European Journal of Operational Research 279 (2019) 673–686

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Invited Review

Load effect on service times

Mohammad Delasay[a,*], Armann Ingolfsson[b], Bora Kolfal[b], Kenneth Schultz[c]

[a] College of Business, Stony Brook University, Stony Brook, New York, USA
[b] Alberta School of Business, University of Alberta, Edmonton, Alberta, Canada
[c] Ohio, USA

ARTICLE INFO

Article history:
Received 19 September 2017
Accepted 15 December 2018
Available online 23 December 2018

Keywords:
Queueing
Behavioral OR
OR in the service industries
Service time
Load

ABSTRACT

In this paper, we develop a general framework to analyze the influence of system load on service times in queueing systems. Our framework unifies previous results and ties them to possible future studies to help empirical and analytical researchers to investigate and model the ways in which load impacts service times. We identify three load characteristics: *changeover*, *instantaneous load*, and *extended load*. The load characteristics induce behaviors, or *mechanisms*, in at least one of the system components: the *server*, the *network*, and the *customer*. A mechanism influences the service-time determinants: the *work content*, *service speed*, or *in-process delay*. We identify and define mechanisms that cause service times to change with load and use the framework to categorize them. We argue that an understanding of the relationship between load and service times can come about only by understanding the underlying mechanisms.

1. Introduction

An understanding of queueing systems is critical to the management of service, production, and supply chain systems. Queueing theory informs the planning of customer service, capacity, processing times, flow times, and delivery schedules. The queueing literature has clearly documented the influence of service times on system load. What is less well understood is the influence of load on service times.

Consistent with the notion of service time in the empirical research that we review, we use "service time" in this paper to mean the time spent inside the process boundaries and "processing time" to mean the part of that time spent serving the customer. Most queueing theory models assume that service times are exogenous. That is, they assume service times are independent of the system state. Recent empirical studies have made it clear that service times are endogenous: They depend on load. The direction and magnitude of the relationship are not clear, however, and the underlying mechanisms vary across applications. The following quotes exemplify a sample of findings that at first glance appear contradictory, where ↗ denotes service times increase with load, ↘ denotes service times decrease with load, and sequences of these symbols denote non-monotone patterns:

↘ It can be seen that [service] time is appreciably longer at low volumes of traffic... than it is at high volumes (Edie, 1954, p. 120).

↗ High hospital occupancy has a significant and quantifiable negative influence on ED [emergency department] throughput, affecting patients both discharged and hospitalized (Hillier, Parry, Shannon, and Stack, 2009, p. 767).

↘↗ We find that workers accelerate the service rate as load increases.... Long periods of increased load (overwork) have the effect of decreasing the service rate (Kc and Terwiesch, 2009, p. 1486).

↗↘... we show that the aggregate effect of load on service time is an inverted U-shaped response, but of modest magnitude (< 10% change) (Batt and Terwiesch, 2016, p. 32).

↗↘↗... we find evidence that patient length of stay... increases as occupancy increases, until a tipping point, after which patients are discharged early to alleviate congestion. More interestingly, we find a second tipping point—at 93% occupancy—beyond which additional occupancy leads to a longer LOS [length of stay].... Collectively, we find that the underlying relationship between occupancy and LOS is N-shaped (Berry Jaeker and Tucker, 2017, abstract).

In interpreting these quotes, we follow the implicit assumption in much of the literature that we survey, that mean service times are inversely related to service rates and to throughput. We revisit this assumption in Section 5.

* Corresponding author.
E-mail addresses: mohammad.delasay@stonybrook.edu (M. Delasay), armann.ingolfsson@ualberta.ca (A. Ingolfsson), bora.kolfal@ualberta.ca (B. Kolfal), BopSchultz@gmail.com (K. Schultz).

https://doi.org/10.1016/j.ejor.2018.12.028
0377-2217/

We do not *rectify* the theories in the above quotes in the sense that we do not choose one above the other. Rather, we demonstrate that choosing one above the other would be in error. We reconcile them in the sense that we show how they fit together into a common body of knowledge, a task which has not previously been attempted.

Our review of the literature indicates that different changes in service time are caused by responses to different load-driven mechanisms. The extant research in this area has been exemplary, and we believe that a framework that unifies previous results and ties them to possible future studies will enrich this body of work. We propose a framework that conceptualizes a relationship between queueing elements, load, and service time. We refer to the framework as the Load Effect on Service Times (LEST) framework.

The quotes above do not represent *competing* theories so much as *complementary* theories. However, that was not the common interpretation before we proposed the LEST framework. When we began this work a central theme from published work and colleagues was the existence of a predominant reaction that, with sufficient work, could be characterized and defined. Oliva and Sterman (2001, p. 896) wrote of a "formal model [of] 'high-contact' service" and Kc and Terwiesch (2009, p. 1488) proposed a "general model of service operations," both arguing for only one, or two, of the mechanisms identified in this paper. We argue that there is no such predominant response. Rather there are mechanisms that are activated in different situations, caused by different factors, and have different effects. Further we argue that research should move away from seeking a dominant response to better understanding the mechanisms.

The LEST framework provides an organizing structure to a growing body of empirical work on the relation between load and service times. Some of the papers we analyzed were not intended to be investigations of the effects of load on service times. We know this because we wrote some of them (Schultz, Juran, Boudreau, McClain, & Thomas, 1998; Schultz, McClain, & Thomas, 2003). It was only after we began this research that we saw the connection. Making those connections, bringing separate, previously unconnected, work into one frame is one of the contributions of this paper.

The answer to the question "What is the effect of load on service time?" is "It depends." The more difficult question, on what does it depend, is the focus of much current research. We propose that an understanding of these questions can occur only by understanding the mechanisms involved. Furthermore, we argue that advancement in this area will come about through a study of these mechanisms, the size and shape of their effects, the operational situations in which they are commonly activated, and the models describing how they interact with other system components. We contribute to further research by providing common names and definitions for mechanisms, giving organization to the body of work and allowing future researchers to know what has come before and how their work fits into the whole.

The LEST framework breaks down the question of how load impacts service time into its component parts and provides a language to formulate questions about the component parts. In this respect, LEST is similar to Kingman's (1961) equation, which decomposes the impacts of variability, utilization, and average service time on average delay; or Vroom's Expectancy Theory, which provides a language for formulating questions about the role of beliefs and motives in work performance (Vroom, 2005). Much of the art of managing operations lies in knowing what questions to ask. If we ask better questions, we can find better solutions. This paper presents a conceptual model that focuses on the skill of suggesting questions. We demonstrate a set of relationships, the value of which is contained in the questions they help us to generate.

The LEST framework is general, in the sense that it is applicable to any type of production or service system. The framework provides a comprehensive and systematic basis to investigate and explain how system components react and interact in response to system load and how the reactions and interactions cause variations in service times. We justify the generality of the framework—in part, by scrutinizing published empirical studies and using the framework to explain the observed relationships between service times and system load.

The LEST framework can help empirical researchers by identifying promising questions for future research and assisting them in placing their work within the broader picture. The framework conceptualizes a thinking process that an empirical researcher can use by provoking such questions as: How is load characterized? Which system components react to load? What are the mechanisms that relate load variations to system component reactions? Which parts of the service time increase or decrease with which mechanisms? By focusing on mechanisms, empiricists can contribute to the field by building our understanding of the scope, frequency, and impact of particular mechanisms. For example, is the effect of *peer pressure* linear, concave, or convex in load? Does peer pressure affect the work content, the service speed, or the in-process delay (the three determinants of the service time)? Is the effect different between servers and customers?

The study of mechanisms will benefit analytical researchers as well, by making models richer and more closely connected to observed system behavior. The LEST framework can help analytical researchers to improve understanding of queues by answering two fundamental questions: "What are the factors on which service times depend? And, how can these factors be translated into state variables?" The proposed framework also emphasizes the importance of certain queueing model characteristics that appear to be important in the empirical literature but that are less frequently discussed in the analytical literature, including single-node queueing systems vs. queueing networks, human vs. inanimate servers or customers, dedicated vs. shared servers, and single vs. multiple customer types.

In the remainder of the paper, we provide further background in Section 2; propose the LEST framework in Section 3; identify, define, and categorize mechanisms in Section 4; discuss modeling implications in Section 5; and lay out our conclusions and discuss future research directions in Section 6.

2. Background

A. K. Erlang developed the classical Erlang *C* and *B* queueing models in the 1910s to quantify traffic congestion in telephone systems (Brockmeyer, Halstrøm, Erlang, & Jensen, 1948). Such classical models are used for capacity planning in manufacturing, telecommunication, and service systems and are used extensively in research on production and service systems. These models are characterized by the assumption that service time distribution parameters are exogenous—independent of the system state.

Exogeneity was called into question, initially based on anecdotal evidence (e.g., Gomersall, 1964). Formal empirical research on queueing systems gained momentum in the 1990s (Gupta, Verma, & Victorino, 2006; Scudder & Hill, 1998) with the use of data collected from field research, archival records, or laboratory experiments. That work has increasingly called into question the validity of the exogeneity assumption (e.g., Inman, 1999; Robbins, Medeiros, & Harrison, 2010).

As we have already demonstrated, a valuable stream of empirical research has supported the dependence of service times on load. A field study of toll collection for the Port Authority of New York found, for example, that drivers who wait longer in line are more likely to have their change ready, leading to shorter

average payment times (Edie, 1954). A laboratory experiment of a low-inventory serial line (Schultz et al., 2003) found that subjects worked at a slower pace during a warmup period after an unintended break caused by a job shortage (an absence of items to work on). Regression analysis of archival data from several hospitals (Kuntz, Mennicken, Scholtes et al., 2011) suggested an inverted U-shape relationship between bed occupancy and length of hospital stay: The LOS increased with occupancy up to a tipping point as patients waited longer for diagnosis, and the LOS dropped after the tipping point because doctors discharged patients earlier to accommodate incoming patients.

These empirical findings represent some fundamental differences. In Edie (1954), for example, it was the behavior of the driver (the customer) in response to load that affected payment time, whereas in Schultz et al. (2003), it was the worker (the server) who behaved adaptively. Another example is the way in which system load is represented: Edie (1954) viewed load as the queue length (number of cars in line), whereas Schultz et al. (2003) characterized load based on whether the amount of work in process (WIP) was zero (idle period) or positive (busy period). Some studies showed a negative relationship (e.g., Edie, 1954), some a positive relationship (e.g., Schultz et al., 2003), and some both a positive and a negative relationship (e.g., Kuntz et al., 2011) between service times and system load. In this paper, we propose a general framework that incorporates these and other controversies.

Although the voluminous body of research in queueing theory since the days of Erlang has extended the classical models in many ways, few modelers have relaxed the exogeneity assumption. Jackson (1963), Welch (1964), Harris (1967), and Graves (1986), among others, have modeled state-dependent queues. The mean service rate in these models depends on the system state, which could be either the queue length or the amount of unfinished work (Dshalalow, 1997). Other theorists have developed vacation models (e.g., Levy & Yechiali, 1975) to capture the type of load characterization that Schultz et al. (1998) observed: lower service rates after a break (vacation) due to setup.

Despite these efforts, there has been limited progress in modeling the effects of load on service times. We postulate that the limited progress in this area is, partly, due to the fact that the nature of the dependency of service times on load is not clear—the issue that the LEST framework can address. For the sake of model tractability, state-dependent models often disregard the central characteristics of real queueing systems. Our proposed framework highlights such characteristics. Most state-dependent models assume a single server, for example, and therefore overlook behaviors like *social loafing* (Karau & Williams, 1993) in multi-server systems. And state-dependent models typically ignore interactions among nodes in a queueing network, such as the impact of hospital occupancy on the LOS of patients in a hospital emergency department (ED) (Hillier et al., 2009).

Advances in numerical techniques and the growing empirical evidence about queueing systems provides opportunities for queueing modelers to include significant characteristics of real systems and allow for more flexible interactions among system components. For example, phase-type distributions facilitate viewing service times as the outcome of a dynamic process of customer–server interaction (Kingman, 2009, and Khudyakov, Gorfine, and Mandelbaum, 2018, as reported in Gans, Liu, Mandelbaum, Shen, & Ye, 2010). Matrix-analytic methods (Neuts, 1981) allow for different load characteristics to affect service times simultaneously (e.g., Azriel, Feigin, and Mandelbaum, 2014; Delasay, Ingolfsson, and Kolfal, 2016a). In this respect, OR is starting to achieve the kind of fertile interplay between experiment and theory that one sees in other sciences such as physics (Fisher, 2007).

3. Load effect on service times framework

Our framework features a chain of effects that connects system load to service time (Fig. 1). We identify three load characteristics: *changeover*, *instantaneous load*, and *extended load*. (We abbreviate "instantaneous load" to "load" in the remainder of the paper.) The load characteristics induce behaviors, or *mechanisms*, in one of the system components: the *server*, the *network*, or the *customer*. The mechanism influences one of the service-time determinants: *work content*, *service speed*, or *in-process delay*. In Sections 3.1–3.3, we explain each box in Fig. 1 and define our terminology.

A mechanism is a link between load characteristics and service time due to a specific cause. We borrow the term from Batt and Terwiesch (2016). One can visualize a mechanism (Fig. 1) as a path from one of the load characteristics, through one of the system components and one of the service-time determinants, to the service time.

Mechanisms are exclusive to one path but paths are not exclusive to one mechanism: Multiple mechanisms can share the same path, but differ in cause. For example, we will see several mechanisms associated with the load–server–work content–service time path (Table 1). Although they share the same path they are different mechanisms because they have different causes. Changes to the three load characteristics invoke different behaviors or induce mechanisms in three system components: the server, the network, or the customer. The mechanism *fatigue*, for example, is caused by high load for an extended period and causes servers to reduce speed, resulting in longer service times (Kc & Terwiesch, 2009). The causes are often behavioral but they need not be. In this paper mechanisms affect service time but the term could be used for other effects on queues as well. For instance, different causes of jockeying could be referred to as mechanisms.

3.1. Load characteristics

Load characteristics are the indices, measures, and conditions by which system load is characterized. We identify three system load characteristics:

Changeover refers to a change from one type of customer to another, including from idle to busy. We use changeover as a separate load characteristic because it involves fundamentally different mechanisms than other changes in load.

Load refers to a measure or set of measures that identifies how busy or congested a system is at a given time. Load is usually measured as the number of jobs in the system, the caseload or number of jobs assigned to a server (the mutitasking level), the amount of unfinished work, or the occupancy rate or occupied capacity. The number of patients in an ED waiting room is one way to measure ED load.

Extended load tracks the history of system load. It usually refers to a situation in which the system has been under a heavy load for an extended period. Gans et al. (2010) measure extended load as the number of calls an agent has answered since the last service gap of longer than an hour; these authors show that agents slow down during periods of extended load.

3.2. System components

Server: We use this term generically, without necessarily implying that servers are human. The server is the person, the resource, or the bundle of people and other resources that provides service. Some systems—diagnostic imaging for hospital physicians or computer and telecommunication infrastructure for a call center (Akşin & Harker, 2003), for example—have shared resources that do not belong exclusively to any single server.

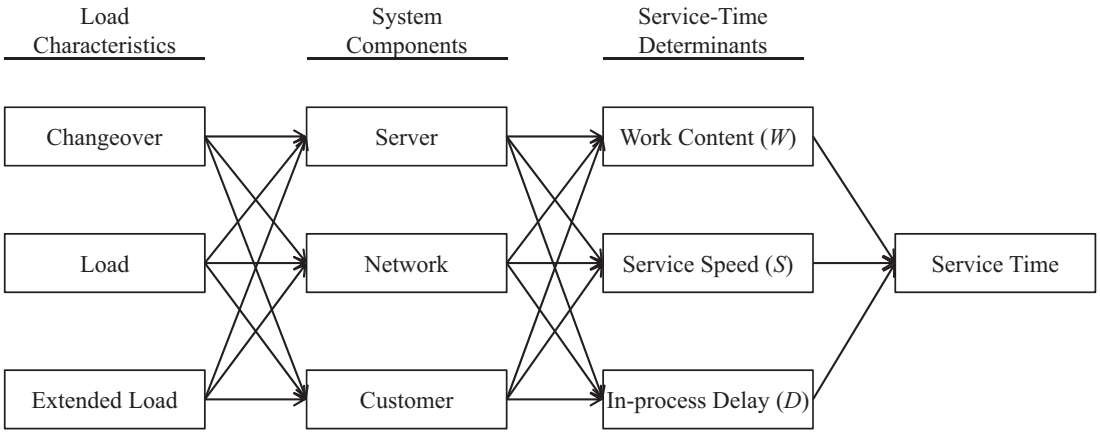


Fig. 1. The LEST framework.

Table 1
Framework and mechanisms.

		System components		
		Server Section 4.1	Network Section 4.2	Customer Section 4.3
Load characteristics	Changeover Section 4.X.1	Work content *Physical setup* (↑) (Schultz et al. 2003) *Forgetting* (↑) (Bailey 1989; Mark et al. 2005; Schultz et al. 2003; Ibanez et al. 2017) Service speed *Loss of rhythm* (↑) (Schultz et al. 2003; Staats & Gino 2012)	Work content *Network arrangement* (↓) (Delasay et al. 2016b)	Work content *Customer early task initiation* (↓) (Edie 1954; Wang & Zhou 2018)
	Load Section 4.X.2	Work content *Task reduction* (↓) (Batt & Terwiesch 2016; Delasay et al. 2016b; Forster et al. 2003; KC 2014; Kc & Terwiesch 2009; KC & Terwiesch 2012; Kuntz et al. 2014; Oliva 2001; Oliva & Sterman 2001; Mæstad et al. 2010; Long & Mathews 2017; Chan et al. 2018) *Engagement* (↑) (Delasay et al. 2016b; Tan & Netessine 2014) *Server early task initiation* (↓) (Batt & Terwiesch 2016; Delasay et al. 2016b) *Multitasking–cognitive sharing* (↑) (Aral et al. 2012; KC 2014; Lu 2013)	Work content *Geographical dispersion* (↑) (Delasay et al. 2016b)	Work content *Return* (↑) (KC 2014; KC & Terwiesch 2012; Long & Mathews 2017⁻)
		Service speed *Social speedup pressure* (↓) (Edie 1954; Kc & Terwiesch 2009; Lu 2013; Mas & Moretti 2009; Schultz et al. 1998; Shunko et al. 2018; Staats & Gino 2012; Tan & Netessine 2014; Wang & Zhou 2018) *Social loafing* (↑) (Berry Jaeker & Tucker 2012; Mas & Moretti 2009; Wang & Zhou 2018)	Service speed *Geographical speedup* (↓) (Delasay et al. 2016b)	
		In-process delay *Multitasking–time sharing* (↑) (Aral et al. 2012; Goes et al. 2018; KC 2014; Lu 2013; Tan & Netessine 2014) *Multitasking–interruptions* (↑) (Chisholm et al. 2000; KC 2014; Lu 2013) *Workload smoothing* (↓) (Berry Jaeker & Tucker 2012; Kim et al. 2014; Long & Mathews 2017)	In-process delay *Resource sharing* (↑) (Hillier et al. 2009) *Downstream system congestion* (↑) (Asaro et al. 2007; Delasay et al. 2016b; Forster et al. 2003; Hillier et al. 2009; Kim et al. 2014; Long & Mathews 2017; Louriz et al. 2012)	In-process delay *Abandonment* (↓) (Batt & Terwiesch 2015; Lu et al. 2013; De Vries et al. 2018)
	Extended Load Section 4.X.3	Work content *Service cancelation* (↓) (Brown et al. 2005) Service speed *Learning by doing* (↓) (Lu 2013) *Fatigue* (↑) (Gans et al. 2010; Kc & Terwiesch 2009; Staats & Gino 2012)	Work content *Network chaos* (↑) (Delasay et al. 2016b)	Work content *Deterioration* (↑) (Kc & Terwiesch 2009)

Network: A system may consist of multiple subsystems. When we analyze a subsystem or a "node," consisting of a set of servers performing the same task fed by one or multiple queues, we consider a network mechanism to be any mechanism that originates from outside the node of interest but impacts service times in the node of interest. We interpret the term "network" broadly, to refer to any influencing component or agent from outside the node of current interest. We provide three examples to illustrate types of components that we intend to include: (1) upstream or downstream nodes, (2) a transportation network, for systems with mobile servers, and (3) managerial action that impacts incentives or targets that servers face.

Customer: The customer is the person or inanimate object that receives service. Patients are the customers in an ED, for example, and unfinished products are the customers in a manufacturing line.

To illustrate these definitions, consider a call center: Servers are agents with associated resources (computers, desks, cubicles), customers are callers, and the network could include an interactive voice response unit with which callers interact prior to entering a queue of callers waiting to talk to an agent. In an emergency medical service (EMS) system, servers are ambulances with crews, customers are patients, and the network could include the road network or the EDs to which ambulances transport patients.

3.3. Service–time determinants

Mechanisms that originate from a system component in response to a load characteristic either increase or decrease one of the service-time determinants (work content, service speed, or in-process delay). We define *service time* as the length of time the customer spends inside the process. The process boundaries are defined by the user or investigator doing the study. The service time T begins when the customer arrives inside the process boundaries and ends when the customer leaves and consists of a combination of *processing time* P and *in-process delay* D, that is, $T = P + D$. We define an in-process delay as any period during which the customer is inside the process but no work for that customer is being done; therefore, some periods during which a customer is waiting fall under our definition of in-process delay and others do not.

We express the processing time as $P = W/S$, where W is a random amount of work that each customer brings and that needs to be completed during the service time, and the service speed S is measured in units of work per time unit. Thus, the service time is $T = P + D = W/S + D$. The mechanisms that we are interested in can impact W, S, or D.

It is often useful to decompose a service into either stages (single or multi-stage, as in Gross, Shortie, Thompson, & Harris, 2008) or phases (access, check-in, diagnosis, service delivery, and checkout, as in Bitran & Lojo, 1993). Denoting the work content, service speed, and in-process delay for Stage or Phase i by W_i, S_i, and D_i, respectively, the total service time can be presented as $T = \sum_i (W_i/S_i + D_i)$.

Decomposition into stages is especially useful for complex services (e.g., ED and EMS) where the service constitutes various individual tasks requiring different resources, each responding differently to workload. Developing an understanding of the effect of load for each service stage separately and then aggregating the effects provides a better understanding of the overall effect of system load on service times because identifying mechanisms is easier for simpler tasks. For example, Batt and Terwiesch (2016) decompose the ED process into "waiting," "treatment," and "boarding" stages. As another example, the service time for an ambulance call can be decomposed into several stages, including "travel time to scene" and "scene time." The travel time has natural measures for work content (the travel distance) and service speed (the average speed of the ambulance), as discussed in Delasay, Ingolfsson, and Schultz (2016b). Although the work content of scene time is more difficult to quantify, it could be related to the patient priority category assessed using standard triage scales.

Decomposition of the processing time for a stage into work content and speed could lead to better fitting or more parsimonious empirical models. For example, processing time = (work content)/speed = $(a + b \times \text{load})/(c + d \times \text{load})$ might provide a well-fitting model of how processing time depends on load, even if "processing time = $e + f \times$ load" provides a poor fit. That is, work content and speed might each have a simple relation with load even if processing time does not.

Depending on the purpose of the analysis and availability of data, modelers may define the servers and the customers differently for the same physical system. A particular time interval could be viewed as processing time in one model and in-process delay in another model. For example, time spent by an ED patient waiting for test results can differ depending on one's definition of the server. The time is in-process delay if one views the patient's bed as the server, but processing time if the server is seen as the resource that prepares and delivers the test results.

4. Mechanisms

Having developed the LEST framework, we analyze the existing empirical literature to show how this body of work fits into the framework and to define mechanisms which cause service times to vary. It is possible to understand the relationship between load and service times only by understanding the mechanisms. Different mechanisms could be involved at different stages of a service encounter, and their identification is crucial in explaining why the overall relationship between load and service times is positive, negative, or non-monotonic. We explore the mechanisms by reviewing published empirical papers that document dependency of service times on system load and demonstrate how these papers are connected through the identified mechanisms and the LEST framework. We attempted to be comprehensive in listing possible mechanisms and in reviewing the papers that document them. For this, we used a snowball search strategy, starting with empirical articles published in top OR/OM journals during the last 10 years, such as Kc and Terwiesch (2009). We systematically reviewed references that were cited in or cited by these articles, in search of articles, working papers, or books that reported on empirical studies of the impact of load on service time. Using this strategy, we found relevant sources covering a long time span (dating back to 1954), from a wide range of outlets (including top OM/OR journals but also articles from the medical sciences, psychology, ergonomics, and economics). Although our focus was on reviewing empirical evidence, we also cite several analytically-focused articles, in order to explain ideas and mechanisms.

The nine cells in Table 1 correspond to all combinations of the three load characteristics and the three system components. In each cell, we classify the hypothesized mechanisms that we identified based on the service-time determinants. After the name of each mechanism, we indicate whether the corresponding mechanism increases (↑) or decreases (↓) service time and we cite authors who discuss this mechanism. Some citations have a "-" superscript to indicate that data analysis failed to support the hypothesized mechanism. Mæstad, Torsvik, and Aakvik (2010) is an example of a study with "-" superscript. These authors hypothesized that physicians exert less effort on diagnosis when responsible for more patients but their data failed to support this hypothesis.

We discuss the mechanisms in the Server, Network, and Customer columns of Table 1 in Sections 4.1, 4.2, and 4.3, respectively. In listing the mechanisms, we use (S), (W), and (D) to indicate whether the mechanism impacts the service speed, the work content, or the in-process delay.

4.1. Server mechanisms

In total we identify fifteen server mechanisms. We review three server–changeover mechanisms in Section 4.1.1, nine server–load mechanisms in Section 4.1.2, and three server–extended load mechanisms in Section 4.1.3.

4.1.1. Server–changeover mechanisms

(W) Physical setup: Additional tasks required when changing to service a different customer class or switching from non-idle to idle. When servers run out of customers of a certain type they may incur a time penalty as they switch to serve a new customer type. We use the term "physical setup" to differentiate it from setup due to cognitive sharing, which we discuss later partly under "forgetting" and partly under "multitasking–cognitive sharing." Researchers have long argued for the productivity benefits of reducing physical setups through strategies like *specialization* and *mass production* (Cellier & Eyrolle, 1992; Schultz et al., 2003).

(W) Forgetting: Loss of required information from immediate memory. Forgetting involves adding tasks, rereading a chart for instance, which increases the work content and service time. When servers take a break from their main duty, they may forget particulars relevant to the task. Forgetting is relevant in situations that involve multitasking but also in situations where the server has responsibility for only one customer at a time. Time taken to remember the operation involves extra work (including mental setup) that increases work content (Steedman, 1970). In their field study in an information technology (IT) and accounting company, Mark, Gonzalez, and Harris (2005) found that workers who switch tasks could experience an average resumption lag of 25 minutes when they return to their original task because of the time needed to remember what they were doing originally. Bailey (1989) and Lu (2013) revealed that forgetting is a function of break time: The longer the break, the longer the processing time penalty. Schultz et al. (2003) tested the forgetting mechanism in a laboratory setting of a low-inventory serial production line. Although their experiments showed that breaks lead to significantly longer processing times, they did not support an association between time penalty and length of break for short breaks (another example of a study that failed to support a hypothesised mechanism). Ibanez, Clark, Huckman, and Staats (2017) found that radiologists take twice as long to read the digital images for a case immediately after a break than otherwise, and they take 1.7% less time for a case that is a repetition of the prior case type, rather than a new case type. It is not clear, however, whether the increase in radiologist reading time is because of an increase in work content or a decrease in speed. Possibly, this is an example of the loss of rhythm mechanism, which we discuss next.

(S) Loss of rhythm: Time penalty due to a break in the rhythm of work. The time penalty is independent of the break length. Breaks interrupt that rhythm and lower service speed until the rhythm is regained (Rubinstein, Meyer, & Evans, 2001). Staats and Gino (2012) analyzed loan-processing times in a bank and found that the assignment of a variety of tasks to employees results in higher average completion times. Schultz et al. (2003) provided evidence for loss of rhythm in a low-inventory serial line, noting that the time penalty appeared independent of the break length.

4.1.2. Server–load mechanisms

(W) Task reduction: Terminating service before completion or eliminating one or more discretionary service steps. This is also known as *cutting corners* (Oliva & Sterman, 2001). Task reduction is more common in professional services with discretionary task completion criteria; thus, professionals use their subjective judgment to decide which tasks need to be completed. Hopp, Iravani, and Yuen (2007) formulated an analytical model of a service with discretionary tasks and proved that task reduction is optimal if service value is concave-increasing and cost is increasing in service time. Stidham Jr. and Weber (1989), George and Harrison (2001), and Alizamir, de Véricourt, and Sun (2013) reached similar conclusions.

Early discharge (referred to as "demand-driven discharge" in Chan, Green, Lekwijit, Lu, & Escobar, 2018) is the manifestation of task reduction in healthcare systems through which healthcare professionals ration the capacity of medical units during busy periods. Kc and Terwiesch (2009) and KC and Terwiesch (2012) associated the shorter LOS of cardiothoracic surgery and intensive care unit (ICU) patients at high occupancy levels with early discharge decisions made to increase bed availability for incoming patients. Kuntz, Mennicken, and Scholtes (2014) and Berry Jaeker and Tucker (2017) observed a negative association between hospital LOS and bed occupancy above a tipping point, which they explained as being due to early discharges. Berry Jaeker and Tucker (2017) further observed that at very high occupancy, the supply of patients that can be discharged early might be exhausted, leading to a second tipping point—an effect that they refer to as saturation. Delasay et al. (2016b) mentioned that "ED surge capacity protocols" could encourage early discharging of ED patients to accelerate admission of a patient transferred by ambulance when the EMS system is under high load. Not all studies have found empirical support for early discharge, however—Chan et al. (2018) did not find evidence for an effect of congestion on ICU LOS, and Long and Mathews (2017) showed that the shorter LOS in response to higher load does not affect the amount of provided care (work content) and it only affects the *boarding time* (the time patients await transfer to other hospital units). We elaborate on this when we discuss *workload smoothing*; a load–server–in-process delay mechanism.

Batt and Terwiesch (2016) found that the number of diagnostic tests for low-acuity patients decreases with the ED waiting room census. Similarly, KC (2014) showed that physicians spend less time on patient diagnosis when they are assigned to treat several patients simultaneously. This could have negative effects on care quality. In contrast, Forster, Stiell, Wells, Lee, and Van Walraven (2003) failed to find evidence to support the impact of hospital occupancy on the proportion of ED patients who are referred to hospital consultants for diagnosis. Mæstad et al. (2010) also observed no association between physician multitasking and diagnostic effort per patient, measured by the number of relevant questions asked and the number of examinations performed.

Outside healthcare, Oliva and Sterman (2001) and Oliva (2001) found that workers in back-office operations for a bank spent less time per order when load was higher and they attribute this to eliminating such tasks as post-service documentation.

(W) Engagement: Increased attention to all tasks caused by increased workload. When lines are short, an increase in the load causes servers to become engaged, spending more time and expending greater effort to improve quality or earn more income. Hopp et al. (2007) formulated this mechanism in their model for discretionary services where there is a tradeoff between quality and speed, and Debo, Toktay, and Van Wassenhove (2008) included it in a model of credence services in which servers could spend extra time on unnecessary tasks to earn more income and customers cannot verify the appropriateness of the amount of provided service (e.g., medical and car repair). The extra effort involved in the engagement mechanism could also be stimulated in response to a perceived challenge of managing tasks in higher workloads (Bendoly, 2011; Deci, Connell, & Ryan, 1989). Tan and Netessine (2014) reported that assigning more diners to a waiter prolongs the duration of the diners' meal (as long as the restaurant is not highly congested) as it encourages the waiter to exert more upselling effort; the adjusted hourly sales per waiter increase with load. Delasay et al. (2016b) related the longer average scene time as the fraction of busy ambulances increases (when the EMS load is not critically high) to the engagement mechanism by the paramedics to prevent the need for hospital transportation.

(W) Server early task initiation: Performing some stages or tasks of a service earlier than usual. Batt and Terwiesch (2016) found that ED triage nurses order more diagnostic tests for patients when the ED waiting room is more crowded, in order to shorten the

LOS by making the test results ready by the time a physician sees the patient. Delasay et al. (2016b) observed that chute times, the preparation for the ambulance crew after receiving the dispatch notification, are shorter when EMS load is high, which is consistent with ambulance crews anticipating the receipt of a dispatch notification when most of the other ambulances are in service.

(W) Multitasking–cognitive sharing: Loss of required information due to having simultaneous responsibility for multiple customers with different service needs. Humans tend to multitask under high load. For example, an ED physician who treats several patients at the same time can see one patient while waiting for test results for another patient. Despite the presumed productivity benefits of multitasking (Lindbeck & Snower, 2000), it can hurt productivity due to three distinct mechanisms "multitasking–cognitive sharing," "multitasking–time sharing," and "multitasking–interruptions." We discuss cognitive sharing here and time sharing and interruptions later in this section under server–load–*in-process delay*. Researchers who studied multitasking have not always distinguished clearly among these three mechanisms. We separate them because their causes are different and to encourage future researchers to distangle their effects.

Psychological studies recognize additional effort needed to refocus on an interrupted task (cognitive sharing) as the main reason for productivity loss due to multitasking (Gladstones, Regan, & Lee, 1989; Pashler, 1994; Rubinstein et al., 2001). KC (2014) related the productivity loss in ED physicians with case load (the number of patients assigned to a physician) of more than six patients partly to cognitive sharing. Studying multitasking in a recruiting firm, Aral, Brynjolfsson, and Alstyne (2012) found that a small amount of multitasking improves the number of vacancies filled per unit time but that excessive multitasking results in longer unit time to fill a specific vacancy. They related this, partly, to cognitive switching among multiple projects. Lu (2013) related longer request completion times for agents with higher case loads, partly, to cognitive sharing involved in frequent task suspensions. She reported more pronounced productivity losses for longer suspension times.

It is important to make the distinction between the multitasking–cognitive sharing and forgetting mechanisms clear. As a server's case load increases, the task of reviewing status when switching between customers increases (Monsell, 2003; Rubinstein et al., 2001). To the extent that this is a *variable* cost based on load (as in multitasking–cognitive sharing), this is a server–*load*–work content mechanism. To the extent this is a fixed cost of changing between customers (as in forgetting), this is a server–*changeover*–work content mechanism.

(S) Social speedup pressure: People feel pressure to speed up in order to avoid delaying the service of others. *Speedup* is common in systems in which server performance is visible to others. Slower servers work faster when performance feedback is available (Bandiera, Barankay, & Rasul, 2013; Schultz et al., 2003). Edie (1954) demonstrated that under the pressure of backed-up traffic, toll collectors at the George Washington Bridge expedited service, by limiting conversation with drivers, and Mas and Moretti (2009) found that slow supermarket cashiers speed up when customers are backed up and the slow cashiers are seen by faster cashiers. In a supermarket with dedicated queues, Wang and Zhou (2018) replicated the finding that supermarket cashiers speed up when more customers are waiting. Kc and Terwiesch's (2009) regression models revealed that in-hospital patient transporters speed up in response to load, defined as the fraction of busy transporters, and researchers have observed speedups when more requests are assigned to agents of a bank (Staats & Gino, 2012) or an IT firm (Lu, 2013). Schultz et al. (1998) demonstrated through laboratory experiments that workers in a low-inventory serial line, where the WIP in between-station buffers can be traced, work faster when they are causing blockage of a preceding station or starvation of a succeeding station. Using data from an e-mail contact center, capacity estimation models in Hasija, Pinker, and Shumsky (2010) showed that the processing rate of any server could roughly double as the center is staffed with fewer servers (more load on each server results in speedup). Workers have been shown to work faster in systems with visible queues (Shunko, Niederhoff, & Rosokha, 2018). Though waiters were shown to increase upselling as load increases from low to medium (as we discussed under engagement), they accelerate service by reducing upselling efforts as load increases past some tipping point (Tan & Netessine, 2014).

(S) Social loafing: Servers exert less effort when their effort is difficult to monitor. *Social loafing*, a.k.a *free riding*, occurs when servers decrease their efforts to avoid pulling the weight of a fellow team member (Karau & Williams, 1993). Social loafing is prevalent in congested systems where individual effort is difficult to monitor (Latane, Williams, & Harkins, 1979). Mas and Moretti (2009) argued that supermarket cashiers slow down to let other cashiers handle the additional workload during peak periods. Wang and Zhou (2018) observed that when the queue configuration in the supermarket in their study was switched from dedicated queues to a pooled queue (but a longer one), social loafing dominates social pressure speedup resulting in the average service time that increases with the queue length. Berry Jaeker and Tucker (2012) found that hospital nurses work slower intentionally to avoid being assigned new patients when they predict that a large number of patients will be admitted from the ED.

(D) Multitasking–time sharing: Sharing server capacity among multiple customers. At each instant, one customer has the attention of the server. Tan and Netessine (2014) mentioned time sharing as a possible reason for prolonged meal duration of diners assigned to a waiter serving several tables simultaneously. The relationship between case load and productivity may not be linear. Aral et al. (2012) found in a recruiting firm that excessive multitasking results in longer duration to fill a specific vacancy. Besides relating it to cognitive sharing, they also mentioned the reason as the delay recruiters face in returning to the activities of one project while cycling through activities of other projects. Lu (2013) found that a higher case load increases the revisit time for services that were suspended due to interruptions. Goes, Ilk, Lin, and Zhao (2018) confirmed that time sharing due to multitasking causes delays in live-chat agents' responses to customers, and showed that the marginal effect of more multitasking on delays is increasing.

(D) Multitasking–interruptions: Momentary pauses in a service interaction with a customer because the server needs to attend to requests from other customers. Using a time and motion study, Chisholm, Collison, Nelson, and Cordell (2000) found a positive correlation between physician case load and number of interruptions that require the physician's attention. Though Lu (2013) found that a higher case load increases the duration of interruptions (as we discussed under multitasking–time sharing), she did not find evidence for the effect of case load on the frequency of interruptions. KC (2014) also discussed the negative impacts of interruptions on physicians' productivity and longer LOS.

To explain how time sharing, interruptions, and cognitive sharing interact, suppose that a server with case load N meets twice with each customer. A simple representation of each customer's total service time is $T_1 + T_2 + T_3$, where T_1 and T_3 are time intervals the server spends with the customer, and T_2 is the time interval between the two meetings with the server, during which the customer might receive service from another resource or wait for the server to finish meeting with some of the other $N-1$ customers. In this fashion, service time is the total time until the end of the last meeting with the server, and T_2 is the in-process delay. A higher case load N increases T_2 through the time sharing

mechanism, because the server needs to serve a larger number of the other customers, on average, before returning to the customer of interest. A higher case load causes T_1 and T_3 to be longer, on average, through the interruptions mechanism, because the frequency of interruptions from other customers increases. And finally, a higher case load can increase the work content of the task of reviewing status when switching between customers, through cognitive sharing, causing T_3 to be longer.

(D) Workload smoothing: Completing the discharge of customers whose processing is complete earlier, in anticipation of incoming demand. Long and Mathews (2017) observed that ICU patients often board while awaiting transfer to other hospital units. By decomposing ICU LOS into care time and boarding time, they found that the shorter LOS during higher ICU occupancy is not related to early discharge decisions (in contrast to findings by other researchers about early discharge, which we cited under the task reduction mechanism) but is instead caused by decreased boarding times of patients that are discharged in order to free ICU capacity. This study highlights the importance of decomposing service time into processing time and in-process delay, as we propose in the LEST framework. Similarly, Kim, Chan, Olivares, and Escobar (2014) found that the impact of load on ICU care time depends on the medical condition and the origin of patients that are about to be admitted. Berry Jaeker and Tucker (2012) reported that hospital medical teams react to a high volume of incoming scheduled patients from a surgery unit by discharging patients earlier.

4.1.3. Server–extended load mechanisms

(W) Service cancelation: Denying service to a customer to obtain extra rest or improve metrics. *Overwork* and the consequent productivity deterioration is the natural outcome of working for long periods (Çakir, Hart, & Stewart, 1980; Setyawati, 1995). Overworked servers may simply refuse to serve a customer in order to obtain extra rest. Brown et al. (2005) discovered this phenomenon when they encountered call times of less than 10 seconds in a call center's data—caused by overworked agents who hung up on customers in order to reduce workload.

(S) Learning by doing: Productivity gains through learning over short horizons (within a shift, for instance). Extended load can result in productivity gains through learning by doing, in contrast to fatigue. Learning can occur over time horizons as short as the length of a shift or as long as months or even years. The higher cumulative number of service completions, for instance, the greater the productivity gains through long-term learning for medical teams (Pisano, Bohmer, & Edmondson, 2001) and call center agents (Gans et al., 2010). In support of short-term learning productivity gains, Lu (2013) reported shorter completion times of requests handled later in a shift in a call center.

(S) Fatigue: The inability of servers to maintain high levels of effort over extended periods. Speeding up cannot be sustained indefinitely; when servers are overworked, they slow down (Dietz, 2011; Sze, 1984). As Kc and Terwiesch (2009) showed, hospital transporters slow down after experiencing extended high-load periods and overworked physicians delay discharge decisions for surgery patients. Staats and Gino (2012) observed the same kind of slowdown behavior by loan processors in a bank. Similarly, Gans et al. (2010) found call times in a call center to be positively associated with the number of calls an agent has answered since the last break period.

4.2. Network mechanisms

In total we identify six network mechanisms. We review one network–changeover mechanism in Section 4.2.1, four network–load mechanisms in Section 4.2.2, and one network–extended load mechanism in Section 4.2.3.

4.2.1. Network–changeover mechanisms

(W) Network arrangement: Planned positioning of servers to improve response for future customers. Delasay et al. (2016b) considered an EMS system as a network of ambulance locations, patient addresses, and hospital locations connected via city roads. *Network arrangement* refers to the positioning of servers (ambulances) at planned locations in order to reduce average travel time to future customers (patients).

4.2.2. Network–load mechanisms

(W) Geographical dispersion: Longer response times because of fewer available servers that are more widely dispersed. Delasay et al. (2016b) proposed *geographical dispersion* as the reason for longer ambulance travel distance to a scene when EMS load is higher. High EMS load means fewer available ambulances to cover a city. Geographic dispersion is also relevant for an array of services, including repair- and tow-truck services, hospital porter services, taxi and delivery services, and fire and police.

(S) Geographical speedup: Higher average travel speed for longer trips. *Geographical speedup* mitigates geographical dispersion by enabling ambulance crews to travel at higher speeds on longer trips that involve at least some highway or main artery travel (Budge, Ingolfsson, & Zerom 2010; Delasay et al. 2016b).

(D) Resource sharing: Congestion due to sharing a resource with other processes. If multiple nodes in a network share a common resource, then an increase in load can prolong in-process delay at the individual nodes, through *resource sharing*. An example of resource sharing, as modeled in Akşin and Harker (2003), is the simultaneous use of a common information processing resource by several agents in a call center, which incurs additional delay in call durations. Hillier et al. (2009) found that high hospital occupancy not only prolongs ED LOS of admitted patients to the hospital but also increases LOS of patients discharged from the ED; they indicate one possible explanation could be that the ED and the hospital share such resources as treatment areas, lab services, and care providers.

(D) Downstream system congestion: Congestion in a service due to back up of customers waiting to proceed to a downstream service. When the various stages of a service are provided at interrelated nodes of a queueing network, load at one node may impact service times at other nodes. *Downstream system congestion* involves resources being tied up serving customers who cannot be admitted to a congested downstream node. Viewing an ED and a hospital as nodes of a network, Forster et al. (2003) and Hillier et al. (2009) showed that extremely high hospital ward occupancy rates increase the LOS of those ED patients who have been admitted to a hospital ward. Asaro, Lewis, and Boxerman (2007), Louriz et al. (2012), Kim et al. (2014), and Long and Mathews (2017) isolated the boarding time component of the LOS and documented the positive association between boarding time in a hospital unit and the occupancy of downstream units. In the context of EMS, Delasay et al. (2016b) observed ambulance paramedics need to wait longer to offload patients when EDs are crowded.

4.2.3. Network–extended load mechanisms

(W) Network chaos: Increased deviation of server locations from planned positions because of extended load. *Network chaos* is the opposite of network arrangement. If EMS load remains high for a long period, ambulance locations are more likely to deviate from their planned positions, leading to longer travel times (Delasay et al., 2016b).

4.3. Customer mechanisms

In total we identify four customer mechanisms. We review one customer–changeover mechanism in Section 4.3.1, two customer–

load mechanisms in Section 4.3.2, and one customer–extended load mechanism in Section 4.3.3.

4.3.1. Customer–changeover mechanisms

(W) Customer early task initiation: Shorter processing times due to customers using waiting time (pre-process delay) to perform tasks that would otherwise be done during service time. Edie (1954) proposes this mechanism as the reason for shorter holding times at toll booths at higher volumes. If there is no line, then drivers have to search to find their tolls after they drive up to the booth. Wang and Zhou (2018) postulate that the same mechanism applies to supermarket customers waiting to be served by a cashier.

4.3.2. Customer–load mechanisms

(W) Return: Customers returning to the process because they were not adequately served during the previous encounter. As discussed under the task reduction mechanism, servers may end a service encounter prematurely, which may degrade service quality and cause customers to return to the system, thereby prolonging total service time. KC and Terwiesch (2012) and KC (2014) documented returns by showing that the likelihood of a patient revisiting EDs and ICUs soon after discharge increases with the load at discharge time. Lower quality of care because of excessive multitasking is cited as the reason for returns in KC (2014), and early discharge decisions are associated with returns in KC and Terwiesch (2012). As we discussed under task reduction, unlike KC and Terwiesch (2012) and KC (2014), Long and Mathews (2017) isolated the boarding time from ICU LOS and found no evidence for an effect of load on revisit rates. Motivated by these empirical results, Chan, Yom-Tov, and Escobar (2014) formulated queueing models to investigate the implications of early discharge decisions and associated returns for long-term system behavior.

(D) Abandonment: Customers leaving the process, without receiving complete service, due to long waiting time (in-process delay) and customers' limited patience. Balking can be viewed as a special case of abandonment, in which a customer abandons immediately upon arrival. Balking (a customer mechanism) is the counterpart to service cancelation (a server mechanism). Abandonment during an in-process delay not only affects the service time of the current customer but also affects the service time of the following customers, and hence the system-wide average service time.

Abandonment is mostly documented as a mechanism occurring during pre-process delay (before service begins). For example, Batt and Terwiesch (2015) showed that queue length (among other measures of load) affects abandonment from an emergency department waiting room (a semi-observable multi-class queue), De Vries, Roy, and De Koster (2018) found that queue length affects abandonment from a restaurant (a semi-observable single-class queue), and Lu, Musalem, Olivares, and Schilkrut (2013) related the nonlinear and decreasing effect of queue length on purchasing incidents at a grocery deli counter (an observable single-class queue) to the balking decisions of the customers in response to longer wait times. Though we are not aware of any empirical studies on abandonment during an in-process delay, the same factors causing abandonments during pre-process delay could hypothetically impact abandonment during in-process delay; for example, in multi-stage systems and during the wait to receive service from a downstream node. This could be a future research direction worth exploring.

4.3.3. Customer–extended load mechanisms

(W) Deterioration: Additional processing required when overwork results in reduced service quality. When overwork is associated with reduced service quality, additional processing may be required. After showing that system-level overwork increases the LOS for surgery patients, Kc and Terwiesch (2009) argued that fatigued care providers are more prone to making medical errors, leading to complications that call for additional processing.

We note that the return and deterioration mechanisms involve a chain of load-induced behaviors, first in the servers and subsequently in the customer. We classify these mechanisms based on the final system component in the chain, that is, the customer. The reason is that in order to understand these mechanisms, researchers need to study conditions experienced by customers that would cause them to return, or that would cause their condition to deteriorate.

4.4. Summary of mechanisms

Server mechanisms. Load effects on servers have been mixed. All server–changeover mechanisms that we identified increased service time. Yet, it is not difficult to imagine a case in which the server takes advantage of a break in order to prepare for the next service. We observed mixed impacts for server–load mechanisms: four mechanisms decrease the service time and five mechanisms increase it. A common feature of server–load mechanisms is the availability of performance feedback; a feature of queueing systems rarely included in mathematical models. Servers react to load if their performance is observable by others. Some server–load mechanisms decrease service time for a single service encounter, but they cause deterioration in service quality that may require customers to revisit the system later. If we track the effects of these mechanisms over multiple service encounters, therefore, we may find longer total service time. Most studies in this area document servers' reactions to extended load to be slowdown due to overwork, which increases service time. Although servers learn and gain a rhythm when doing a specific job over a long period, fatigue is likely to be the dominant mechanism if the high-load period is sufficiently long.

Network mechanisms. Although analytical models of queueing networks abound, empirical research on network mechanisms is not common. Most of the empirical papers that we reviewed are in medical journals. These papers demonstrate the impact of downstream system congestion and resource sharing on prolonging service times when two systems are connected in series, as in a standard tandem queueing network.

Customer mechanisms. We identified far fewer customer mechanisms than server mechanisms. The reasons could include (1) customers having less control over service encounters, (2) greater interest in servers because they are subject to managerial control, and (3) customer data being more difficult to obtain, because of privacy regulations. In addition, we found no customer mechanisms that influence service speed. We expect they exist but have not been studied in the literature.

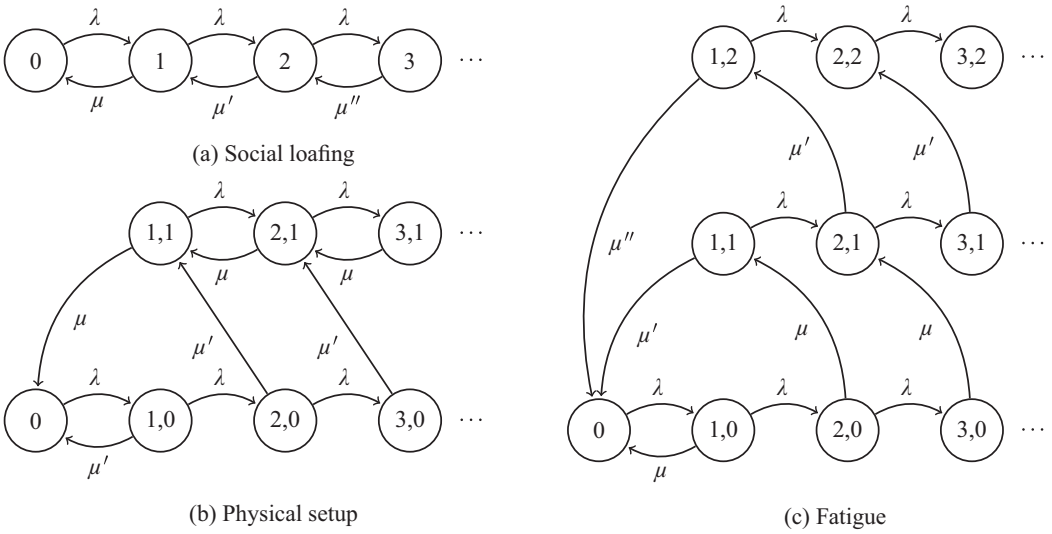
We conclude this section by summarizing the definitions of the identified mechanisms in Table 2. To the extent that these definitions become standard, and the LEST framework is accepted, we now have the ability to categorize future research. This will allow researchers to identify those papers that extend theory on known mechanisms and those papers that attempt to identify new mechanisms. To the extent that these mechanism definitions are applied and are searchable, empirical and analytical researchers will be able to gather and review relevant studies for each mechanism, helping us to identify the frequency, significance, and breadth of these phenomena.

5. Implications for modeling

Translating findings from empirical research into mathematical queueing models is not always straightforward. Most empirical

Table 2
Mechanisms.

Mechanism	Framework Cell	Definition
Physical setup	Changeover, Server, Work content (↑)	Additional tasks required when changing to service a different customer class. (Schultz et al. 2003)
Forgetting	Changeover, Server, Work content (↑)	Loss of required information from immediate memory. Involves a fixed cost in speed of switching from one task type to another. The delay is a function of changeover. (Bailey 1989; Ibanez et al. 2017; Mark et al. 2005; Schultz et al. 2003)
Loss of rhythm	Changeover, Server, Service speed (↑)	Time penalty due to a break in the rhythm of work. The time penalty is independent of the break length. (Schultz et al. 2003; Staats & Gino 2012)
Task reduction	Load, Server, Work content (↓)	Terminating service before completion or eliminating one or more discretionary service steps. Also known as cutting corners. (Batt & Terwiesch 2016; Chan et al. 2018; Delasay et al. 2016b; Forster et al. 2003; KC 2014; Kc & Terwiesch 2009; KC & Terwiesch 2012; Kuntz et al. 2014; Long & Mathews 2017; Mæstad et al. 2010; Oliva 2001; Oliva & Sterman 2001)
Engagement	Load, Server, Work content (↑)	Increased attention to all tasks caused by increased workload. (Delasay et al. 2016b; Tan & Netessine 2014)
Server early task initiation	Load, Server, Work content (↓)	Performing some stages or tasks of a service earlier than usual. (Batt & Terwiesch 2016; Delasay et al. 2016b)
Multitasking–cognitive sharing	Load, Server, Work content (↑)	Loss of required information due to having simultaneous responsibility for multiple customers with different service needs. (Aral et al. 2012; KC 2014; Lu 2013)
Social speedup pressure	Load, Server, Service speed (↓)	People feel pressure to speed up in order to avoid delaying the service of others. (Batt & Terwiesch 2016; Edie 1954; Kc & Terwiesch 2009; Lu 2013; Mas & Moretti 2009; Schultz et al. 1998; Shunko et al. 2018; Staats & Gino 2012; Tan & Netessine 2014; Wang & Zhou 2018)
Social loafing	Load, Server, Service speed (↑)	Servers exert less effort when their effort is difficult to monitor. (Berry Jaeker & Tucker 2012; Mas & Moretti 2009; Wang & Zhou 2018)
Multitasking–time sharing	Load, Server, In-process delay (↑)	Sharing server capacity among multiple customers. At each instant, one customer has the attention of the server. (Aral et al. 2012; Goes et al. 2018; KC 2014; Lu 2013; Tan & Netessine 2014)
Multitasking–interruptions	Load, Server, In-process delay (↑)	Momentary pauses in a service interaction with a customer because the server needs to attend to requests from other customers. (Chisholm et al. 2000; KC 2014; Lu 2013)
Workload smoothing	Load, Server, in-process delay (↓)	Completing the discharge of customers whose processing is complete earlier, in anticipation of incoming demand. (Berry Jaeker & Tucker 2012; Kim et al. 2014; Long & Mathews 2017)
Service cancelation	Extended load, Server, Work content (↓)	Denying service to a customer to obtain extra rest or improve metrics. (Brown et al. 2005)
Learning by doing	Extended load, Server, Service speed (↓)	Productivity gains through learning over short horizons (within a shift, for instance). (Lu 2013)
Fatigue	Extended load, Server, Service speed (↑)	The inability of servers to maintain high levels of effort over extended periods. (Gans et al. 2010; Kc & Terwiesch 2009; Staats & Gino 2012)
Network arrangement	Changeover, Network, Work content (↓)	Planned positioning of servers to improve response for future customers. (Delasay et al. 2016b)
Geographical dispersion	Load, Network, Work content (↑)	Longer response times because of fewer available servers that are more widely dispersed. (Delasay et al. 2016b)
Geographical speedup	Extended load, Network, Service speed (↑)	Higher average travel speed for longer trips. (Delasay et al. 2016b)
Resource sharing	Load, Network, In-process delay (↑)	Congestion due to sharing a resource with other processes. (Hillier et al. 2009)
Downstream system congestion	Load, Network, In-process delay (↑)	Congestion in a service due to back up of customers waiting to proceed to a downstream service. (Asaro et al. 2007; Delasay et al. 2016b; Forster et al. 2003; Hillier et al. 2009; Kim et al. 2014; Long & Mathews 2017; Louriz et al. 2012)
Network chaos	Extended load, Network, Work content (↓)	Increased deviation of server locations from planned positions because of extended load. (Delasay et al. 2016b)
Customer early task initiation	Changeover, Customer, Work content (↓)	Shorter processing times due to customers using waiting time (pre-process delay) to perform tasks that would otherwise be done during service time. (Edie 1954; Wang & Zhou 2018)
Return	Load, Customer, Work content (↑)	Customers returning to the process because they were not adequately served during the previous encounter. (KC 2014; KC & Terwiesch 2012; Long & Mathews 2017)
Abandonment	Load, Customer, In-process delay (↓)	Customers leaving the process, without receiving complete service, due to long waiting time (in-process delay) and customers' limited patience. (Batt & Terwiesch 2015; De Vries et al. 2018; Lu et al. 2013)
Deterioration	Extended load, Customer, Work content (↑)	Additional processing required when overwork results in reduced service quality. (Kc & Terwiesch 2009)



(a) Social loafing

(b) Physical setup

(c) Fatigue

Fig. 2. Illustrative Markov chain models.

research focuses on how load (and other factors) impacts service *times*, whereas mathematical queueing models are typically formulated in terms of service *rates*. Nevertheless, empirical research that documents fundamental mechanisms through which service times increase or decrease with load can help modelers formulate models that at least capture directional effects, if one assumes that service rates move in the opposite direction to service times—an assumption that we will revisit at the end of this section.

For example, empirical work has shown that average service time increases when the *social loafing* mechanism is in effect (Section 4.1.2), when a *physical setup* is incurred (Section 4.1.1), and when servers become *fatigued* (Section 4.1.3). Although the direction of change in the average service time is the same, modeling these three mechanisms requires different approaches. Starting with a standard $M/M/1$ system with arrival rate λ and service rate μ and a single state variable, I, representing the number of customers in the system, the following modifications provide possible Markov chain models of these three mechanisms.

Social loafing: The service rate decreases with I (Fig. 2a), to capture the fact that the server decreases her speed when the queue builds up ($\mu > \mu' > \mu''$). See Do, Shunko, Lucas, and Novak (2018) for a model of a multi-server parallel-queue system consisting of multiple copies of this type of model, one for each server, with a service rate that decreases with the number of servers.

Physical setup: A binary state variable indicates whether the server was busy when the customer that is currently receiving service arrived (Fig. 2b). The service rate, $\mu > \mu'$, is higher for such customers. See Welch (1964) for a more general version of this model.

Fatigue: An additional state variable, J, counts the number of customers served during the current busy period, as a measure of fatigue (Fig. 2c). The service rate decreases as J increases ($\mu > \mu' > \mu''$). See Delasay et al. (2016a) for a multi-server version of this model.

Fig. 2a–2c illustrate how queueing modelers have revisited traditional queueing models to capture empirical findings. Each of these models captures a single mechanism. More complex models can capture multiple mechanisms—for example, Delasay et al. (2016a) model social speedup pressure and fatigue together, and Do et al. (2018) model social speedup pressure and social loafing together.

In all of these illustrative models, the empirical finding was that service time increased with a change in a load characteristic, and the models captured this with a service rate that decreases with the load characteristic change. Wang and Zhou (2018) demonstrate that if the service rates increase with I, then the mean service times decrease with I, for a model that has the structure shown in Fig. 2a. Future research should investigate whether similar results hold for models with multiple state variables, such as the ones in Fig. 2b–2c.

Future research should also investigate how one can go beyond translating directional effects for mean service times to directional effects for service rates towards also translating the *magnitude* of effects on mean service times to magnitude for effects on service rates. One approach to that challenge is to use statistical methods that directly estimate how service rates (rather than mean service times) depend on the system state, as Azriel et al. (2014) do. Another possible approach is to develop computational methods for determining state-dependent service rates that will result in a set of desired values for state-dependent mean service times. This approach has not been investigated, to our knowledge.

6. Conclusion and future directions

Empirical researchers have recently challenged the assumption of exogenous service times in queueing models by providing evidence for dependence of service times on load in such systems as call centers, emergency rooms, and banks. Studies typically focused on the most obvious manifestation of load in queueing systems: the current congestion level and its effect on servers. A few researchers also tracked load history and found evidence for such behaviors as slowdown in response to overwork.

We proposed the LEST framework that can be employed by empirical and analytical researchers to investigate and model service time dependencies on load. The LEST framework has three dimensions: (1) load characteristics, (2) system components, and (3) service-time determinants.

In the first dimension, we identified three load characteristics: *changeover*, *load*, and *extended load*. Changeover refers to the switch from idle to busy or from one task to another, which induces mechanisms like setup. Load is the instantaneous system congestion level, and extended load is the past history of load. We found it interesting that, in a general sense, reactions to extended load relate to past events, changeover relates to the present, and load relates to customers who will receive service in the future.

In the second dimension of the LEST framework, we identified three system components: *server*, *customer*, and *network*. We recognized that servers are not the only system components that react to load. Customers do as well. The service time at a particular node in a queueing network can also depend on load at upstream and downstream nodes. Therefore, we include the network as the third system component.

In the third dimension of the LEST framework, we decomposed service time into three determinants—*work content*, *service speed*, and *in-process delay*. Researchers should strive to measure work content and service speed separately, for each stage of service, to help identify and separate mechanisms with different causes.

Organizing frameworks can have a significant impact on the direction and progress of an academic field. In queueing theory, for example, the notation introduced by Kendall (1953), has served as a powerful organizing framework for over half a century. We contend that the LEST framework provides a structure for researchers and practitioners to consider the implications of load on service times. Through queueing theory, we already have a reasonably clear understanding of the effects of congestion on waiting time. When considering the effects of congestion on service times, the LEST framework gives both researchers and practitioners the opportunity to ask the right questions by considering the mechanisms identified in this paper that may be relevant to their context. In addition, LEST helps them to undertake deeper analysis by giving them a framework for asking questions about possible new mechanisms that may apply. Delasay et al. (2016b) provides an example of a way the framework can be used to break down the relationship between load and EMS system service time, to allow for better analysis and to look for new and notable mechanisms.

While engaged in this research, we observed an apparent disconnect between analytical papers and empirical papers. Most analytical research and most textbooks are heavily influenced by the modeling tractability focus on interarrival times, processing times, and queueing disciplines. The empirical work shows the importance of single vs. multitasking servers, visibility of queues, visibility of coworkers, interruptions, single vs. multiple queues, and shared resources. The difference between the factors important in models, those that get taught, and those that show up in empirical work is uncomfortable. Although there certainly are queueing models that consider some of these factors, we believe that it is valuable to go further in developing analytical models of these phenomena. These models will not be easy to formulate or analyze, but they may prove to be of great value.

Accordingly, we return to what we consider the most important contribution of this paper. As we have demonstrated, there is no single answer to the question of "What is the effect of load on service times?" The answer is: "It depends." We believe that further research in this area should leave aside the general question and focus on the more specific ones. Although a universal theory of the effects of load on service times would be laudable, we do not believe that such a thing exists.

Rather, we would direct research into specific mechanisms. A mechanism is a link between a change in load and a change in service times. Each mechanism activates a particular set of parameters and a particular cause for that relationship. In some cases multiple papers have different names for what is, in effect, the same mechanism. This leads to serious issues with continuity of research if later researchers can only find part of the previous literature. In this paper, we identified 25 mechanisms and explained them through the LEST framework. We demonstrated how the findings of previous empirical papers are connected through the identified mechanisms and the LEST framework. By classifying the published studies according to the LEST framework, research gaps in the empirical literature are highlighted. The literature focuses primarily on the aggregate effect of load on server behavior, for example, ignoring server heterogeneity. A promising avenue for further research is an investigation of the distribution of skill levels across servers—the degree of cross training for human servers—and its impact on load mechanisms.

Not all mechanisms are equally important. Important mechanisms are those that occur frequently, have a significant impact, and improve predictability. In our opinion, task reduction, engagement, social speedup pressure, downstream system congestion, fatigue, and abandonment may qualify as important mechanisms. However identified, researchers should investigate a set of key mechanisms. Editors should recognize that significant contributions come not just from the identification of new mechanisms but also from the exploration of important, previously identified, mechanisms. Empirical researchers may ask how the parameters of the service time distributions change, identify the moderators and mediators, or examine the ways in which different people react. Analytical researchers could use that information to build models to help us understand the effects of these mechanisms on system performance. By combining the contributions of two different research methods, we can use the strengths of each, overcome the weaknesses of both, and build a better understanding of how queues work in practice.

We hope that researchers can use the LEST framework to develop new understanding of load effects. One contribution of our paper is to show a way forward for the development of different reactions of service time to load and to fit all the parts together into a comprehensive whole. We do not present theory that closes out the development of this topic. Rather, we open the doors to show how different work fits together in previously unknown ways.

The LEST framework leads to a new and important question, namely, when does the impact of one type of mechanism dominate another? It seems clear that there is not one dominating mechanism. Rather the dominant mechanism depends on the particular situation. Further research on situational factors that lead to the domination of which mechanism should be done. Such research requires a framework with which to think about and compare mechanisms and situations.

Our paper does not close the question of how service times depend on load, it repositions it; we do not provide the answers, we provide the questions.

Acknowledgments

The authors thank two anonymous referees for their constructive comments, which helped to improve the paper. This work was partially supported by the Canadian Natural Science and Engineering Research Council (Discovery Grant 203534 and the CREATE Program in Healthcare Operations and Information Management program). This support is gratefully acknowledged.

References

Akşin, O. Z., & Harker, P. T. (2003). Capacity sizing in the presence of a common shared resource: Dimensioning an inbound call center. *European Journal of Operational Research, 147*(3), 464–483. doi:10.1016/S0377-2217(02)00274-6.

Alizamir, S., de Véricourt, F., & Sun, P. (2013). Diagnostic accuracy under congestion. *Management Science, 59*(1), 157–171. doi:10.1287/mnsc.1120.1576.

Aral, S., Brynjolfsson, E., & Alstyne, M. V. (2012). Information, technology, and information worker productivity. *Information Systems Research, 23*(3-part-2), 849–867. doi:10.1287/isre.1110.0408.

Asaro, P. V., Lewis, L. M., & Boxerman, S. B. (2007). The impact of input and output factors on emergency department throughput. *Academic Emergency Medicine, 14*(3), 235–242. doi:10.1197/j.aem.2006.10.104.

Azriel, D., Feigin, P. D., & Mandelbaum, A. (2018). Erlang S: A data-based model of servers in queueing networks. Forthcoming in *Management Science*.

Bailey, C. D. (1989). Forgetting and the learning curve: A laboratory study. *Management Science, 35*(3), 340–352. doi:10.1287/mnsc.35.3.340.

Bandiera, O., Barankay, I., & Rasul, I. (2013). Team incentives: Evidence from a firm level experiment. *Journal of the European Economic Association, 11*(5), 1079–1114. doi:10.1111/jeea.12028.

Batt, R. J., & Terwiesch, C. (2015). Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science, 61*(1), 39–59. doi:10.1287/mnsc.2014.2058.

Batt, R. J., & Terwiesch, C. (2016). Early task initiation and other load-adaptive mechanisms in the emergency department. *Management Science, 63*(11), 3531–3551. doi:10.1287/mnsc.2016.2516.

Bendoly, E. (2011). Linking task conditions to physiology and judgment errors in RM systems. *Production and Operations Management, 20*(6), 860–876. doi:10.1111/j.1937-5956.2010.01213.x.

Berry Jaeker, J. A., & Tucker, A. L. (2012). Hurry up and wait: Differential impacts of congestion, bottleneck pressure, and predictability on patient length of stay. *Technical report*.

Berry Jaeker, J. A., & Tucker, A. L. (2017). Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science, 63*(4), 1042–1062. doi:10.1287/mnsc.2015.2387.

Bitran, G., & Lojo, M. (1993). A framework for analyzing the quality of the customer interface. *European Management Journal, 11*(4), 385–396. doi:10.1016/0263-2373(93)90002-Y.

Brockmeyer, E., Halstrøm, H., Erlang, A., & Jensen, A. (1948). The life and works of A.K. Erlang. *Transactions (Akademiet for de tekniske videnskaber (Denmark))*. Akademiet for de Tekniske Videnskaber.

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. *Journal of the American Statistical Association, 100*(469), 36–50. doi:10.1198/016214504000001808.

Budge, S., Ingolfsson, A., & Zerom, D. (2010). Empirical analysis of ambulance travel times: The case of Calgary emergency medical services. *Management Science, 56*(4), 716–723. doi:10.1287/mnsc.1090.1142.

Çakir, A., Hart, D. J., & Stewart, T. F. (1980). *Visual display terminals: A manual covering ergonomics, workplace design, health and safety, task organisation*. New York: John Wiley & Sons.

Cellier, J.-M., & Eyrolle, H. (1992). Interference between switched tasks. *Ergonomics, 35*(1), 25–36. doi:10.1080/00140139208967795.

Chan, C. W., Green, L. V., Lekwijit, S., Lu, L., & Escobar, G. (2018). Assessing the impact of service level when customer needs are uncertain: An empirical investigation of hospital step-down units. *Management Science*. doi:10.1287/mnsc.2017.2974.

Chan, C. W., Yom-Tov, G., & Escobar, G. (2014). When to use speedup: An examination of service systems with returns. *Operations Research, 62*(2), 462–482. doi:10.1287/opre.2014.1258.

Chisholm, C. D., Collison, E. K., Nelson, D. R., & Cordell, W. H. (2000). Emergency department workplace interruptions: Are emergency physicians "interrupt-driven" and "multitasking"? *Academic Emergency Medicine, 7*(11), 1239–1243. doi:10.1111/j.1553-2712.2000.tb00469.x.

De Vries, J., Roy, D., & De Koster, R. (2018). Worth the wait? How restaurant waiting time influences customer behavior and revenue. *Journal of Operations Management*. doi:10.1016/j.jom.2018.05.001.

Debo, L. G., Toktay, L. B., & Van Wassenhove, L. N. (2008). Queueing for expert services. *Management Science, 54*(8), 1497–1512. doi:10.1287/mnsc.1080.0867.

Deci, E. L., Connell, J. P., & Ryan, R. M. (1989). Self-determination in a work organization. *Journal of Applied Psychology, 74*(4), 580.

Delasay, M., Ingolfsson, A., & Kolfal, B. (2016a). Modeling load and overwork effects in queueing systems with adaptive service rates. *Operations Research, 64*(4), 867–885. doi:10.1287/opre.2016.1499.

Delasay, M., Ingolfsson, A., & Schultz, K. (2016b). *Inventory is people: How load affects service times in emergency response* (pp. 21–49). World Scientific/NOW Publishers.

Dietz, D. C. (2011). Practical scheduling for call center operations. *Omega, 39*(5), 550–557. doi:10.1016/j.omega.2010.12.001.

Do, H. T., Shunko, M., Lucas, M. T., & Novak, D. C. (2018). Impact of behavioral factors on performance of multi-server queueing systems. *Production and Operations Management, 27*(8), 1553–1573. doi:10.1111/poms.12883.

Dshalalow, J. H. (1997). Queueing systems with state dependent parameters. *Frontiers in queueing: Models and applications in science and engineering* (pp. 61–116). CRC Press.

Edie, L. C. (1954). Traffic delays at toll booths. *Journal of the Operations Research Society of America, 2*(2), 107–138. doi:10.1287/opre.2.2.107.

Fisher, M. (2007). Strengthening the empirical base of operations management. *Manufacturing & Service Operations Management, 9*(4), 368–382. doi:10.1287/msom.1070.0168.

Forster, A. J., Stiell, I., Wells, G., Lee, A. J., & Van Walraven, C. (2003). The effect of hospital occupancy on emergency department length of stay and patient disposition. *Academic Emergency Medicine, 10*(2), 127–133. doi:10.1197/aemj.10.2.127.

Gans, N., Liu, N., Mandelbaum, A., Shen, H., & Ye, H. (2010). *Service times in call centers: Agent heterogeneity and learning with some operational consequences* (pp. 99–123). Institute of Mathematical Statistics.

George, J. M., & Harrison, J. M. (2001). Dynamic control of a queue with adjustable service rate. *Operations Research, 49*(5), 720–731. doi:10.1287/opre.49.5.720.10605.

Gladstones, W. H., Regan, M. A., & Lee, R. B. (1989). Division of attention: The single-channel hypothesis revisited. *The Quarterly Journal of Experimental Psychology Section A, 41*(1), 1–17. doi:10.1080/14640748908402350.

Goes, P. B., Ilk, N., Lin, M., & Zhao, J. L. (2018). When more is less: Field evidence on unintended consequences of multitasking. *Management Science, 64*(7), 3033–3054. doi:10.1287/mnsc.2017.2763.

Gomersall, E. R. (1964). The backlog syndrome. *Harvard Business Review, 42*(5), 105–115.

Graves, S. C. (1986). A tactical planning model for a job shop. *Operations Research, 34*(4), 522–533. doi:10.1287/opre.34.4.522.

Gross, D., Shortle, J. F., Thompson, J. M., & Harris, C. M. (2008). *Fundamentals of queueing theory* (4th ed.). John Wiley & Sons. doi:10.1002/9781118625651.

Gupta, S., Verma, R., & Victorino, L. (2006). Empirical research published in production and operations management (1992–2005): Trends and future research directions. *Production and Operations Management, 15*(3), 432–448. doi:10.1111/j.1937-5956.2006.tb00256.x.

Harris, C. M. (1967). Queues with state-dependent stochastic service rates. *Operations Research, 15*(1), 117–130. doi:10.1287/opre.15.1.117.

Hasija, S., Pinker, E., & Shumsky, R. A. (2010). OM practice–work expands to fill the time available: Capacity estimation and staffing under Parkinson's law. *Manufacturing & Service Operations Management, 12*(1), 1–18. doi:10.1287/msom.1080.0250.

Hillier, D. F., Parry, G. J., Shannon, M. W., & Stack, A. M. (2009). The effect of hospital bed occupancy on throughput in the pediatric emergency department. *Annals of Emergency Medicine, 53*(6), 767–776. doi:10.1016/j.annemergmed.2008.11.024.

Hopp, W. J., Iravani, S. M., & Yuen, G. Y. (2007). Operations systems with discretionary task completion. *Management Science, 53*(1), 61–77. doi:10.1287/mnsc.1060.0598.

Ibanez, M. R., Clark, J. R., Huckman, R. S., & Staats, B. R. (2017). Discretionary task ordering: Queue management in radiological services. *Management Science, 64*(9), 4389–4407. doi:10.1287/mnsc.2017.2810.

Inman, R. R. (1999). Empirical evaluation of exponential and independence assumptions in queueing models of manufacturing systems. *Production and Operations Management, 8*(4), 409–432. doi:10.1111/j.1937-5956.1999.tb00316.x.

Jackson, J. R. (1963). Jobshop-like queueing systems. *Management Science, 10*(1), 131–142. doi:10.1287/mnsc.10.1.131.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology, 65*(4), 681. doi:10.1037/0022-3514.65.4.681.

KC, D. S. (2014). Does multitasking improve performance? Evidence from the emergency department. *Manufacturing & Service Operations Management, 16*(2), 168–183. doi:10.1287/msom.2013.0464.

Kc, D. S., & Terwiesch, C. (2009). Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science, 55*(9), 1486–1498. doi:10.1287/mnsc.1090.1037.

KC, D. S., & Terwiesch, C. (2012). An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management, 14*(1), 50–65. doi:10.1287/msom.1110.0341.

Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *The Annals of Mathematical Statistics, 24*(3), 338–354. doi:10.1214/aoms/1177728975.

Khudyakov, P., Gorfine, M., & Mandelbaum, A. (2018). Phase-type models of service times. In preparation.

Kim, S.-H., Chan, C. W., Olivares, M., & Escobar, G. (2014). ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Science, 61*(1), 19–38. doi:10.1287/mnsc.2014.2057.

Kingman, J. (1961). The single server queue in heavy traffic. *Mathematical Proceedings of the Cambridge Philosophical Society, 57*(4), 902–904. doi:10.1017/S0305004100036094.

Kingman, J. F. C. (2009). The first Erlang century—and the next. *Queueing Systems, 63*(1), 3–12. doi:10.1007/s11134-009-9147-4.

Kuntz, L., Mennicken, R., & Scholtes, S. (2014). Stress on the ward: Evidence of safety tipping points in hospitals. *Management Science, 61*(4), 754–771. doi:10.1287/mnsc.2014.1917.

Kuntz, L., Mennicken, R., Scholtes, S., et al. (2011). *Stress on the ward: An empirical study of the nonlinear relationship between organizational workload and service quality*. RUB, Department of Economics.

Latane, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology, 37*(6), 822. doi:10.1037/0022-3514.37.6.822.

Levy, Y., & Yechiali, U. (1975). Utilization of idle time in an $M/G/1$ queueing system. *Management Science, 22*(2), 202–211. doi:10.1287/mnsc.22.2.202.

Lindbeck, A., & Snower, D. J. (2000). Multitask learning and the reorganization of work: From Tayloristic to holistic organization. *Journal of Labor Economics, 18*(3), 353–376. doi:10.1086/209962.

Long, E. F., & Mathews, K. S. (2017). The boarding patient: Effects of ICU and hospital occupancy surges on patient flow. *Production and Operations Management*. doi:10.1111/poms.12808.

Louriz, M., Abidi, K., Akkaoui, M., Madani, N., Chater, K., Belayachi, J., et al. (2012). Determinants and outcomes associated with decisions to deny or to delay intensive care unit admission in Morocco. *Intensive Care Medicine, 38*(5), 830–837.

Lu, Y. (2013). *Data-driven system design in service operations*. Columbia University (Ph.D. thesis).

Lu, Y., Musalem, A., Olivares, M., & Schilkrut, A. (2013). Measuring the effect of queues on customer purchases. *Management Science, 59*(8), 1743–1763. doi:10.1287/mnsc.1120.1686.

Mæstad, O., Torsvik, G., & Aakvik, A. (2010). Overworked? On the relationship between workload and health worker performance. *Journal of Health Economics, 29*(5), 686–698. doi:10.1016/j.jhealeco.2010.05.006.

Mark, G., Gonzalez, V. M., & Harris, J. (2005). No task left behind? Examining the nature of fragmented work. In *Proceedings of the SIGCHI conference on human factors in computing systems*. In *CHI '05* (pp. 321–330). New York, NY, USA: ACM. doi:10.1145/1054972.1055017.

Mas, A., & Moretti, E. (2009). Peers at work. *The American Economic Review, 99*(1), 112–145. doi:10.1257/aer.99.1.112.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences, 7*(3), 134–140.

Neuts, M. F. (1981). *Matrix-geometric solutions in stochastic models: An algorithmic approach*. Dover Publications.

Oliva, R. (2001). Tradeoffs in responses to work pressure in the service industry. *California Management Review, 43*(4), 26–43.

Oliva, R., & Sterman, J. D. (2001). Cutting corners and working overtime: Quality erosion in the service industry. *Management Science, 47*(7), 894–914. doi:10.1287/mnsc.47.7.894.9807.

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin, 116*(2), 220. doi:10.1037/0033-2909.116.2.220.

Pisano, G. P., Bohmer, R. M., & Edmondson, A. C. (2001). Organizational differences in rates of learning: Evidence from the adoption of minimally invasive cardiac surgery. *Management Science, 47*(6), 752–768. doi:10.1287/mnsc.47.6.752.9811.

Robbins, T. R., Medeiros, D. J., & Harrison, T. P. (2010). Does the Erlang C model fit in real call centers? In *Proceedings of the winter simulation conference* (pp. 2853–2864). doi:10.1109/WSC.2010.5678980.

Rubinstein, J. S., Meyer, D. E., & Evans, J. E. (2001). Executive control of cognitive processes in task switching. *Journal of Experimental Psychology: Human Perception and Performance, 27*(4), 763. doi:10.1037/0096-1523.27.4.763.

Schultz, K. L., Juran, D. C., Boudreau, J. W., McClain, J. O., & Thomas, L. J. (1998). Modeling and worker motivation in JIT production systems. *Management Science, 44*(12-part-1), 1595–1607. doi:10.1287/mnsc.44.12.1595.

Schultz, K. L., McClain, J. O., & Thomas, L. J. (2003). Overcoming the dark side of worker flexibility. *Journal of Operations Management, 21*(1), 81–92. doi:10.1016/S0272-6963(02)00040-2.

Scudder, G. D., & Hill, C. A. (1998). A review and classification of empirical research in operations management. *Journal of Operations Management, 16*(1), 91–101.

Setyawati, L. (1995). Relation between feelings of fatigue, reaction time and work productivity. *Journal of Human Ergology, 24*(1), 129–135.

Shunko, M., Niederhoff, J., & Rosokha, Y. (2018). Humans are not machines: The behavioral impact of queueing design on service time. *Management Science, 64*(1), 453–473. doi:10.1287/mnsc.2016.2610.

Staats, B. R., & Gino, F. (2012). Specialization and variety in repetitive tasks: Evidence from a Japanese bank. *Management Science, 58*(6), 1141–1159. doi:10.1287/mnsc.1110.1482.

Steedman, I. (1970). Some improvement curve theory. *The International Journal of Production Research, 8*(3), 189–206. doi:10.1080/00207547008929840.

Stidham Jr. , S., & Weber, R. R. (1989). Monotonic and insensitive optimal policies for control of queues with undiscounted costs. *Operations Research, 37*(4), 611–625. doi:10.1287/opre.37.4.611.

Sze, D. Y. (1984). OR Practice–A queueing model for telephone operator staffing. *Operations Research, 32*(2), 229–249. doi:10.1287/opre.32.2.229.

Tan, T. F., & Netessine, S. (2014). When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Science, 60*(6), 1574–1593. doi:10.1287/mnsc.2014.1950.

Vroom, V. H. (2005). On the origins of expectancy theory. In *Great minds in management: The process of theory development* (pp. 239–258). New York: Oxford University Press.

Wang, J., & Zhou, Y.-P. (2018). Impact of queue configuration on service time: Evidence from a supermarket. *Management Science, 64*(7), 2973–3468. doi:10.1287/mnsc.2017.2781.

Welch, P. D. (1964). On a generalized $M/G/1$ queuing process in which the first customer of each busy period receives exceptional service. *Operations Research, 12*(5), 736–752. doi:10.1287/opre.12.5.736.