denote the virtual service-time distribution of a customer, conditioned on that customer’s offered wait, namely,  $\Psi^{en}(x, z) \triangleq P(S_i^{en} \leq x \mid Z_i^{en} = z)$ . Define the conditional ccdf  $\bar{\Psi}^{en}(x, z) \triangleq 1 - \Psi^{en}(x, z)$ .

**Assumption 1.**  $\Psi^{en}(x, z)$  is differentiable in  $x$  and in  $z$ .

Under Assumption 1, the pdf of the virtual service time exists and satisfies  $\psi^{en}(x, z) = \frac{\partial \Psi^{en}(x, z)}{\partial x}$ .

### 2.2. Systems’ Dynamics

Because we do not assume that the arrival process  $A(t)$  is Poisson and that the service and the patience times are exponentially distributed (or that the service time of a customer is independent of his patience in the system with exogenous dependence), the number-in-system process is non-Markov. (See evidence of nonexponential service and patience times in Brown et al. 2005 and Mandelbaum and Zeltyn 2004.) A Markovian representation for the queuing dynamics is achieved by keeping track, at each time  $t \geq 0$ , of the remaining time until the next arrival, as well as the remaining service time of each customer in service and the elapsed waiting time of each customer that is waiting in queue.



Specifically, for  $\mathcal{M} \in \{ex, en\}$  and  $t \geq 0$ , let  $Y^{\mathcal{M}}(t) \in \{0, 1, \dots, n\}$  and  $Q^{\mathcal{M}}(t) \in \mathbb{N}_+ \triangleq \{0, 1, \dots\}$  denote the number of customers in service and in queue, respectively. Let  $B^{\mathcal{M}}(t)$  denote the forward recurrence time for the arrival process  $A(t)$ , namely, with  $\kappa_i$  being the time of the  $i$ th arrival after time 0,  $B^{\mathcal{M}}(t) = \kappa_{A(t)+1} - t$ . Let  $U^{\mathcal{M}}(t) \in \mathbb{R}_+^n$  denote the remaining service times of customers in service, sorted in increasing order. Specifically, if  $Y^{\mathcal{M}}(t) < n$ , so there are  $n - Y^{\mathcal{M}}(t)$  idle servers, then for  $j \in \{1, \dots, n - Y^{\mathcal{M}}(t)\}$ , let  $U_j^{\mathcal{M}}(t)$ , the  $j$ th entry of  $U^{\mathcal{M}}(t)$ , be zero; for  $j \in \{n - Y^{\mathcal{M}}(t) + 1, \dots, n\}$ , let  $U_j^{\mathcal{M}}(t)$  be the  $(Y^{\mathcal{M}}(t) + j - n)$ th (weakly) smallest remaining service time of customers in service. Similarly, let  $V^{\mathcal{M}}(t) \in \mathbb{R}_+^{\infty}$  denote the elapsed waiting times of customers in queue, sorted in decreasing order (this leads to ranking waiting customers in ascending order of their arrivals). If  $Q^{\mathcal{M}}(t) \geq 1$ , then for  $j \in \{1, \dots, Q^{\mathcal{M}}(t)\}$ , let  $V_j^{\mathcal{M}}(t)$  be the  $j$ th (weakly) largest elapsed waiting time of customers in queue; for  $j > Q^{\mathcal{M}}(t)$ , let  $V_j^{\mathcal{M}}(t) = 0$ . Otherwise, if  $Q^{\mathcal{M}}(t) = 0$ , then let all entries of  $V^{\mathcal{M}}(t)$  be zero. In general, the entries of  $U^{\mathcal{M}}(t)$  and  $V^{\mathcal{M}}(t)$  can be ordered in an arbitrary manner; we choose the orderings above because they are easy to interpret. Then,

$$\mathbf{X}^{\mathcal{M}}(t) = \left( Y^{\mathcal{M}}(t), Q^{\mathcal{M}}(t), U^{\mathcal{M}}(t), V^{\mathcal{M}}(t), B^{\mathcal{M}}(t) \right), \quad t \geq 0,$$

is a Markov process, describing the queuing dynamics of system  $\mathcal{M}$ .

**Remark 1.** An alternative Markov representation for a system without dependence can be achieved by tracking the remaining time to abandon of each customer waiting in queue. It is significant that this alternative representation cannot be employed in our setting, because the information regarding the actual waiting time of each customer is required in order to determine the service-time distribution of that customer.

### 3. Main Results

To formally state our main result, we define an equivalence relation between exogenous and endogenous dependencies; see Definition 1. We first note that, in addition to the arrival process  $A$  and the number of agents  $n$ , the queuing dynamics in a system with exogenous dependence are completely determined by the joint service-patience distribution  $f^{ex}$ , whereas the dynamics in a system with endogenous dependence are completely determined by the patience distribution  $f_T^{en}$  together with the virtual service-time distribution  $\psi^{en}$ .

Let  $\mathcal{L}^{ex}(A, n; f^{ex})$  denote the probability law of  $\mathbf{X}^{ex}$  in the system with exogenous dependence characterized by  $(A, n, f^{ex})$ ; similarly, let  $\mathcal{L}^{en}(A, n; f_T^{en}, \psi^{en})$  denote the probability law of  $\mathbf{X}^{en}$  in the system with endogenous dependence described by  $(A, n; f_T^{en}, \psi^{en})$ . Here, by “the law of  $\mathbf{X}^{\mathcal{M}}$ ,” we mean, as usual, the joint distribution

of the family of random variables  $\mathbf{X}^{\mathcal{M}}(t)$ , indexed by  $t \geq 0$ , which is determined by the family of finite-dimensional distributions of  $\mathbf{X}^{\mathcal{M}}$ ; see, for example, Billingsley (2013) and Whitt (2002). Let  $\stackrel{d}{=}$  denote equality in distribution.

**Remark 1.** We say that an exogenous dependence with a joint pdf  $f^{ex}$  is equivalent to an endogenous dependence with  $(f_T^{en}, \psi^{en})$ , if  $\mathcal{L}^{ex}(A, n; f^{ex}) = \mathcal{L}^{en}(A, n; f_T^{en}, \psi^{en})$ , for any arrival process  $A$  and capacity  $n$ , whenever  $\mathbf{X}^{ex}(0) \stackrel{d}{=} \mathbf{X}^{en}(0)$ .

We now state the main result of the paper.

**Theorem 1.** For any exogenous dependence characterized by a joint pdf  $f^{ex}$ , there is an equivalent endogenous dependence  $(f_T^{en}, \psi^{en})$ . Further, the two dependencies are related via

$$f_T^{en}(z) = f_T^{ex}(z) \quad \text{and} \quad \psi^{en}(x, z) = \frac{\int_z^{\infty} f^{ex}(x, y) dy}{\int_z^{\infty} f_T^{ex}(y) dy}, \quad (1)$$

for all  $x, z \geq 0$ .

To prove Theorem 1, we define a discrete-time offered-wait process, denoted by  $\{Z_i^{\mathcal{M}}, i \geq 1\}$ , such that  $Z_i^{\mathcal{M}}$  is customer  $i$ 's offered wait. We characterize the offered-wait process  $\{Z_i^{\mathcal{M}}, i \geq 1\}$  by describing an auxiliary  $n$ -dimensional stochastic process  $\{\mathbf{Z}_i^{\mathcal{M}}, i \geq 1\}$ , which tracks the residual service times at other servers not serving customer  $i$ . Imagine that  $n - 1$  virtual customers with infinite patience arrive simultaneously with customer  $i$ , whom we index by  $\{2, 3, \dots, n\}$ . Denoting the  $k$ th entry of  $\mathbf{Z}_i^{\mathcal{M}}$  by  $Z_i^{\mathcal{M}}(k)$ , let  $Z_i^{\mathcal{M}}(1)$  be the offered wait of the actual customer  $i$ —that is,  $Z_i^{\mathcal{M}}(1) = Z_i^{\mathcal{M}}$ . For  $2 \leq k \leq n$ , we stipulate that virtual customer  $k$  should be served by a distinct server other than those who have already served the real customer  $i$  and virtual customers  $2, \dots, k - 1$ , and we let  $Z_i^{\mathcal{M}}(k)$  be his waiting time. In other words,  $Z_i^{\mathcal{M}}(k) - Z_i^{\mathcal{M}}(k - 1)$  represents the time between the  $(k - 1)$ st and  $k$ th distinct servers becoming available.

The representation of the offered-wait process follows Moyal (2019), employing the sorting (reordering) operator  $\mathcal{R}: \mathbb{R}_n^+ \mapsto \mathbb{R}_n^+$ , defined as follows. Let  $x_{(j)}$  denote the  $j$ th (weakly) smallest component of the vector  $x \in \mathbb{R}^n$ :  $x_{(j-1)} \leq x_{(j)} \leq x_{(j+1)}$ ,  $2 \leq j \leq n - 1$ . Then,  $\mathcal{R}(x) = (x_{(1)}, \dots, x_{(n)})$ —that is,  $\mathcal{R}$  sorts each vector  $x$  in increasing order, with  $\mathcal{R}_1(x)$  being the smallest and  $\mathcal{R}_n(x)$  being the largest components of  $x$ .

To achieve a unified representation for the process  $\{Z_i^{\mathcal{M}}, i \geq 0\}$ , we employ the notation  $S_i^{\mathcal{M}}(\mathbf{Z}_i^{\mathcal{M}}(1))$  for  $\mathcal{M} = en$ , as well as for  $\mathcal{M} = ex$ . This notation is clear for the endogenous case ( $\mathcal{M} = en$ ), because the service time of each customer is a function of the delay. It is redundant for the exogenous case ( $\mathcal{M} = ex$ ), but can be justified by treating  $S_i^{ex}$  as a constant function of  $\mathbf{Z}_i^{ex}(1)$ , namely,  $S_i^{ex}(z) \equiv S_i^{ex}$  for all  $i \geq 1$  and  $z \geq 0$ . With

this notation, the offered-wait process evolves according to the following recursive formula:

$$\mathbf{Z}_{i+1}^{\mathcal{M}} = \mathcal{R}([\mathbf{Z}_i^{\mathcal{M}} + \mathbb{I}_{\{Z_i^{\mathcal{M}}(1) \leq T_i^{\mathcal{M}}\}} S_i^{\mathcal{M}}(\mathbf{Z}_i^{\mathcal{M}}(1)) \mathbf{e}_1 - \alpha_{i+1}^{\mathcal{M}} \mathbf{1}]^+),$$

$$\mathcal{M} \in \{en, ex\}, \quad (2)$$

where  $\mathbf{e}_1 \triangleq (1, 0, \dots, 0)$ ,  $\mathbf{1} \triangleq (1, 1, \dots, 1)$ , and  $\alpha_{i+1}^{\mathcal{M}}$  is the time between the arrivals of customers  $i$  and  $i + 1$ , namely,  $\alpha_{i+1}^{\mathcal{M}} = \kappa_{i+1}^{\mathcal{M}} - \kappa_i^{\mathcal{M}}$ . The initial value  $\mathbf{Z}_1^{\mathcal{M}}$  is fully characterized by  $\mathbf{X}^{\mathcal{M}}(0)$ .

We are now prepared to prove Theorem 1. The proof follows two steps: a “construction step,” in which we construct two systems jointly via a coupling argument, such that one of the systems has the endogenous dependence, and the coupled system has the same dynamics as the endogenous one; and a “verification step,” in which we show that the coupled system constructed in the first step has the law of a system with exogenous dependence. Because the verification step requires a tedious computation, we omit it in the proof for brevity.

**Proof of Theorem 1.** As described above, we first couple two systems on the same probability space, among which one has an endogenous dependence specified in Section 2.1, such that both systems have the same sample paths with probability one (w.p.1.). We then argue that the second system (that is coupled with the endogenous one) has the desired exogenous dependence, therefore proving the claim of the theorem. The proof focuses on establishing an equivalence between two systems initialized empty and can be easily extended to prove systems with arbitrary initial conditions. We use a  $\sim$  (tilde) to denote the stochastic processes and random variables on the new probability space.

For each sample path describing the dynamics of a system with an endogenous dependence, we construct a coupled system in the following steps.

We first generate  $(\tilde{t}_i^{en}, \tilde{s}_i^{en}(\cdot))$  and  $\tilde{\alpha}_i^{en}$  for each new arrival in the system with endogenous dependence. The offered-wait process for these new arrivals is fully characterized by the Recursion (2). We next construct a coupled system and use superscript “c” to denote the random variables in it.

i. Set  $\tilde{\alpha}_i^c = \tilde{\alpha}_i^{en}$ , so the  $i$ th customer arrives at the same time in the system with endogenous dependence and the coupled system. Hence, both systems have the same realized arrival process.

ii. If  $\tilde{t}_i^{en} \leq \tilde{z}_i^{en}$ , then the  $i$ th customer abandons in the system with endogenous dependence. Set  $\tilde{T}_i^c = \tilde{t}_i^{en}$  and generate  $\tilde{S}_i^c$  from the density

$$\frac{f^{ex}(\cdot, \tilde{t}_i^{en})}{f_T^{ex}(\tilde{t}_i^{en})},$$

namely, from the conditional distribution of  $S_i^{ex}$  conditioned on  $T_i^{ex} = \tilde{t}_i^{en}$ .

iii. If  $\tilde{t}_i^{en} > \tilde{z}_i^{en}$ , then the  $i$ th customer is served in the system with endogenous dependence and requires a service time  $\tilde{s}_i^{en}(\tilde{z}_i^{en})$ . Set  $\tilde{S}_i^c = \tilde{s}_i^{en}(\tilde{z}_i^{en})$  and generate  $\tilde{T}_i^c$  from the density

$$\frac{f^{ex}(\tilde{s}_i^{en}(\tilde{z}_i^{en}), \cdot) \mathbb{I}_{\{\cdot > \tilde{z}_i^{en}\}}}{\int_{\tilde{z}_i^{en}}^{\infty} f^{ex}(\tilde{s}_i^{en}(\tilde{z}_i^{en}), y) dy},$$

namely, from the conditional distribution of  $T_i^{ex}$  conditioned on  $S_i^{ex} = \tilde{s}_i^{en}(\tilde{z}_i^{en})$  and  $T_i^{ex} > \tilde{z}_i^{en}$ .

Steps (i)–(iii) above guarantee the coupled system has exactly the same dynamics as the system with endogenous dependence we start with. Specifically, we can first argue that  $\tilde{z}_i^{en} = \tilde{z}_i^c$  for all  $i$ . This is because step (iii) guarantees that each customer, if he enters service (when  $\tilde{z}_i^{\mathcal{M}} < \tilde{t}_i^{\mathcal{M}}$ , for  $\mathcal{M} \in \{en, c\}$ ), requires the same service time in both the coupled system and the system with endogenous dependence, namely,  $\tilde{s}_i^{en}(\tilde{z}_i^{en}) = \tilde{s}_i^c$ . Then, using (2), it follows by induction that the auxiliary processes  $\{\tilde{\mathbf{Z}}_i^{\mathcal{M}} : i \geq 1\}$  are the same in both systems, which implies  $\tilde{z}_i^{en} = \tilde{z}_i^c$  because  $\tilde{z}_i^{\mathcal{M}} = \tilde{\mathbf{Z}}_i^{\mathcal{M}}(1)$ . This further implies  $\tilde{\mathbf{X}}^{en} = \tilde{\mathbf{X}}^c$  path by path. For example, consider the queue process  $\{\tilde{Q}^{\mathcal{M}}(t), t \geq 0\}$ . We have  $\tilde{Q}^{\mathcal{M}}(t) = \sum_{\tilde{\kappa}_i^{\mathcal{M}} \leq t} \mathbb{I}_{\{t - \tilde{\kappa}_i^{\mathcal{M}} < \min(\tilde{t}_i^{\mathcal{M}}, \tilde{z}_i^{\mathcal{M}})\}}$ , where  $\tilde{\kappa}_i^{\mathcal{M}} = \sum_{j=1}^i \tilde{\alpha}_j^{\mathcal{M}}$  is the arrival time of the  $i$ th customer. Because  $\tilde{\kappa}_i^{en} = \tilde{\kappa}_i^c$  by step (i), it follows that  $\tilde{Q}^{en}(t) = \tilde{Q}^c(t)$  for all  $t$ . Similarly, we can verify the other component processes in  $\tilde{\mathbf{X}}^{en}$  and  $\tilde{\mathbf{X}}^c$  are equal w.p.1.

Using basic computations, we can show that the coupled system has the desired exogenous dependence, consistent with the one described in Section 2.1. Specifically, it can be shown that  $\{(\tilde{S}_i^c, \tilde{T}_i^c) : i \geq 1\}$  generated in steps (i)–(iii) are i.i.d. and have the same joint distribution  $f^{ex}$ . Because the law of the process  $\mathbf{X}^{ex}$  is uniquely determined by  $f^{ex}$ , given  $A$  and  $n$ , it follows that the coupled system has the same queuing dynamics as the system with exogenous dependence, which further implies  $\mathcal{L}^{ex}(A, n; f^{ex}) = \mathcal{L}^{en}(A, n; f_T^{en}, \psi^{en})$ .  $\square$

The proof of Theorem 1 shows that the two dependence mechanisms impact the queuing dynamics only via the conditional service-time distribution conditioned on the waiting times, namely, via  $P(S_i^{\mathcal{M}} \leq s | T_i^{\mathcal{M}} > Z_i^{\mathcal{M}}, Z_i^{\mathcal{M}} = z)$ . Now, if  $f_T^{en}$  and  $\psi^{en}$  in an endogenous dependence satisfy (1), then

$$\bar{\Psi}(x, z) \bar{F}_T^{en}(z) = \int_{v=z}^{\infty} \int_{u=x}^{\infty} f^{ex}(u, v) du dv,$$

which is decreasing in  $x$  and in  $z$ . As we show in Theorem 2, this latter monotonicity property also implies that the endogenous dependence has an equivalent exogenous dependence. To prove this result, we need

the following lemma, whose proof appears in the appendix.

**Lemma 1.** For  $k = 1, 2$  let  $f_T^k$  denote a pdf of the patience-time distribution and  $\psi^k$  denote the conditional pdf of the virtual waiting time. Then,  $\mathcal{L}^{en}(A, n; f_T^1, \psi^1) = \mathcal{L}^{en}(A, n; f_T^2, \psi^2)$  for any arrival process  $A$ , capacity  $n$ , and initial condition  $\mathbf{X}^{en}(0)$  if and only if  $(f_T^1, \psi^1) = (f_T^2, \psi^2)$ .

Lemma 1 shows that if two systems with endogenous dependencies have the same queuing dynamics under any arbitrary arrival process, capacity, and initial conditions, then the two dependencies must be characterized by the same patience-time and conditional service-time distributions. Because each exogenous dependence has an equivalent endogenous dependence, Lemma 1 implies that the latter equivalent endogenous dependence is *unique*. Moreover, Lemma 1 allows us to characterize the condition for an endogenous dependence to have an equivalent exogenous dependence.

**Theorem 2.** An endogenous dependence with  $(f_T^{en}, \psi^{en})$  is equivalent to some exogenous dependence  $f^{ex}$  if and only if  $\bar{\Psi}^{en}(x, z)\bar{F}_T^{en}(z)$  is decreasing in  $x$  and in  $z$ .

Theorem 1 implies that the family of systems with exogenous dependence forms a subclass of the family of systems with endogenous dependence, in the sense that the queue process under any exogenous dependence is equal in distribution to the queue process under some (specific) endogenous dependence. Theorem 2 further implies this subclass is *proper*, because there exist endogenous dependencies for which the condition in Theorem 2 fails to hold, so that no equivalent exogenous dependence exists.

**Example 1.** Consider the following endogenous-dependence model: The patience times are exponentially distributed with mean  $1/\gamma$ , and the service times are, conditional on the offered wait being  $z$ , exponentially distributed with mean  $1/\mu(z)$ . Then,  $\bar{\Psi}^{en}(x, z)\bar{F}_T^{en}(z) = \exp(-(\gamma z + \mu(z)x))$  which is decreasing in  $x$ , but is not necessarily decreasing in  $z$  when  $\mu(z)$  is strictly decreasing in  $z$ . For example, consider  $\mu(z) = \max\{\bar{a} - bz, \underline{a}\}$ , where  $\bar{a} > \underline{a} > 0$ . Then, for small  $z$  such that  $\bar{a} - bz > \underline{a}$ , it holds that

$$\begin{aligned} \bar{\Psi}^{en}(x, z)\bar{F}_T^{en}(z) &= \exp(-(\gamma z + (\bar{a} - bz)x)) \\ &= \exp(-(\bar{a}x + (\gamma - bx)z)). \end{aligned}$$

For  $x > \gamma/b$ ,  $\bar{\Psi}^{en}(x, z)\bar{F}_T^{en}(z)$  is increasing in  $z$  for  $z < (\bar{a} - \underline{a})/b$ , thus violating the condition in Theorem 2.

It is significant that estimating an endogenous dependence is, in general, simpler than estimating an exogenous dependence. Therefore, for distribution-fitting purposes, we advocate that the system should be considered as having an endogenous dependence, even if it is known to possess an exogenous dependence.

To elaborate on this latter point, note that under the endogenous dependence, the problem of fitting a two-sided censored distribution is replaced by the problem of estimating the univariate patience distribution, for which (unlike the multivariate case) an efficient estimator exists, as well as estimating the uncensored service-time distribution corresponding to each delay time.

## 4. A Simple Estimation Procedure for Systems with Dependence

We now propose a simple procedure to estimate endogenous dependencies. First, we use the Kaplan-Meier (K-M) estimator (Kaplan and Meier 1958) to estimate the patience-time distribution (e.g., Zohar et al. 2002). Let  $N$  denote the number of customers in the sample and  $J$  denote the number of customers that abandoned the queue, so that  $N - J$  is the number of customers that received service. We rank the waiting times of the *abandoned* customers in increasing order,  $0 \triangleq t_0 < t_1 < t_2 < \dots < t_j < t_{j+1} \triangleq \infty$ . The K-M estimator for  $F_T^{en}$  is then

$$\hat{F}_T^{en}(x) = \prod_{t_j \leq x} \left( 1 - \frac{\# \text{ customers who abandon at } t_j}{\# \text{ customers who have not abandoned by } t_j} \right). \quad (3)$$

Second, the conditional service-time distribution can be estimated from the service times observed from *served* customers. For  $i = 1, \dots, N$ , let  $a_i$  denote whether customer  $i$  was served:  $a_i = 1$  if the customer was served and  $a_i = 0$  otherwise. If  $a_i = 1$ , let  $s_i$  and  $w_i$  be the service and waiting times observed from that customer. We split the observations of served customers into  $M$  separate bins  $\{\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_M\}$  based on their waiting times. Each bin is set to be a disjoint interval such that there are sufficiently many observations in each bin to allow for an estimator that approximates  $\Psi^{en}$  well. For each  $z \in \mathbb{R}_+$ , find the bin  $\mathcal{B}_m$ ,  $1 \leq m \leq M$  containing  $z$ . We use the following empirical service-time distribution to approximate  $\Psi^{en}(x, z)$ ,

$$\hat{\Psi}^{en}(x, z) = \frac{\sum_{i=1}^N \mathbb{I}\{a_i = 1, w_i \in \mathcal{B}_m, s_i \leq x\}}{\sum_{i=1}^N \mathbb{I}\{a_i = 1, w_i \in \mathcal{B}_m\}}. \quad (4)$$

As the length of each bin is set to be sufficiently small, we expect the approximation in (4) to be close enough to the true  $\Psi^{en}$ . We leave the rigorous statistical analysis for future research.

### 4.1. Implementation of the Proposed Estimation Procedure

We demonstrate via simulations that the proposed estimation procedure is effective, even when the actual dependence in the system is exogenous. We start by simulating two systems initialized empty, each having

**Table 1.** Performance Comparison of Estimation Procedures

Agents	Queue length			Throughput		
	Actual	Endogenous	Independent	Actual	Endogenous	Independent
Positive dependence, $r = 0.5$						
8	12.37	-3.0%	-12.9%	5.81	+2.6%	+13.8%
10	7.74	+0.6%	-1.4%	8.13	-0.7%	+0.9%
12	3.64	+4.1%	+31.0%	10.19	-0.8%	-5.3%
Negative dependence, $r = -0.3$						
8	5.75	+0.0%	+20.2%	9.12	+0.3%	-6.0%
10	3.37	-0.2%	+5.5%	10.31	+0.2%	-0.7%
12	1.76	-0.4%	-13.3%	11.12	+0.2%	+1.2%

*Notes.* The steady-state performance metrics are computed by taking averages of 500 independent runs, each lasting 3,000 time units with the first 1,000 time units serving as a warm-up period. The 95% confidence interval half-width is less than 0.3% for all reported metrics.

$n = 10$  agents, a Poisson arrival process with rate 12, and marginal service and patience times that are exponentially distributed with means one and two, respectively. Both systems have an exogenous dependence: one with a positive dependence and one with a negative dependence between the service and patience times. We use Gaussian copulas to generate the joint service-patience distributions (see Wu et al. 2019, appendix A for background). We denote by  $r$  the correlation of the generated service and patience times. We take  $r = -0.3$  and  $r = 0.5$  to represent a negative and a positive dependence, respectively. For each system, we simulate one sample path over 50,000 time units and collect the service and waiting times of customers served by an agent, as well as the patience times of customers who abandoned the queue. We employ (3) and (4) to estimate the patience-time distribution and the conditional service-time distribution for equivalent endogenous dependencies.

We compare our estimations to an estimation procedure that ignores the dependence, namely, a procedure that treats the service times as being independent of all other random variables in the system, as well as the system’s state. In the latter estimation procedure, we use the K-M estimator (3) to estimate the patience-time distribution. Because the service times are assumed to be i.i.d. and independent of the waiting times, we estimate the (unconditional) service-time distribution using the empirical distribution for the service times observed from served customers.

To demonstrate the performance of the estimations for the equivalent endogenous dependencies as opposed to the estimations that ignore the dependencies, we vary the number of agents while keeping the arrival process fixed. Specifically, we first simulate systems with exogenous dependencies to produce the true steady-state metrics. We then fix the estimations for the equivalent endogenous dependencies, as well as the estimations that ignore the dependencies, both obtained from the observations produced by a system

with an exogenous dependence and 10 agents. We then simulate systems with the two different estimations, varying the number of agents from 8 to 12.

We compare in Table 1 the steady-state queue length and throughput (the average number of service completions per unit time) of the systems with true exogenous dependencies (“Actual” column), the systems with estimations for the equivalent endogenous dependencies (“Endogenous” column), and the systems with estimations that ignore the dependencies (“Independent” column).

We find that treating the stochastic processes as being independent when they are, in fact, dependent can lead to substantial errors in estimations and predictions. In contrast, the estimation procedure for equivalent endogenous dependence performs relatively well, despite its simplicity. More efficient econometric methods can be developed to refine the estimations for endogenous dependencies, and we leave these methods for future research.

## 4.2. Optimal Staffing

In this section, we use simulation examples to demonstrate how to utilize our estimates in Section 4.1 to make predictions for the optimal staffing level if the arrival rate is about to change, with no data available (e.g., service time or waiting time) for that new rate. Specifically, we consider two systems, each having an exogenous dependence (marginal service and patience times are exponentially distributed with means one and two, respectively, and their joint distributions are generated via Gaussian copulas with correlation  $r = 0.5$  and  $r = -0.3$ , respectively), 10 agents and a Poisson arrival process with rate 12, and use our proposed procedure to estimate the equivalent endogenous dependence. We then consider a new system with the same exogenous dependence as before, but increase the arrival rate to 24. The goal is to identify the optimal number of agents for the new system to maximize profit. We follow Wu et al. (2019) and define the profit



**Table 2.** Staffing Using Estimations of Equivalent Endogenous Dependence Compared with Optimal Staffing and Staffing Using Estimations That Assume Independent Service and Patience Times

$p$	Optimal		Endogenous		Independent	
	Capacity	Profit	Capacity	Profit	Capacity	Profit
Positive dependence, $r = 0.5$						
1.5	25	8.16	26	-1.3%	23	-9.8%
3	27	42.35	28	-0.3%	30	-2.6%
Negative dependence, $r = -0.3$						
0.95	9	2.74	10	-2.0%	15	-30.5%
1.5	18	11.96	18	0.0%	21	-3.8%
3	24	44.51	24	0.0%	24	0.0%

Notes.  $c = 1$  in all examples. When  $p = 0.95$ , service is unprofitable under positive dependence  $r = 0.5$ .

as the difference between the revenue generated from customers served and cost of allocating servers. Let  $p$  denote the revenue of serving a customer and  $c$  denote the unit cost of capacity.

In Table 2, we compare the optimal capacity and profit (“Optimal” column) to the corresponding optimal values obtained from our estimations of the equivalent endogenous dependence (“Endogenous” column) and from estimations that ignore the dependence (“Independent” column). We find that our estimations of the equivalent endogenous dependence lead to accurate prescriptions of the optimal staffing level, whereas the estimations that treat the dependence to be nonexistent may lead to substantial loss in profit, especially when there is a negative dependence and the reward of serving customers is low,  $p < c$  (recall the unconditional service rate of customers not delayed in queue is  $\mu = 1$ ). When  $p < c$ , service is unprofitable in the system with positive dependence because the throughput in such a system is lower than the capacity. However, service can be profitable in a system with negative dependence, because the throughput in such a system is higher than the capacity; see Wu et al. (2019). In this latter case, our estimations of the equivalent endogenous dependence significantly outperform those that ignore the dependence in prescribing the staffing level.

In ending, we remark that a similar staffing problem is considered in Wu et al. (2019). In the setting of Wu et al. (2019), management is assumed to know that (i) the dependence is exogenous, and (ii) the joint service-patience distribution. However, as discussed above, it is hard in practice to determine the form of the dependence mechanism and to estimate the joint distribution, even if one believes that the dependence is exogenous. It is also significant that Wu et al. (2019) solves the staffing problem by utilizing a fluid model that is useful as an approximation for *large systems*. In contrast, our estimates of the endogenous dependence (whether it is the true dependence mechanism or is equivalent to an actual exogenous dependence) do not require any prior knowledge regarding the exact form of the underlying

dependence. Moreover, unlike Wu et al. (2019), our results hold for stochastic systems themselves, and not for their asymptotic approximations.

### Acknowledgments

The authors thank the area editor Amy Ward, the associate editor, and two anonymous reviewers for their careful reading of the paper and for providing constructive feedbacks.

### Appendix A. A Model with Generalized Dependence Mechanism

In this section, we consider a generalized dependence mechanism, in which each customer’s service time depends on his patience and his delay in queue. This generalized model naturally subsumes exogenous and endogenous dependencies as special cases. As we show below, the queuing dynamics under such generalized dependence are again equivalent to the queuing dynamics under certain endogenous dependence, so that, once again, endogenous dependence is all one needs to consider in practice.

#### The Model

Letting  $T_i^G$  denote the patience time of customer  $i$ , we assume that  $\{T_i^G : i \geq 1\}$  are i.i.d. continuous random variables that are independent from the system’s state. We denote the cdf and pdf of  $T_i^G$  by  $F_T^G$  and  $f_T^G$ , respectively, with  $\bar{F}_T^G \triangleq 1 - F_T^G$  denoting the corresponding cdf. The service-time distribution of each customer depends on the customer’s patience, as well as on his delay in queue. Specifically, letting  $Z_i^G$  denote the offered wait of customer  $i$ , we assume that the service times of arriving customers are described by a stochastic process,  $\{S_i^G(T_i^G, Z_i^G)\}$ , where  $S_i^G(T_i^G, Z_i^G)$  denotes a random variable representing the virtual service time of customer  $i$ , given that his patience time is  $T_i^G$  and offered wait is  $Z_i^G$ . We assume that  $\{S_i^G(t, z) : i \geq 1\}$  are independent across customers, are identically distributed for each realized value  $(t, z)$  of  $(T_i^G, Z_i^G)$ , and are also independent of all other random variables comprising the system. Let  $\Xi^G$  denote the virtual service-time distribution of a customer, namely,  $\Xi^G(x, t, z) \triangleq P(S_i^G \leq x \mid T_i^G = t, Z_i^G = z)$ .

We assume that the pdf of the virtual service time exists and satisfies  $\xi^G(x, t, z) = \frac{\partial \Xi^G(x, t, z)}{\partial x}$ .

Clearly, the exogenous and endogenous dependence mechanisms are special cases of the generalized-dependence mechanism just described: The dependence is exogenous if  $\xi(x, t, z)$  does not depend on  $z$  and is endogenous if  $\xi(x, t, z)$  does not depend on  $t$ . For a system with generalized dependence, we can follow Section 2.2 to formulate the queuing dynamics using a Markov process and define its probability law. We can define an equivalence relation between generalized dependence and endogenous dependence analogously to Definition 1 by replacing the law under exogenous dependence by the law under generalized dependence. The following theorem shows that for each generalized dependence, there exists an equivalent endogenous dependence, and it must be unique, as implied by Lemma 1.

**Theorem A.1.** *For any generalized dependence mechanism characterized by  $(f_T^G, \xi^G)$ , there is an equivalent endogenous dependence  $(f_T^{en}, \psi^{en})$ . The two dependencies are related via*

$$f_T^{en}(z) = f_T^G(z) \text{ and } \psi^{en}(x, z) = \frac{\int_z^\infty \xi^G(x, t, z) f_T^G(t) dt}{\bar{F}_T^G(z)}$$

for all  $x, z \geq 0$ .

The fact that the endogenous dependence is subsumed by the generalized dependence implies that the family of systems with endogenous dependence forms a subclass of the family of systems with generalized dependence. Theorem A.1 shows that these two classes are in fact identical. Therefore, the generalized dependence is *not more general than the endogenous dependence*, in the sense that both mechanisms give rise to the same family of distributions of queuing dynamics. As a result, our main insight remains valid: Statistical analyses of systems can be carried out by assuming that the service requirement of each customer depends on his delay in queue, even if the service time depends (solely, or additionally) on his patience.

## Appendix B. Proofs

**Proof of Lemma 1.** Sufficiency follows trivially from the fact that the law of the process  $\mathbf{X}^{en}$  is uniquely determined by  $(f_T, \psi)$ , given  $A$  and  $n$ .

To prove the necessity, we show that, if  $\mathcal{L}^{en}(A, n; f_T^1, \psi^1) = \mathcal{L}^{en}(A, n; f_T^2, \psi^2)$  for an arbitrary arrival process  $A$ , capacity  $n$ , and initial conditions  $\mathbf{X}(0)$ , then  $(f_T^1, \psi^1) = (f_T^2, \psi^2)$ . To this end, we focus on the second arriving customer ( $i = 2$ ) in the single-server system ( $n = 1$ ), where we construct proper arrival process  $A$  and initial conditions  $\mathbf{X}(0)$  to demonstrate that having  $Z_2^1 =^d Z_2^2$  is sufficient to lead to  $(f_T^1, \psi^1) = (f_T^2, \psi^2)$ .

The offered wait of customer 2 is described via a recursive formula (Baccelli et al. 1984, equation (2.1)),  $Z_2^j = [Z_1^j + \mathbb{I}_{\{T_1^j > x\}} S_1^j(Z_1^j) - \alpha_2^j]^+$ ,  $j = 1, 2$ . For any  $z > 0$ , let  $Z_1^1 = Z_1^2 = z$  w.p.1, (e.g., by letting  $U^1(0) = U^2(0) = 2z$ ,  $Q^1(0) = Q^2(0) = 0$  and  $\alpha_1^1 = \alpha_1^2 = z$  w.p.1) and  $\alpha_2^1 = \alpha_2^2 = z/2$  w.p.1. It follows that  $Z_2^j = [z/2 + \mathbb{I}_{\{T_1^j > z\}} S_1^j(z)]^+$  w.p.1. Hence,  $P(Z_2^j \leq z/2) = P([z/2 + \mathbb{I}_{\{T_1^j > z\}} S_1^j(z)]^+ \leq z/2) = P(z/2 + \mathbb{I}_{\{T_1^j > z\}} S_1^j(z) \leq z/2) = P(T_1^j \leq z)$ . Because the queuing dynamics in the two systems have the same law, it must hold that  $P(Z_2^1 \leq z/2) = P(Z_2^2 \leq$

$z/2)$  for all  $z$ , implying  $P(T_2^1 \leq z) = P(T_2^2 \leq z)$  for all  $z$ . Because  $z$  is arbitrary, it must hold that  $f_T^1 = f_T^2$ .

For any  $z, x > 0$ , let  $Z_1^1 = Z_1^2 = z$  and  $\alpha_2^1 = \alpha_2^2 = z/2$  w.p.1. It follows that

$$\begin{aligned} P\left(Z_2^j > \frac{z}{2} + x\right) &= P\left(\left[\frac{z}{2} + \mathbb{I}_{\{T_1^j > z\}} S_1^j(z)\right]^+ > \frac{z}{2} + x\right) \\ &= P\left(\frac{z}{2} + \mathbb{I}_{\{T_1^j > z\}} S_1^j(z) > \frac{z}{2} + x\right) \\ &= P(\mathbb{I}_{\{T_1^j > z\}} S_1^j(z) > x) \\ &= P(S_1^j(z) > x)P(T_1^j > z), \end{aligned}$$

where the last equality follows because  $T_1^j$  and  $S_1^j(\cdot)$  are independent. Because we have established earlier that  $P(T_1^1 > z) = P(T_1^2 > z)$  for all  $z > 0$ , it follows that  $P(S_1^1(z) > x) = P(S_1^2(z) > x)$ . Because  $z$  and  $x$  are arbitrary, we thus have  $\psi^1 = \psi^2$ .  $\square$

**Proof of Theorem 2.** Sufficiency: Suppose  $\bar{\Psi}^{en}(x, z) \bar{F}_T^{en}(z)$  is decreasing in both  $z$  and  $x$ . We show by construction that there exists an equivalent exogenous dependence. Notice that  $\lim_{z \rightarrow \infty} \lim_{x \rightarrow \infty} \bar{\Psi}^{en}(x, z) \bar{F}_T^{en}(z) = 0$  and  $\lim_{z \rightarrow 0} \lim_{x \rightarrow 0} \bar{\Psi}^{en}(x, z) \bar{F}_T^{en}(z) = 1$ . The differentiability of  $\Psi^{en}$  in Assumption 1 implies that  $\bar{\Psi}^{en}(x, z) \bar{F}_T^{en}(z)$  is jointly continuous in  $x$  and in  $z$ . Hence,  $\bar{\Psi}^{en}(x, z) \bar{F}_T^{en}(z)$  can be represented as a continuous two-dimensional ccdf. The differentiability of  $\Psi^{en}(x, z)$  in  $z$  also implies  $\bar{\Psi}^{en}(x, z) \bar{F}_T^{en}(z)$  is differentiable in  $z$ . Hence, there exists a bivariate joint density  $\hat{f}^{ex} : \mathbb{R}_+^2 \mapsto \mathbb{R}_+$  such that  $\bar{\Psi}^{en}(x, z) \bar{F}_T^{en}(z) = \int_{y=z}^\infty \int_{u=x}^\infty \hat{f}^{ex}(u, y) du dy$ . One can verify that  $\hat{f}^{ex}$  defined above, together with  $(f_T^{en}, \psi^{en})$ , satisfies the conditions in (1). Theorem 1 then implies the endogenous dependence  $(f_T^{en}, \psi^{en})$  is equivalent to an exogenous dependence with joint service-patience distribution  $\hat{f}^{ex}$ .

Necessity: We prove by contradiction. Suppose there exists an endogenous dependence  $(f_T^{en}, \psi^{en})$  such that  $\bar{\Psi}^{en}(x, z) \bar{F}_T^{en}(z)$  is not decreasing in  $x$  and in  $z$ . Further, it is equivalent to some exogenous dependence  $\hat{f}^{ex}$ . Theorem 1 and Lemma 1 imply the latter exogenous dependence  $\hat{f}^{ex}$  is equivalent to a *unique* endogenous dependence  $(\hat{f}_T^{en}, \hat{\psi}^{en})$  such that  $\hat{\Psi}^{en}(x, z) \hat{F}_T^{en}(z)$  is decreasing in  $x$  and in  $z$ . This leads to a contradiction to the assumed equivalence between  $(f_T^{en}, \psi^{en})$  and  $\hat{f}^{ex}$ .  $\square$

**Proof of Theorem A.1.** The arguments to prove this result are similar to those of Theorem 1. We follow two steps: A construction step, in which we construct two systems jointly via a coupling argument; and a verification step, in which we show that the coupled system constructed in the first step has the law of a system with generalized dependence. Because the verification step requires a tedious computation, we omit it for brevity.

Following the notation in the proof of Theorem 1, we use a  $\sim$  (tilde) to denote the stochastic processes and random variables on the new probability space, where we couple two systems. For each sample path describing the dynamics of a system with an endogenous dependence, we construct a coupled system in the following steps. We first generate  $(\tilde{t}_i^{en}, \tilde{s}_i^{en}(\cdot))$  and  $\tilde{\alpha}_i^{en}$  for each new arrival in

the system with endogenous dependence. The offered-wait process for these new arrivals is fully characterized by the Recursion (2). We next construct a coupled system and use superscript “c” to denote the random variables in it.

i. Set  $\tilde{\alpha}_i^c = \tilde{\alpha}_i^{en}$ , so that the  $i$ th customer arrives at the same time in the system with endogenous dependence and the coupled system. Hence, both systems have the same realized arrival process.

ii. If  $\tilde{t}_i^{en} \leq \tilde{z}_i^{en}$ , then the  $i$ th customer abandons in the system with endogenous dependence. Set  $\tilde{T}_i^c = \tilde{t}_i^{en}$  and generate  $\tilde{S}_i^c$  from the density  $\xi^G(\cdot, \tilde{t}_i^{en}, \tilde{z}_i^{en})$ , namely, from the conditional distribution of  $S_i^G(T_i^G, Z_i^G)$  conditioned on  $T_i^G = \tilde{t}_i^{en}$  and  $Z_i^G = \tilde{z}_i^{en}$ .

iii. If  $\tilde{t}_i^{en} > \tilde{z}_i^{en}$ , then the  $i$ th customer is served in the system with endogenous dependence and requires a service time  $\tilde{s}_i^{en}(\tilde{z}_i^{en})$ . Set  $\tilde{S}_i^c = \tilde{s}_i^{en}(\tilde{z}_i^{en})$  and generate  $\tilde{T}_i^c$  from the density

$$\frac{\xi^G(\tilde{s}_i^{en}(\tilde{z}_i^{en}), \cdot, \tilde{z}_i^{en}) f_T^{en}(\cdot) \mathbb{I}_{\{\cdot > \tilde{z}_i^{en}\}}}{\int_{\tilde{z}_i^{en}}^{\infty} \xi^G(\tilde{s}_i^{en}(\tilde{z}_i^{en}), t, \tilde{z}_i^{en}) f_T^{en}(t) dt},$$

namely, from the conditional distribution of  $T_i^G$  conditioned on  $S_i^G = \tilde{s}_i^{en}(\tilde{z}_i^{en})$  and  $T_i^G > \tilde{z}_i^{en}$ .

The coupled system has the same queuing dynamics as the system with endogenous dependence. Using basic computations, we can further show that the coupled system has the desired generalized dependence described by  $(f_T^G, \xi^G)$ .  $\square$

## References

- Akritis MG, Keilegom IV (2003) Estimation of bivariate and marginal distributions with censored data. *J. Roy. Statist. Soc. Ser. B Statist. Methodology* 65(2):457–471.
- Baccelli F, Boyer P, Hebuterne G (1984) Single-server queues with impatient customers. *Adv. Appl. Probab.* 16(4):887–905.
- Bassamboo A, Randhawa RS (2015) Scheduling homogeneous impatient customers. *Management Sci.* 62(7):2129–2147.
- Billingsley P (2013) *Convergence of Probability Measures* (John Wiley & Sons, New York).
- Boxma OJ, Vlasiou M (2007) On queues with service and interarrival times depending on waiting times. *Queueing Systems* 56(3–4): 121–132.
- Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L (2005) Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* 100(469): 36–50.
- Browne S, Yechiali U (1990) Scheduling deteriorating jobs on a single processor. *Oper. Res.* 38(3):495–498.
- Chan CW, Farias VF, Escobar GJ (2017) The impact of delays on service times in the intensive care unit. *Management Sci.* 63(7): 2049–2072.
- Dabrowska DM (1988) Kaplan-Meier estimate on the plane. *Ann. Statist.* 16(4):1475–1489.
- De Vries J, Roy D, De Koster R (2018) Worth the wait? How waiting influences customer behavior and their inclination to return. *J. Oper. Management* 63(1):59–78.
- Glazebrook K (1992) Single-machine scheduling of stochastic jobs subject to deterioration or delay. *Naval Res. Logistics* 39(5): 613–633.
- Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* 53(282):457–481.
- Lopez O, Saint-Pierre P (2012) Bivariate censored regression relying on a new estimator of the joint distribution function. *J. Statist. Planning Inference* 142(8):2440–2453.
- Mandelbaum A, Zeltyn S (2004) The impact of customers’ patience on delay and abandonment: Some empirically-driven experiments with the M/M/n + G queue. *OR Spectrum* 26(3):377–411.
- Mosheiov G (1991) V-shaped policies for scheduling deteriorating jobs. *Oper. Res.* 39(6):979–991.
- Moyal P (2019) Coupling in the queue with impatience: Case of several servers. *Discrete Event Dynam. Systems* 29(2):145–162.
- Reich M, Mandelbaum A, Ritov Y (2010) The workload process: Modelling, inference and applications. Working paper, Technion-Israel Institute of Technology, Haifa, Israel.
- Sugawa S, Takahashi M (1965) On some queues occurring in an integrated iron and steel works. *J. OR Society Japan* 8(1):16–23.
- Whitt W (1990) Queues with service times and interarrival times depending linearly and randomly upon waiting times. *Queueing Systems* 6(1):335–351.
- Whitt W (2002) *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*, Springer Series in Operations Research and Financial Engineering (Springer Science & Business Media, New York).
- Wu CA, Bassamboo A, Perry O (2019) Service systems with dependent service and patience times. *Management Sci.* 65(3): 1151–1172.
- Zohar E, Mandelbaum A, Shimkin N (2002) Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Sci.* 48(4):566–583.

**Chenguang (Allen) Wu** is an assistant professor in the Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology. His research interests include service operations, bundling, and supply chain management.

**Achal Bassamboo** is the Charles E. Morrison Professor at Kellogg School of Management, Northwestern University. He joined the faculty at the Kellogg School of Management in 2005. His research interests lie in the areas of service systems, revenue management, and information sharing.

**Ohad Perry** is an associate professor in the Department of Industrial Engineering and Management Science, Northwestern University. His research focuses on applying methodologies from applied probability and dynamical systems’ control to approximate and analyze complex stochastic systems with applications to service and inventory systems.