Doctors Under Load: An Empirical Study of State-Dependent Service Times

Robert J. Batt, Christian Terwiesch
The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, batt@wharton.upenn.edu, terwiesch@wharton.upenn.edu

We present an empirical study of a service system in which the servers have discretion over both selecting which tasks a customer requires and the duration of task completion. Using operational data from a hospital emergency department, we show that when crowded, multiple mechanisms act to retard patient treatment, but care providers adjust their clinical behavior to accelerate the service. We show that load-induced slowdown is present in many common tasks of the treatment process such as lab-specimen collection time and time-to-first-order. We identify two mechanisms that servers use to accelerate the system: early task initiation and task reduction. In contrast to other recent works, we find the net effect of these countervailing forces to be an increase in service time when the system is crowded. Further, we use simulation to show that ignoring state-dependent service times leads to modeling errors that could cause hospitals to overinvest in human and physical resources.

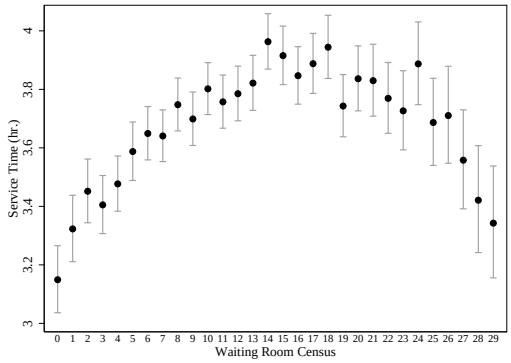
Key words: Healthcare operations; empirical; emergency department; dynamic queue control
History: Working Paper: November, 2012

1. Introduction

The Operations Management community has long been concerned with how crowding affects the performance of queuing systems. Basic queuing theory shows that crowding and high utilization of queues lead to exponentially increasing wait times. Since long waits are generally undesirable, it seems reasonable that, when possible, workers in human-paced service systems would attempt to accelerate the system, a phenomenon we call *Speedup*. Indeed, this has been shown to be true both in the lab and in practice (Schultz et al. 1998, Kc and Terwiesch 2009, Chan et al. 2011). These papers show that workers in settings as varied as data-entry and hospital intensive care units accelerate service under high load conditions.

In contrast, in domains such as transportation and telecommunications, high load conditions are well known to lead to service time increases or *Slowdown* (Chen et al. 2001, Gerla and Kleinrock 1980). A hallmark of Slowdown-prone systems is that service involves shared resources and/or servers that are not independent. For example, a highway lane is a shared resource for all the cars traveling in it and its performance can also be impacted by the traffic in adjacent lanes. Likewise, each node

Figure 1 Service Time as a Function of Census



Notes: Mean and 95% confidence interval of mean shown. ED patients between 3pm and 11pm (second shift). Census is measured at the time a patient enters a treatment room.

in a telecom network is a shared resource for many users, and it can be impacted by spillover from other nearby nodes (Gerla and Kleinrock 1980).

We bring these two viewpoints together by empirically analyzing a service system where both Speedup and Slowdown effects are present: a hospital emergency department (ED). The ED provides an excellent study environment for several reasons. First, the service (medical care) is provided by humans and as such is worker paced. Further, the required work for each patient is largely determined by the server (nurse or doctor) and the patient has limited knowledge of his or her own needs. This creates an environment in which the servers have a great deal of discretion over the encounter. This freedom can be used to alter both the service content (the specific tasks performed for the patient) and the service time (the total time to complete all tasks). Lastly, the ED is interesting because it is a complex service environment with many shared resources (nurses, doctors, equipment, hallways, laboratory, etc.). This suggests that the ED is prone to Slowdown.

Figure 1 previews our data, and motivates our study of Speedup and Slowdown mechanisms. The figure plots the mean service time of ED patients that arrive during second shift (3pm to 11pm) as a function of the waiting room census. Here, and throughout the paper, we define service time to be the time from when a patient is placed in a treatment bed to when treatment in the ED is complete as indicated by the patient either departing to go home or an inpatient bed request is placed in preparation for admission to the hospital. Thus service time does not include any time spent in the waiting room. The figure shows that mean service time rises from about 3.2 hours to 3.9 hours and then falls to 3.3 hours as the waiting room census ranges from low to high. If Speedup and Slowdown effects are monotone in census level, then the non-monotone form of Figure 1 suggests that both Speedup and Slowdown are at work in the ED.

Prior empirical work on state-dependent service times has largely focused on the presence of state-dependent service times but not the mechanisms generating the state dependencies. In this

paper, we identify and test for several state-dependent mechanisms including task reduction, early task initiation, multitasking, and interference. The first two are Speedup mechanisms and the latter two are Slowdown mechanisms.

Our study hospital has the additional feature of an "express lane" or FastTrack (FT) for low-acuity patients that is open only certain hours of the week. The FT is partially isolated from the rest of the ED operations; it uses dedicated treatment rooms and care providers. However, it relies on the same auxiliary services, such as the pathology lab and x-ray machines, as the main ED. We compare the effects of crowding on the ED and the FT.

We conduct a detailed econometric analysis of the service times and service content during more than 100,000 emergency department visits at a major U.S. hospital. We observe patient-level characteristics (age, gender, race, etc.) as well as timestamps of the progress of each visit including patient location and all laboratory, radiology, and medication orders. Survival analysis techniques are used to estimate the effects of Slowdown on service time and several common tasks. Count model regression techniques are used to identify various forms of service Speedup. Lastly, we use discrete event simulation to determine if these state-dependencies have a meaningful impact on the system. This research design allows us to make the following four contributions:

1. We examine several common ED tasks and find evidence of Slowdown in all. For example, time to first order (a measure of doctor speed) and medication delivery time time (a measure of nurse speed) increase by 26% and 11% respectively under high load.

2. We test for two Speedup mechanisms: early task initiation and task reduction. We find strong evidence of early task initiation with the expected number of triage tests increasing from 0.3 to 0.9 in the ED. We find only limited use of task reduction in the ED, while task reduction is more common in the FT.

3. We show that the net effect of Speedup and Slowdown is different in the ED and the FT. In the ED, service time first increases then decreases with load as the relative strength of Speedup and Slowdown mechanisms shifts. In the FT, Speedup and Slowdown balance out leading to little change in service time with increased crowding.

4. We show that models which ignore the state-dependent service times overestimate the system utilization and congestion.

These findings offer several operational insights for managers. For example, we show that implementing early task initiation by increasing the number of tests ordered at triage is an effective way to reduce service time. This suggests that care providers should consider incorporating state-dependencies into ED care protocols. For both the healthcare domain and other domains, our findings show that understanding the micro-level mechanisms behind state-dependent service rates is important for properly modeling service systems where the server has discretion over the service

speed and the service content. Our results, particularly regarding task reduction and task time increases, suggest an operational explanation for the many studies that have shown a link between crowding and reduced clinical quality in the ED (e.g., Fee et al. 2007, Pines and Hollander 2008). However, in this paper we remain focused on the effect of crowding on service time.

2. Clinical Setting

Our study is based on data from a large, urban, teaching hospital with an average of 4,700 ED visits per month over the study period of January, 2009 through December, 2011. The ED has 25 treatment rooms and 15 hallway beds for a theoretical maximum treatment capacity of 40 beds. However, the actual treatment capacity at any given moment can fluctuate for various reasons. The hospital also operates an express lane or FastTrack (FT) for low acuity patients. The FT is generally open from 8am to 8pm on weekdays, and from 9am to 6pm on weekends. The FT operates somewhat autonomously from the rest of the ED in that it utilizes seven dedicated beds and is usually staffed by dedicated group of Certified Registered Nurse Practitioners (CRNP) rather than Medical Doctors (MD)[1] .

In our analysis, we focus solely on patients that are classified as "walk-ins" or "self" arrivals, as opposed to ambulance, police, or helicopter arrivals. This is because the walk-ins go through a more standardized process of triage, waiting, and treatment, as described below. In contrast, ambulance arrivals tend to jump the queue for bed placement, regardless of severity, and often do not go through the triage process or wait in the waiting room. More than 70% of ED arrivals are walk-ins. Note, however, that the non-walkin patients are included in the relevant census measures.

The study hospital operates in a manner similar to many hospitals across the United States. Upon arrival, patients are checked in and an electronic patient record is initiated for that visit. Only basic information (name, age, complaint) is collected at check-in. Shortly thereafter, the patient is seen by a triage nurse who assesses the patient, measures vital signs, and records the official chief complaint. The triage nurse also assigns a triage level which indicates acuity. The hospital uses a five-level Emergency Severity Index triage scale with 1 being most severe and 5 being least severe. The triage nurse also has the option of ordering pathology lab tests (e.g., urinalysis, blood test) and certain types of radiology imaging scans (e.g., x-rays).

After triage, all patients wait in a common waiting room to be taken to a treatment room. Patients are called for service when a treatment bed is available. If only the ED is open, patients are generally (but not strictly) called for service in first-come-first-served (FCFS) order by triage level. If the FT

[1] We interchangeably use the term ED to refer to the entire Emergency Department inclusive of the FastTrack or to just the main emergency department treatment area exclusive of the FastTrack. The use is generally clear from the context, but we use the term "main ED" to clarify and indicate the primary ED treatment space when necessary.

is open, then the FT will serve triage level 4 and 5 patients in FCFS order by triage level and the ED will serve patients of triage levels 1 through 3 in FCFS order by triage level. These routing procedures are flexible, however. For example, the ED might serve a triage level 4 patient if the patient has been waiting a long time and there are not more acute patients that need immediate attention. Similarly, the FT might serve a triage level 3 patient if the patient has been waiting a long time and the patient's needs can be met by the nurse practitioners in the FT. The mean and median wait times for ED patients are are 1.6 hours and 0.84 hours, respectively. The mean and median wait times for FT patients are 1.1 hours and 0.9 hours, respectively.

Patients served by the main ED are eventually assigned to a treatment room by the charge nurse.[2] This marks the beginning of the service time. Soon after being moved to a treatment room, a physician meets with and examines the patient.[3] At this point, the physician generates a mental list of possible diagnoses, called a differential diagnosis, and decides the trajectory of the diagnosis and treatment process. Frequently, orders for diagnostic tests, medications, or both are made at this point. All lab test, radiology scan, and medication orders are recorded electronically in the patient tracking system, but orders are frequently conveyed orally to the nurses as well.

Lab specimens are drawn by the nurse and most are sent to the hospital's central pathology lab by pneumatic tube for processing. A small subset of pathology tests are performed locally in the ED by the nurse. Similarly, the nurse is responsible for delivering medications to the patient. When the nurse finishes either of these tasks, the order is closed out and timestamped in the electronic patient record. Orders for radiology scans trigger a patient transport request. Transporters work in a first-come-first-served manner through the request queue to transport patients to the appropriate scanner and then back to the treatment room.

Eventually, the physician decides that either the patient can leave or the patient needs to be admitted. If the patient is to be admitted, a bed request is entered in the inpatient bed management system. At this point, ED service is considered complete. The patient waits for an available inpatient bed and is considered a "boarder" in the ED. This boarding period can be quite long with a mean of 3.6 hours. During this time, the patient continues to occupy a treatment room and requires some attention from the nursing staff, but the physician is effectively done with the patient. The number of boarding patients in the ED ranges from zero to 20 with a mean of six. For patients that are discharged, service time ends when the patient leaves the ED. Mean service time for admitted and discharged patients is 3.6 hours and 3.8 hours respectively.

[2] The treatment location is sometimes a hallway bed rather than a room, but we use the word "room" for ease of exposition.

[3] Because the study hospital is a teaching hospital, a medical student or a resident physician may also be involved in the care of the patient.

For patients served by the FT, the care process is quite similar to that in the ED, except with a dedicated group of rooms and providers. Once in a treatment room, the care provider evaluates the patient, orders any necessary tests and medicines, and attempts to provide treatment as rapidly as possible. Just as in the ED, all lab test, radiology scan, and medication orders are recorded electronically in the patient tracking system. One difference between the FT and the ED is that there is a less clear demarcation between provider and nurse tasks. For example, a CRNP treating a FT patient may order and deliver medications him or herself, whereas in the ED, the doctor would order the medicine and the nurse would deliver it. However, as in the ED, FT labs are generally drawn by a nurse and scan orders enter the same transport queue as the ED patients. When treatment is complete, the patient is discharged. In rare cases, the FT provider can reroute the patient to the ED or admit the patient to the hospital. Mean service time for FT patients is 1.3 hours.

3. Framework & Hypotheses

We are interested in examining the mechanisms of state-dependent service times at the server level. We begin with the assumption from classical queuing theory that the service time distribution is not affected by the system state (Wolff 1989). However, as seen in Figure 1, it appears that this assumption is false in our setting, and that there is a dependence between the system state and the service time. Similarly, Armony et al. (2012) includes an empirical examination of an ED at the system level and finds evidence of both Speedup and Slowdown. However, in contrast to what we show in Figure 1, Armony et al. (2012) finds that the ED first speeds up and then slows down as load increases from low to high. Armony et al. (2012) muses (but does not test) that Speedup may be the result of rushing as care providers respond to a mild increase in congestion, and that Slowdown could also be caused by factors such as fatigue, shared resources being spread thin, or nurses having to devote too much time to caring for boarding patients.

We posit that there are several mechanisms that may be at work and that these can be classified by the direction of their impact on service times and by the number of resources involved. In the following we describe these mechanisms, their related prior research, and the hypotheses they motivate.

3.1. Slowdown

We focus first on Slowdown, or mechanisms that increase service time. Prior literature has shown that both fatigue and multitasking can lead to Slowdown in individual servers. For example, several studies in medical and ergonomics journals have shown that fatigue leads to diminished productivity (e.g., Setyawati 1995, Caldwell 2001). Similarly, Kc and Terwiesch (2009) finds that fatigue caused by extended periods of high workload leads to decreased productivity in both hospital transportation and cardiac ICU care.

In our setting, multitasking refers to a single resource, such as a nurse, being simultaneously responsible for multiple patients, but individual tasks are not necessarily performed simultaneously. For example, a nurse may deliver a medication to one patient and then draw blood from a second patient. In effect, the nurse acts as a single channel server performing tasks for different patients in rapid succession. As the nurse becomes responsible for more patients and gets "spread thin," the arrival rate of tasks to the nurse's virtual queue increases leading to longer completion times for each individual task from the patient's point of view. The Psychology literature on human multitasking shows that multitasking additionally incurs cognitive switching costs which further hinder productivity (Pashler 1994). These switching costs increase with increased levels of multitasking. See KC (2011) for a summary of this literature. KC (2011) empirically examines the effect of ED physician multitasking on service time and finds that multitasking leads to longer service times. A shared resource, like an x-ray machine, can be thought of as multitasking in a similar manner. With more patients in treatment, more x-ray requests are generated, the queue for x-rays grows, and the completion time for each x-ray increases.

Another form of Slowdown can occur with multiple resources. As mentioned in Section 1, the idea of high load causing Slowdown is well established in fields such as transportation and telecommunications (Chen et al. 2001, Gerla and Kleinrock 1980). In these settings, this effect is commonly referred to as congestion. However, we refer to this as *interference* since this is a different effect than is generally referred to in the Operations Management literature by the word "congestion." In the Operations Management literature, congestion usually refers to long queues and long wait and sojourn times, but does not imply any change in service times. In the transportation and telecommunications settings, and in this paper, the Slowdown effect of interest is an increase in the actual service time, regardless of wait time. In the ED, examples of interference are crowded hallways that slow workers down and nurses waiting for computer terminals.

Both multitasking and interference are conceptually similar to queuing models with shared processors (e.g., Yamazaki and Sakasegawa 1987, Aksin and Harker 2001). Shared processor models assume that the server (or servers) splits its processing capacity across all items in service leading to service times increasing as the number of customers in service increases. For example, Aksin and Harker (2001) models a multi-server call center with multiple customer classes and a single shared information management system that slows down as it performs more simultaneous operations. The key finding is that the system throughput decay caused by processor sharing is a function of both the offered load on the system and the proportion of a customer's service that requires use of the shared resource. This is relevant for our ED setting since many resources in the ED are shared resources (e.g., nurses, doctors, equipment) and EDs regularly operate under high offered loads.

To test for Slowdown, it is not sufficient to simply examine total service time for a patient because the service time is affected by both Speedup and Slowdown effects. To isolate and test for the existence of Slowdown, we focus on the durations of a few specific tasks that are common to many ED visits such as lab specimen collection time and x-ray completion time. We suspect that such tasks are susceptible to all the Slowdown mechanisms described above. For example, lab collection time will increase as a nurse juggles more patients, becomes fatigued, and has to wait in line to use the pneumatic tube system to send a sample to the lab. Thus, while we do not attempt to separately identify the Slowdown mechanisms at work, we test for the presence of Slowdown in general, and we expect crowding to lead to increased task times.

Hypothesis 1 Task time increases with load: $\frac{\partial TaskTime}{\partial Load} > 0$

3.2. Speedup

Turning now to Speedup, or mechanisms that decrease service times, the subset of queuing theory focused on optimal control of queues provides theoretical motivation for Speedup behavior. Dynamic control queues dynamically adjust to system state parameters such as the queue length. Going back to Crabill (1972), several papers have explored optimal control policies that minimize average cost per unit time by adjusting the service time, and have proven under increasingly weaker assumptions the existence of an optimal service time policy that is monotone decreasing in queue length (e.g., Stidham and Weber 1989, George and Harrison 2001). The intuition behind such a policy is based on the assumptions that the system waiting cost per unit time increases with queue length and that there is a cost to decreased service time, either in terms of labor, effort, or reduced quality. Thus, as the queue length grows, the waiting costs eventually outweigh the cost of faster service and the optimal response is to reduce the service time.

Perhaps the simplest form of service time reduction is *rushing*. That is, the server simply works faster. Schultz et al. (1998) finds this sort of acceleration behavior in a lab experiment, and Kc and Terwiesch (2009) is the first paper to show this behavior in the field. It finds that hospital transporters work faster when the workload is high. Similarly, Tan and Netessine (2012) and Staats and Gino (2012) find evidence of rushing Speedup under load with restaurant waiters and loan application processors, respectively.

Since rushing affects task time, we are actually testing the net effect of Slowdown and rushing when we test for the effect of load on task time in Hypothesis 1. We have stated Hypothesis 1 as we have ($\frac{\partial TaskTime}{\partial Load} > 0$) because we believe that Slowdown dominates rushing in the ED. In fact, we believe that rushing is not prevalent in many knowledge-intensive services such as the ED. Despite what is portrayed on TV, doctors and nurses are rarely seen running through the halls of the ED or performing specific procedures faster.

3.2.1. Task Reduction Papers by Hopp et al. (2007) and by Alizamir et al. (2011) build on the optimal queue control stream and suggest another Speedup mechanism; *task reduction*. Hopp et al. (2007) describes a service system with discretionary task completion that is concave-increasing in value with time. A holding cost is incurred per unit time for each customer in the system. This leads to an optimal policy that sets a service cutoff time for every value of queue length. This policy is monotone decreasing in queue length. Alizamir et al. (2011) models a diagnostic service as a stochastic sequence of diagnostic tests. Each test informs the server's probability estimation of the customer's type. This specification can lead to an optimal policy that sets a maximum number of tests for each queue length. This maximum is decreasing in queue length. The common element of these papers is that it is a change in the service content, not the service rate (i.e. task completions per time interval), which leads to a change in the service time per customer. Oliva and Sterman (2001), Kc and Terwiesch (2009), and Chan et al. (2011) are all suggestive of this sort of task reduction based Speedup.

The discretionary task completion model of Hopp et al. (2007) forms the basis of our hypotheses regarding task reduction. In the Hopp et al. (2007) framework, the variable under the server's control is service time itself. In our setting, we assume the variable under the physician's control is the service content, that is the quantity of diagnostic tests ordered. Further, we assume that utility is concave increasing with the number of tests. As long as reducing testing quantity reduces service time, the insight from Hopp et al. (2007) that service time should be reduced under crowding translates to the hypothesis that testing should be reduced under crowding. This leads to the following two hypotheses.

Hypothesis 2 Service time increases with diagnostic testing: $\frac{\partial ServiceTime}{\partial Tests} > 0$

Hypothesis 3 Diagnostic testing decreases with load: $\frac{\partial Tests}{\partial Load} < 0$

The idea that service time should be reduced under crowding seems quite reasonable, perhaps even obvious, in the settings proposed in Hopp et al. (2007) such as telemarketers and salespeople. However, in a medical setting such as an ED, the idea of reducing the quantity -and perhaps quality- of care for Mrs. Jones just because she has the bad luck of being in the ED when there is a crowd seems less obvious. We leave that discussion for later and simply draw on the Hopp et al. (2007) model to suggest an interesting hypothesis, that physicians change the thoroughness of their testing based on crowding. We refer to this behavior with the admittedly loaded term "cutting corners."

3.2.2. Early Task Initiation While rushing and task reduction are Speedup mechanisms that can be implemented by a single server, we propose the mechanism of *early task initiation* as a Speedup mechanism that may exist between resources. Early task initiation is similar to concurrent engineering, which for nearly thirty years has been acknowledged as an effective way to speed up

product development cycles. First widely publicized by Imai et al. (1985) and Takeuchi and Nonaka (1986), the concept is to take logically consecutive tasks and execute them with some amount of temporal overlap. This requires the decision makers at each task to make some guesses or bets since the exact needs of the other tasks are not yet known. The fundamental tradeoff is that overlapping the tasks reduces the time to market but that too much overlap leads to rework or poor final design quality (Loch and Terwiesch 1998).

A similar opportunity exists in multi-resource service systems. A service task may be started early, before it is even fully known if the task is required. For example, in the ED, as described in Section 2, triage nurses have the option of ordering some diagnostic tests.[4] If tests are ordered at triage, the tests can be processed while the patient is waiting in the waiting room. Then when the patient sees the physician the tests are already under way or may even be ready for review. This reduces service time. However, the downside of triage testing is that the nurse is "placing bets," in that the nurse may not be certain what tests the doctor will want and may order unneeded tests. This could be due to the nurse having less training and skill than the doctor, or due to the limited information available from a triage examination. This over-testing is undesirable because it increases financial costs, medical risk for the patient (if the test is risky), and load on the diagnostic resources.

Note that the benefits of ordering tests at triage are largest when waiting times are long. This is because much or all of the test processing time occurs in parallel with the patient waiting in the waiting room. Conversely, when waiting times are short, there is little benefit to triage testing since the service time will be reduced by only a few minutes. However, the consequences of over-testing do not scale with load in a similar fashion, and therefore we hypothesize that triage testing will be most common when the system is crowded.

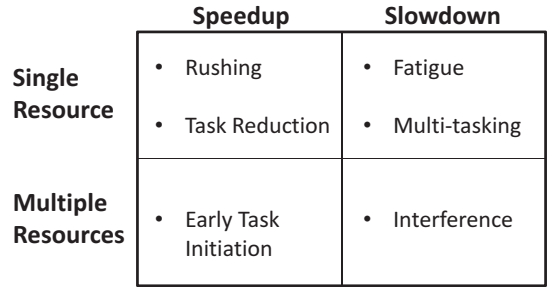
Hypothesis 4 Triage testing increases with load: $\frac{\partial TriageTest}{\partial Load} > 0$

For early task initiation to be beneficial, an increase in triage testing should lead to a decrease in doctor testing. If triage nurses have perfect information we would expect a one for one trade-off between triage and doctor testing; each incremental triage test would lead to a one test reduction in doctor testing. However, if the nurses have imperfect information and "betting" is an apt description, then we would expect the marginal triage test to lead to a reduction in doctor testing of less than one.

Hypothesis 5 Doctor testing decreases less than one unit for each unit increase in triage testing: $-1 < \frac{\partial DocTest}{\partial TriageTest} < 0$

[4] These triage tests are commonly referred to as Advanced Triage Protocols in the medical community.

Figure 2 **State-Dependent Mechanisms**



3.3. Net Impact on Service Time

Figure 2 summarizes the categorization of the mechanisms just described that potentially lead to state-dependent service times. Since Speedup and Slowdown mechanisms work in opposing directions, the net impact is indeterminate a priori. Therefore, we do not posit an hypothesis. Nonetheless, it is worth examining the net change in service time with load to determine the relative magnitudes of the two effects. Based on Figure 1, we suspect that Slowdown dominates but that Speedup effects eventually become large enough such that the marginal effect of load is negative. Stated differently, we believe that for low to mid level loads $\frac{\partial ServiceTime}{\partial Load} > 0$, and for mid to high level loads $\frac{\partial ServiceTime}{\partial Load} < 0$.

3.4. Additional Related Literature

While we have already referenced the prior work to which our study is most closely related, we also point out connections to two other bodies of literature.

Our work is influenced by the portion of the analytical queuing theory literature has been stimulated by problems in the health care domain. Topics such as capacity planning (e.g., Lee and Zenios 2009, Allon et al. 2011), staffing (e.g., deVericourt and Jennings 2011, Yankovic and Green 2012) and patient flow (e.g., Green et al. 2006, Ibrahim and Whitt 2011) have all been studied extensively. We direct the reader to Green (2006) for an overview of this literature. This body of work has largely been focused on characterizing and managing service systems from a high-level or system design point of view.

Our work also relates to the large body of medical literature on crowding's effect on service and quality. Many of these papers have shown the negative impacts of ED crowding on such measures as timing of antibiotic delivery for pneumonia patients, pain medication for patients with severe pain, and nebulizer treatment for patients with asthma (Pines et al. 2006, Fee et al. 2007, Pines and Hollander 2008, Pines et al. 2010). Crowding has also been associated with reduced patient satisfaction (Pines et al. 2008). Results on the impact of crowding on length of stay have been mixed. For example, Pines et al. (2010) report a positive relationship between crowding and length

of stay while Lucas et al. (2009) find no significant relationship. McCarthy et al. (2009) report that crowding drives up wait times but has no effect on service times, a result that agrees with traditional queuing theory.

Our contribution to the literature is in bringing attention to the level of the servers (care providers). We expand on the prior literature by providing detailed evidence of both Speedup and Slowdown mechanisms occurring simultaneously. By focusing at the micro-level, we can identify the underlying mechanisms that lead to the service time changing under load. We hope this will extend the understanding of service system productivity.

4. Data Description & Definitions

Our data include information for each patient visit such as patient demographics, chief complaint, attending physician, and timestamps of all major events and physician orders. Table 1 provides descriptive statistics of the patient population. For much of the analysis, we focus on a single chief

Table 1 Summary Statistics of Patients

Variable	ED Mean	FT Mean
Age	41.2 (0.05)	34.6 (0.08)
Female	61% (0.002)	59% (0.003)
Triage 2	25.1% (0.001)	1.3% (0.001)
Triage 3	59.2% (0.001)	5.3% (0.001)
Race: Black	58.6% (0.002)	64.3% (0.001)
Race: White	24.8% (0.001)	19.8% (0.002)
Diagnostics Ordered	5.38 (0.014)	1.27 (0.010)
Service Time (hr.)	3.77 (0.009)	1.31 (0.006)
N	108,014	36,427

Standard error in parentheses

complaint at a time since the testing patterns and response to crowding can be quite different from one chief complaint to another. Chief complaint is determined by the triage nurse, and our data contains 129 unique chief complaints. The two most common chief complaints in the ED are abdominal pain and chest pain, representing 13% and 9% of the ED visits respectively. The two most common chief complaints in the FT are limb pain and body pain, representing 14% and 9% respectively.

We are primarily concerned with how load affects ED performance. In the ED, there are several census measures that indicate system load. These include waiting room census, ED in-service census, FT in-service census, and ED boarding census. To calculate these census measures, we divide the study period (2009-2011) into 15-minute intervals labeled t, and we use the patient visit timestamps to generate the census variables $WAIT_t$, $EDSERV_t$, $FTSERV_t$, and $BOARD_t$ as the number of patients in the given location during interval t.

When we examine task times (Hypothesis 1), we perform the analysis at the per-hour level and thus we generate the load variables $\overline{WAIT_h}$, $\overline{EDSERV_h}$, $\overline{FTSERV_h}$, and $\overline{BOARD_h}$ as the average for hour h for each of the census measures

For the rest of our analysis, we focus solely on the waiting room census as the measure of ED load. We do this because observation and anecdotal evidence suggests that ED nurses and doctors focus on this number as a key indicator of the crowd level in the ED. Further, the waiting room census is visible to the triage nurses and the rest of the ED staff on electronic dashboards. We also choose to focus on waiting room census because it effectively has no upper bound and thus has a great deal of variability. In contrast, in-service and boarding census measures are limited by the number of beds in the ED. Lastly, we focus on waiting room census because we believe that the effects of crowding in the ED primarily occur when the ED is operating in a highly-loaded or overloaded state with all treatment beds filled.

We assign two load measures to each patient visit: load at arrival, $aLOAD_i$, and load at the start of service, $sLOAD_i$. For example, for patient i who arrives at time interval $t=1$ and is put in a treatment room at time $t=8$, $aLOAD_i = WAIT_1$, and $sLOAD_i = WAIT_8$. We then convert the variables $aLOAD_i$ and $sLOAD_i$ into vectors of dummy variables $\widetilde{\mathbf{aLOAD}}_i$ and $\widetilde{\mathbf{sLOAD}}_i$ corresponding to low, mid, and high census levels. The cut points are set such that 25% of observations are in each of the low and high categories and 50% of the observations are in the mid category. For $\widetilde{\mathbf{aLOAD}}_i$, the cut points are at 5 and 19, while for $\widetilde{\mathbf{sLOAD}}_i$ the cutpoints are at 4 and 18.

One reason for using a categorical load variable is that it allows for a more general response to load than would including just linear and quadratic terms of $LOAD_i$. The other reason is that it greatly simplifies the reporting of results and comparison of various models as will be seen in Section 6.

We examine several dependent variables in this study including task time, service time, and the counts of various categories of diagnostic tests.

To study task timing, we define the variable $\overline{TASKTIME_h}$ as the mean task completion time across all tasks of a given type ordered during hour h. The tasks we examine are as follows:

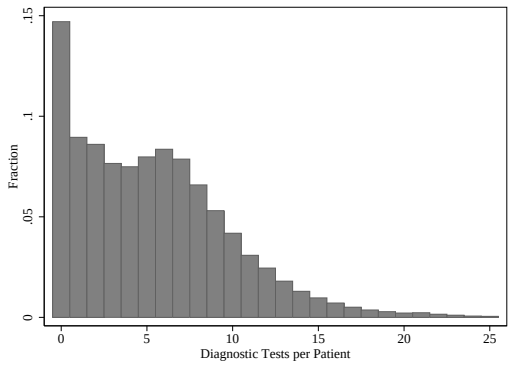
First Order Time: The time from when a patient is put in a treatment room until the first order (lab, scan, or medication) is recorded.

Lab Collection Time: The time from a lab order being placed until the nurse closes out the order indicating that the specimen has been sent for analysis.

Medication Delivery Time: The time from a medication order being placed until the nurse closes out the order indicating the medication has been given to the patient.

Scan Completion Time: The time from a radiology scan order being placed until the patient returns from having the scan performed. This does not include the time required for a radiologist to perform the official "reading" of the scan.

Figure 3 Number of Diagnostic Tests per ED Patient



The first task is a proxy for the physician busyness level. The second and third tasks are proxies for nurse busyness. The fourth task measures the sojourn time for an auxiliary service that is shared by the entire ED and by other parts of the hospital, depending on the scan type.

The service time variable, $SERVTIME_i$, is defined as the time from placement in a treatment room until the patient is either discharged or a bed request is placed for admission to the hospital for patient i. Note that service time does not include any time spent in the waiting room.

The last major dependent variable is the count of diagnostic tests ordered either by the triage nurse or doctor. There are two types of diagnostic tests: lab tests and radiology imaging scans. Lab tests are chemical analyses of patient tissue or fluid such as urinalysis, white blood cell counts, and electrolyte levels. Most of these tests are performed by the hospital's central pathology lab that serves both the ED and the rest of the hospital. Radiology imaging scans include various types of electromagnetic and ultrasonic imaging techniques, such as x-ray, magnetic resonance imaging, and computed tomography, used to view the internal structures of the body. For most of our analyses we aggregate these two types of tests into a single variable $TEST_i$ (Figure 3). We also decompose diagnostic test orders into $TRITEST_i$ and $DOCTEST_i$ based on whether the test was ordered at triage or in the treatment room. The average ED patient receives 0.6 triage tests and 4.8 doctor tests, however 15% receive no diagnostic tests at all. The mean number of diagnostic tests varies significantly by chief complaint and triage level. For some models, we further decompose $TRITEST_i$ and $DOCTEST_i$ into the number of labs and scans ordered at each location.

$$TRITEST_i = TRILAB_i + TRISCAN_i \tag{1}$$

$$DOCTEST_i = DOCLAB_i + DOCSCAN_i \tag{2}$$

5. Econometric Specification

We now develop the econometric specifications for testing our hypotheses. In the discussion below, the index h indicates an hour in the study period, and the index i denotes a patient visit to the emergency department.

To test Hypothesis 1, we are interested in how load impacts the duration of various common ED tasks, thus we turn to survival analysis models. Specifically, we use an accelerated-failure-time (AFT) model with a log-normal distribution. The AFT model relates the log of service time to a vector of covariates and a random error term ϵ through a linear equation. For this analysis, we relate the mean task time in a given hour to a load variable and control variables as follows:

$$ln(\overline{TASKTIME_h}) = \alpha + \beta_1 \overline{WAIT_h} + \beta_2 \overline{EDSERV_h} + \beta_3 \overline{FTSERV_h} + \beta_4 \overline{BOARD_h} + \mathbf{Z}_i\boldsymbol{\phi} + \epsilon_h \quad (3)$$

$\boldsymbol{Z}_i$ is a vector of time related control variables including year, month, day of week, hour of day, and the interaction of day of week and hour of day. Because our dependent variables are estimated means, we use weighted least squares to estimate the model where the weights are equal to the number of tasks ordered in hour h (Wooldridge 2009). Also, because the data forms a time series with possible autocorrelation we use the Newey-West covariance estimator to provide standard errors that are robust to both heteroskedasticity and autocorrelation (Greene 2012). Due to these complications, we must assume that ϵ_h follows a normal distribution. Thus, Equation 3 is an AFT model with a log-normal underlying distribution. In this specification, positive coefficients β or $\boldsymbol{\phi}$ indicate an increase in mean task time, and Hypothesis 1 is supported if $\beta > 0$.

We note that the AFT model implies specific assumptions about the underlying survival and hazard functions. Specifically the log-normal specification implies a hazard function that is first increasing and then decreasing. We choose this distribution because this form resembles the hazard function form of the data and because it allows us to correct for the weighting and autocorrelation as mentioned above. The major advantage of the AFT model over the semi-parametric Cox proportional hazard model is that the AFT model coefficients can be directly interpreted as changes in duration and a prediction of mean task time can be calculated.

Hypothesis 2 examines the effect of testing on service time. We achieve this by using the following AFT model specification which includes variables for both labs and scans ordered at triage and by the doctor.

$$\begin{aligned} ln(SERVTIME_i) = &\alpha + \widetilde{\mathbf{aLOAD}}_i\boldsymbol{\beta} + \delta_1 TRILAB_i + \delta_2 DOCLAB_i \\ &+ \delta_3 TRISCAN_i + \delta_4 DOCSCAN_i + \mathbf{W}_i\boldsymbol{\theta} + \mathbf{Z}_i\boldsymbol{\phi} + \epsilon_i \end{aligned} \quad (4)$$

The dependent variable is now service time for patient i. $\mathbf{W}_i$ is a vector of patient-visit specific covariates such age, gender, race, triage level, and chief complaint. $\boldsymbol{Z}_i$ is again a vector of time related control variables including year, month, hour of day and a weekend indicator variable. $\widetilde{\mathbf{aLOAD}}_i$ is a vector of dummy variables indicating mid and high load with the low load condition as the omitted category. We now assume ϵ follows a log-logistic distribution rather than a log-normal distribution. While the log-logistic and log-normal distributions assume similarly shaped hazard functions, we use the log-logistic function here because it better fits the data based on the Bayesian Information Criterion. Positive values of the δ coefficients support the hypothesis that testing leads to longer service times.

Hypotheses 3, 4, and 5 all require examining how test order quantities change with respect to some load or testing variable. Since the dependent variable is discrete and fairly small, we need to use a count-type model. Further, as seen in Figure 3, the excess of zero counts suggests the need for a zero-inflated model. We use a zero-inflated negative binomial (ZINB) model for all of these studies. The ZINB model combines a binary logit process with probability density $f_1(\cdot)$ and a negative binomial count process with probability density $f_2(\cdot)$ to create the combined density

$$f(y|\mathbf{x}) = \begin{cases} f_1(1|\mathbf{x}_1) + \{1 - f_1(1|\mathbf{x}_1)\} f_2(0|\mathbf{x}_2) & \text{if } y = 0 \\ \{1 - f_1(1|\mathbf{x}_1)\} f_2(y|\mathbf{x}_2) & \text{if } y \geq 1 \end{cases} \tag{5}$$

Note that this formulation is somewhat counterintuitive (albeit standard practice) in that a "success" of the binary process corresponds to $y = 0$, whereas a "failure" corresponds to y being determined by the negative binomial count process. This model has the conditional mean

$$E[y|\mathbf{x}] = \frac{1}{1 + \exp(\mathbf{x_1}\boldsymbol{\eta_1})} \times \exp(\mathbf{x_2}\boldsymbol{\eta_2}) \tag{6}$$

The covariate vectors $\mathbf{x_1}$ and $\mathbf{x_2}$ need not be the same, but for our purposes they are the same unless noted otherwise on the result table. The parameter vectors $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are estimated jointly by maximum likelihood using the log-likelihood function shown in the appendix. For $\boldsymbol{\eta}_1$, a positive coefficient indicates a decrease in the expectation of the dependent variable with an increase in the given independent variable, while the opposite is true for $\boldsymbol{\eta}_2$.

To test for the presence of task reduction (Hypothesis 3) we examine how $DOCTEST_i$ changes with load controlling for $TRITEST_i$. We formulate the linear predictors $\mathbf{x}_{i,1}\boldsymbol{\eta}_1$ and $\mathbf{x}_{i,2}\boldsymbol{\eta}_2$ as follows:

$$\mathbf{x}_{i,j}\boldsymbol{\eta}_j = \alpha_j + \widetilde{\mathbf{sLOAD}}_i\boldsymbol{\beta}_j + \delta_j TRITEST_i + \mathbf{W}_{i,j}\boldsymbol{\theta}_j + \mathbf{Z}_{i,j}\boldsymbol{\phi}_j \text{ for } j = 1, 2 \tag{7}$$

Similar to Equation 4, $\mathbf{W}_{i,j}$ is a vector of patient-visit specific covariates such as age, gender, race, triage level, and chief complaint. $\mathbf{Z}_{i,j}$ is a vector of time related control variables such as year, month, shift, and a weekend indicator variable.[5]

To test for the presence of early task initiation (Hypothesis 4), we switch to $TRITEST_i$ as the dependent variable of the ZINB model. We formulate the linear predictors as follows:

$$\mathbf{x}_{i,j}\boldsymbol{\eta}_j = \alpha_j + \widetilde{\mathbf{aLOAD}}_i\boldsymbol{\beta}_j + \mathbf{W}_{i,j}\boldsymbol{\theta}_j + \mathbf{Z}_{i,j}\boldsymbol{\phi}_j \text{ for } \text{ for } j = 1,2 \tag{8}$$

To test the marginal impact of triage testing on doctor testing (Hypothesis 5), we use the model specified in equation 7 but focus on the marginal effect of $TRITEST$ rather than of $\widetilde{sLOAD}$.

While we do not offer an hypothesis for the net impact of Speedup and Slowdown on service time, we are interested in the empirical result. Since we are again looking at a duration outcome, we use the following AFT model:

$$ln(SERVTIME_i) = \alpha + \widetilde{\mathbf{aLOAD}}_i\boldsymbol{\beta} + \mathbf{W}_i\boldsymbol{\theta} + \mathbf{Z}_i\boldsymbol{\phi} + \epsilon_i \tag{9}$$

This model is the same as equation 4 minus the lab and scan count variables. In this specification, positive coefficients $\boldsymbol{\beta}$, $\boldsymbol{\theta}$,or $\boldsymbol{\phi}$ indicate an increase in service time.

6. Results

To test for evidence of Slowdown effects, we examine the impact of load on task times (Hypothesis 1). Tables 2 and 3 show the results for the ED and the FT respectively. The general pattern we see in

Table 2 Effect of Load on Task Times (ED only)

	(1)		(2)		(3)		(4)	
	1st Order Delay		Lab Collect Time		Med Time		Scan Time	
Wait Census	0.001	(0.001)	0.005***	(0.001)	0.002**	(0.001)	0.004***	(0.001)
ED In-Service	0.031***	(0.001)	0.006***	(0.001)	0.014***	(0.001)	0.013***	(0.002)
FT In-Service	-0.001	(0.003)	0.002	(0.003)	0.008**	(0.004)	0.019***	(0.005)
Boarding	0.007***	(0.001)	0.021***	(0.002)	0.012***	(0.002)	0.011***	(0.002)
N	24,465		21,278		25,344		25,424	

Newey-West HAC robust standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

both the ED and the FT is that task times increase as load increases, which supports Hypothesis 1. We also see that the in-service census for the given area (ED or FT) tends to be the main driver of the increase, which supports the idea of nurse or doctor multitasking leading to increased service

[5] The shift variable indicates the three main physician work shifts: 7:00am-3:00pm, 3:00pm-11:00pm, and 11:00pm-7:00am. We use this shift indicator rather than an hour of day indicator because it captures much of the time of day effect with only two dummy variables rather than twenty three.

Table 3 Effect of Load on Task Times (FT only)

	(1) 1st Order Delay		(2) Lab Collect Time		(3) Med Time		(4) Scan Time	
Wait Census	0.006***	(0.001)	0.002	(0.003)	0.004	(0.004)	0.003	(0.002)
ED In-Service	0.003	(0.002)	0.010*	(0.006)	0.004	(0.006)	-0.005	(0.004)
FT In-Service	0.092***	(0.007)	0.087***	(0.014)	0.052***	(0.017)	0.076***	(0.012)
Boarding	-0.000	(0.003)	0.025***	(0.008)	0.004	(0.007)	0.004	(0.005)
N	10,247		5,449		6,387		7,585	

Newey-West HAC robust standard errors in parentheses
* $p<0.10$, ** $p<0.05$, *** $p<0.01$

times. To get a sense of the magnitude of change in task times, we note that the interquartile range of $EDSERV$ spans from 15.5 patients to 23 patients; a range of 7.5 patients. Multiplying 7.5 by the ED In-Service coefficient and exponentiating the product gives the percent change in the dependent variable. For example, the First Order Delay for ED patients increases by about 26% ($exp(7.5 \times 0.031) = 1.26$) as the number of patients in the ED service beds ranges from the 25th to 75th percentile. That other census measures are significant for some models and not others shows that Slowdown is caused by different factors for different tasks. Still, the general finding remains the same; task times increase with load.

For most of the rest of our analysis, the variable of interest is the three-level load variable. Because of this, we generally report predicted values and pairwise differences between predicted values. This provides a more intuitive interpretation than simply reporting regression coefficients, especially for the ZINB models with two coefficients for each variable. Also, for all models, we run and report the results separately for various subsets of the population. We show results for both the ED and the FT to allow for comparison between these two systems. Also, we show aggregate results for all chief complaints and then for each of the most common chief complaints in the ED and the FT individually. We do this because aggregating patients across chief complaints forces the coefficients of all the variables to be the same across all chief complaints. For example, in the aggregate model, the difference in testing between low and high crowding is the same regardless of whether the patient has a heart attack or a tooth ache. While this is perhaps tolerable for the load variable, it is outright dubious for other variables such as age and gender. By focusing on a single chief complaint at a time we sacrifice sample size but gain tenability.

As we turn our attention to task reduction (Hypothesis 3), we first show that diagnostic tests do indeed increase service time (Hypothesis 2). Table 4 shows the results of estimating Equation 4. All coefficients are positive or insignificant. The exponentiated form of these coefficients can be interpreted as multipliers of the service time. For example, for an abdominal pain patient, each doctor-ordered lab increases the service time by about 4% ($\exp(0.038) = 1.039$). Also note that the doctor-ordered test coefficient is always significantly larger than the related triage-ordered test

Table 4 Effect of Diagnostic Orders on Service Time

	ED			FastTrack		
	(1)	(2)	(3)	(4)	(5)	(6)
	All ED	AP	CP	All FT	LP	BP
TRILAB	-0.001	0.018***	-0.002	0.023***	0.023	0.057***
	(0.002)	(0.004)	(0.006)	(0.007)	(0.043)	(0.020)
DOCLAB	0.024***	0.038***	0.019***	0.142***	0.121***	0.139***
	(0.001)	(0.002)	(0.002)	(0.003)	(0.009)	(0.010)
TRISCAN	0.015***	-0.025	0.108***	0.108***	0.086***	0.216***
	(0.005)	(0.036)	(0.015)	(0.007)	(0.011)	(0.033)
DOCSCAN	0.154***	0.175***	0.186***	0.371***	0.295***	0.514***
	(0.002)	(0.004)	(0.007)	(0.005)	(0.009)	(0.016)
Controls						
Age, Race, Gender	Yes	Yes	Yes	Yes	Yes	Yes
Chief Complaint	Yes	AP only	CP only	Yes	LP only	BP only
Triage	1-5	2,3	2,3	1-5	3-5	3-5
Doctor	Yes	Yes	Yes	No	No	No
Year, Month, Weekend, Hour	Yes	Yes	Yes	Yes	Yes	Yes
N	98,304	12,449	8,499	36,300	5,111	3,103

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
AP: Abdominal Pain, CP: Chest Pain, LP: Limb Pain, BP: Body Pain

coefficient. This speaks to the time savings provided by early task initiation (Hypothesis 4), discussed below.

For task reduction, we examine how the quantity of doctor-ordered tests changes with load, controlling for tests ordered at triage (Table 5). For ED patients in aggregate, column 1 shows a small but statistically significant dip in testing at mid level crowding suggesting some amount of cutting corners. Columns 2 and 3 provide no evidence of cutting corners on specific chief complaints in the ED. The results look quite different for FT patients. Columns 4 and 5 show strong evidence of task reduction for FT patients in aggregate and for limb pain patients in isolation. For example, the predicted mean number of doctor ordered tests drops from 1.13 to 0.89 as load goes from low to high. There is no evidence of cutting corners with body pain patients (Column 6).

To test for early task initiation (Hypothesis 4), we examine how triage testing changes with load (Table 6). Note that in this table we do not separate by ED and FT since that distinction is not made until after triage when the patient is placed in a treatment bed. Thus, we show the results for all patients and for the four most common chief complaints. We see that across the board, triage testing increases with load. For example, the predicted mean number of triage tests for an abdominal pain patient almost triples from 0.397 to 1.019 and roughly quadruples from 0.342 to 1.309 for a chest pain patient as load goes from low to high. This is strong evidence in support of Hypothesis 4. We also examine how doctors and nurse practitioners respond to triage testing (Hypothesis 5). Table 7 shows the marginal effect $\frac{\partial DOCTEST}{\partial TRITEST}$ for several levels of $TRITEST$. Almost

Table 5 Doctor Tests (controlling for triage testing)

	ED			FastTrack		
	(1)	(2)	(3)	(4)	(5)	(6)
	All ED	AP	CP	All FT	LP	BP
Predicted Doctor Orders						
Wait Census: Low	4.81	6.67	5.58	0.96	1.13	1.00
	(0.026)	(0.070)	(0.082)	(0.021)	(0.055)	(0.081)
Wait Census: Mid	4.75	6.63	5.49	0.89	1.04	1.03
	(0.016)	(0.049)	(0.052)	(0.011)	(0.031)	(0.037)
Wait Census: High	4.88	6.73	5.52	0.86	0.89	1.02
	(0.029)	(0.090)	(0.091)	(0.015)	(0.041)	(0.059)
Differences						
Mid vs Low	-0.064**	-0.039	-0.084	-0.065***	-0.088	0.029
	(0.030)	(0.088)	(0.099)	(0.023)	(0.063)	(0.089)
High vs Low	0.065	0.058	-0.059	-0.091***	-0.235***	0.019
	(0.042)	(0.122)	(0.132)	(0.027)	(0.071)	(0.105)
High vs Mid	0.129***	0.097	0.025	-0.026	-0.147***	-0.010
	(0.033)	(0.101)	(0.104)	(0.019)	(0.052)	(0.071)
Controls						
Age, Race, Gender	Yes	Yes	Yes	Yes	Yes	Yes
Triage	1-5	2, 3	2, 3	3-5	3-5	3-5
Triage Test Count	Yes	Yes	Yes	Yes	Yes	Yes
Doctor	Yes	Yes	Yes	Yes	Yes†	Yes
Year, Month, Weekend, Shift	Yes	Yes	Yes	Yes	Yes	Yes
N	98,583	12,482	8,517	35,751	5,113	3,103

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

AP: Abdominal Pain, CP: Chest Pain, LP: Limb Pain, BP: Body Pain

†Variable included in count portion of model only

all of the marginal effects are between negative one and zero indicating that doctors are reducing testing in response to triage testing, but not at a one-for-one ratio. This supports the idea of there being uncertainty in the triage nurse ordering. Further, for ED patients (columns 1 and 2), the marginal effect of $TRITEST$ approaches zero for larger values of $TRITEST$ indicating decreasing marginal benefit of triage testing. This shows that when the triage nurse orders just one test, there is a high probability that this is a useful test and the doctor can reduce her testing orders by one. However, as more triage tests are ordered, the uncertainty in their usefulness increases and each additional test leads to smaller reductions in doctor testing. In contrast, the marginal benefit of triage testing is much smaller for FT patients. This shows that early task initiation is less effective in the FT.

Finally, we look at the net effect of crowding on service time. Table 8 shows the results of the log-logistic AFT regression of service time (Equation 9). Columns 1, 2, and 3 show the results for ED patients. We find evidence of service time first rising and then falling a bit as load moves from low to mid to high. This result matches the pattern seen in Figure 1. This suggests that Slowdown

Table 6 Count of Triage Tests

	(1) All (ED&FT)	(2) AP	(3) CP	(5) LP	(6) BP
Predicted Triage Orders					
Wait Census: Low	0.312	0.397	0.342	0.272	0.276
	(0.005)	(0.015)	(0.021)	(0.013)	(0.014)
Wait Census: Mid	0.547	0.822	0.843	0.470	0.477
	(0.004)	(0.015)	(0.020)	(0.012)	(0.012)
Wait Census: High	0.742	1.019	1.309	0.474	0.550
	(0.009)	(0.031)	(0.042)	(0.020)	(0.023)
Differences					
Mid vs Low	0.235***	0.424***	0.501***	0.197***	0.201***
	(0.007)	(0.022)	(0.0310)	(0.018)	(0.019)
High vs Low	0.430***	0.622***	0.967***	0.201***	0.274***
	(0.011)	(0.036)	(0.048)	(0.025)	(0.028)
High vs Mid	0.195***	0.198***	0.466***	0.004	0.073***
	(0.010)	(0.034)	(0.047)	(0.023)	(0.026)
Controls					
Age, Race, Gender	Yes	Yes	Yes	Yes	Yes
Triage	1-5	1-5	1-5	1-5	1-5
Chief Complaint	Yes	AP only	CP only	LP only	BP only
Year, Month, Weekend, Shift	Yes	Yes	Yes	Yes	Yes
Weekend×Shift	Yes	Yes	Yes	Yes	Yes
N	144,252	14,351	9,689	10,536	10,099

Standard error in parentheses

* $p<0.10$, ** $p<0.05$, *** $p<0.01$

AP: Abdominal Pain, CP: Chest Pain, LP: Limb Pain, BP: Body Pain

Table 7 Marginal Effect of Triage Testing on Doctor Testing

	ED				FastTrack			
	(1) AP		(2) CP		(3) LP		(4) BP	
TRITEST								
0	-1.09	(0.05)	-0.99	(0.05)	-0.27	(0.06)	-0.11	(0.10)
1	-0.96	(0.04)	-0.88	(0.04)	-0.35	(0.04)	-0.14	(0.04)
2	-0.84	(0.03)	-0.79	(0.03)	-0.33	(0.02)	-0.15	(0.06)
3	-0.73	(0.02)	-0.70	(0.02)	-0.21	(0.02)	-0.14	(0.05)
4	-0.64	(0.01)	-0.62	(0.02)	-0.10	(0.02)	-0.12	(0.03)
N	12,482		8,517		5,113		3,103	

Standard error in parentheses

AP: Abdominal Pain, CP: Chest Pain, LP: Limb Pain, BP: Body Pain

effects strongly dominate at first but then as load continues to increase Speedup effects increase and bring the service time back down. However, there is no evidence of Speedup ever being so strong as to reduce the high-load service times below the low-load service times. In contrast, in the FT, there is little evidence of load having any effect on service time. In Column 4 we see an increase of 0.03 hours (1.8 minutes) in service time for all FT patients when going from low to mid load, but

Table 8 Mean Service Time Predictions and Differences

	ED			FastTrack		
	(1)	(2)	(3)	(4)	(5)	(6)
	All ED	AP	CP	All FT	LP	BP
Predicted Mean Service Time						
Wait Census: Low	3.97	5.22	3.68	1.42	1.74	1.48
	(0.02)	(0.06)	(0.06)	(0.02)	(0.04)	(0.06)
Wait Census: Mid	4.13	5.49	3.83	1.45	1.77	1.56
	(0.01)	(0.05)	(0.04)	(0.01)	(0.03)	(0.04)
Wait Census: High	4.04	5.45	3.69	1.46	1.70	1.56
	(0.02)	(0.09)	(0.07)	(0.01)	(0.04)	(0.05)
Differences						
Mid vs Low	0.156***	0.271***	0.150**	0.029*	0.030	0.079
	(0.022)	(0.078)	(0.071)	(0.017)	(0.050)	(.063)
High vs Low	0.063**	0.232**	0.011	0.034	-0.037	0.086
	(0.031)	(0.111)	(0.093)	(0.022)	(0.063)	(0.081)
High vs Mid	-0.093***	-0.038	-0.139*	0.005	-0.068	0.007
	(0.024)	(0.091)	(0.072)	(0.016)	(0.045)	(0.059)
Controls						
Age, Race, Gender	Yes	Yes	Yes	Yes	Yes	Yes
Chief Complaint	Yes	AP only	CP only	Yes	LP only	BP only
Triage	1-5	2,3	2,3	1-5	3-5	3-5
Doctor	Yes	Yes	Yes	No	No	No
Year, Month, Weekend, Hour	Yes	Yes	Yes	Yes	Yes	Yes
N	98,304	12,449	8,499	36,300	5,111	3,103

Standard error in parentheses
Stars displayed for differences only: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
AP: Abdominal Pain, CP: Chest Pain, LP: Limb Pain, BP: Body Pain

no other predicted differences are significant. These results show that in the ED, Slowdown is the dominant result of crowding, while the FT is largely immune from crowding affecting service times.

7. Robustness to Endogenous Treatment and Selection

As with all empirical studies, we must give thought to potential endogeneity issues. There are two potential sources of endogeneity bias in our study: triage testing and patient abandonment. Triage testing is not randomly assigned, but rather is a decision made by a triage nurse based on the characteristics of the patient, some of which are observed (e.g., age, gender, race) and some of which are unobserved to the researcher (e.g., countenance, sweating, pallor). However, triage testing influences the testing decision of the doctor (the coefficient on $TRITEST$ in Equation 7 is significant in all models), and thus it can be considered a "treatment." Just like the triage testing decision, the doctor testing decision is likely driven by many of the same observed and unobserved patient characteristics. A shared unobserved variable could induce correlation in the triage testing and doctor testing models leading to biased estimates of the coefficients. The issue of patient abandonment, also known as Left Without Being Seen (LWBS), further complicates the issue. Patients sometimes

abandon the queue after being triaged but before being seen by a doctor. This abandonment filters the population that the doctor sees. If this filtering changes with crowding, then the doctor is seeing a different patient mix during times of high and low crowding. Further, this filtering is a potential problem because the abandonment rate is affected by triage testing and is possibly driven by the same unobservable covariates affecting triage testing and doctor testing. Thus, there is the potential for a three-way interaction between triage testing, abandonment, and doctor testing. For example, a patient with chest pain who is pale and sweaty may have an increased probability of receiving diagnostic tests both in triage and from the doctor, and might be highly likely to wait to be served since the patient feels quite sick. This would lead to positive uncontrolled correlations among the three equations. Note, however, that all these potential issues only become problematic if the observed covariates are not rich enough to capture the differences between patients. Also, if there is a bias, it is likely that the bias is toward sicker patients remaining and being tested during high crowds. This would be a bias against our hypotheses, and thus our findings are conservative.

The "ideal" test for endogeneity would be a three-equation model that simultaneously estimates the endogenous treatment (triage testing), the self-selection (abandonment), the resulting zero-inflated count outcome (doctor testing) and the respective pairwise correlations. Unfortunately, to the best of our knowledge, no such model exists. The closest model we are aware of is the sample-selection-endogenous-treatment model from Bratti and Miranda (2011). However, this model uses a Poisson model for the final outcome and generally fails to converge with our overdispersed and zero-inflated data. In lieu of an ideal test, we present several pieces of supporting information that point to the conclusion that our results are robust to the potential endogeneity problems.

We begin with the patient abandonment issue. Overall, 6.5% of patients abandon the queue. However, the rate ranges from 3% under low crowding to 12% under high crowding. We use a Heckman-style bivariate probit selection correction model to test for unobserved correlation between patient abandonment and doctor testing (de Ven and Praag 1981, Greene 2012). We treat both the abandonment decision and doctor testing as binary outcomes and formulate the model as follows:

$$\begin{gathered} S^* = \alpha_1 + \mathbf{a\widetilde{LOAD}}\beta_\mathbf{1} + \delta_1 \mathbf{1}(TRITEST > 0) + \mathbf{W_1}\boldsymbol{\theta_1} + \mathbf{Z_1}\boldsymbol{\phi_1} + \varepsilon_1 \\ STAY = 1 \text{ if } S^* > 0,\ 0 \text{ otherwise} \end{gathered} \tag{10}$$

$$\begin{gathered} D^* = \alpha_2 + \mathbf{s\widetilde{LOAD}}\beta_\mathbf{2} + \delta_2 TRITEST + \gamma_2 FT + \mathbf{W_2}\boldsymbol{\theta_2} + \mathbf{Z_2}\boldsymbol{\phi_2} + \varepsilon_2 \\ DOCTEST\_YN = 1 \text{ if } D^* > 0,\ 0 \text{ otherwise} \end{gathered} \tag{11}$$

The vectors $\mathbf{W_1}$and $\mathbf{W_2}$contain the patient covariates age, gender, race, chief complaint, and triage level. The variable FT is a dummy variable indicating if the patient was treated in the FastTrack. The vector $\mathbf{Z_1}$ contains controls for year, month, weekend, and shift, while the vector $\mathbf{Z_2}$ contains controls for only weekend and shift. We drop the year and month variables from the

second equation to provide an exclusion restriction to help with model identification even though the model technically is identified by the non-linearity of the probit equations. ε_1 and ε_2 are assumed to be standard bivariate normally distributed with correlation coefficient ρ, and Equation 11 is only observed if $STAY = 1$. If $\rho = 0$, this indicates that the control variables are adequately controlling for the selected sample and the models can be estimated separately without significant bias. We see in Table 9 that indeed the estimated correlations are insignificantly different from zero for models 2, 3, and 5, but for models 1 and 4, the correlation is positive and significant. The coefficients in

Table 9 Heckman Probit Selection model of Abandonment and Doctor Testing

	(1) All	(2) AP	(3) CP	(4) LP	(5) BP
Stay (Y/N)					
Wait Census Mid	-0.469***	-0.744***	-0.625***	-0.210***	-0.576***
	(0.018)	(0.053)	(0.075)	(0.065)	(0.070)
Wait Census High	-0.890***	-1.411***	-1.174***	-0.538***	-0.995***
	(0.020)	(0.059)	(0.088)	(0.076)	(0.080)
Doctor Test (Y/N)					
Wait Census Mid	-0.043***	-0.018	-0.204***	-0.069**	0.006
	(0.011)	(0.045)	(0.077)	(0.034)	(0.039)
Wait Census High	-0.052***	-0.023	-0.361**	-0.175***	0.072
	(0.016)	(0.073)	(0.148)	(0.044)	(0.058)
ρ	0.148***	-0.201	0.657	0.636***	-0.302
	(0.043)	(0.213)	(0.294)	(0.119)	(0.233)
Age, Race, Gender, Triage	Yes	Yes	Yes	Yes	Yes
Chief Complaint	Yes	AP only	CP only	LP Only	BP Only
FastTrack†	Yes	Yes	Yes	Yes	Yes
Year††, Month††, Weekend, Shift	Yes	Yes	Yes	Yes	Yes
N	144,252	14,351	9,689	10,536	10,099

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

†Variable included in Doctor Test Y/N portion of model only

††Variable included in selection (Stay Y/N) portion of model only

AP: Abdominal Pain, CP: Chest Pain, LP: Limb Pain, BP: Body Pain

the upper panel show that the probability of staying (not abandoning) decreases with load, as one would expect. The coefficients in the lower panel indicate that for all patients in aggregate and for chest pain and limb pain patients (columns 1, 3, and 4), doctors are less likely to order tests during high crowding, whereas in Table 5 we only saw limited evidence of cutting corners under load. These results show that while the observed covariates are controlling for much of the patient differences, correcting for the remaining correlation between self-selected abandonment and doctor testing only strengthens our findings.

We also check for unobserved correlation between triage testing and patient abandonment. We use a bivariate probit model similar to the selection model above, but without needing to adjust

for the selected sample. Table 10 shows that the census coefficients are all significant and in the

Table 10 Bivariate Probit of Triage Test and Stay/LWBS

	(1) All ED	(2) AP	(3) CP	(4) LP	(5) BP
Triage Test (Y/N)					
Wait Census Mid	0.496***	0.631***	0.685***	0.426***	0.434***
	(0.011)	(0.031)	(0.037)	(0.038)	(0.039)
Wait Census High	0.646***	0.726***	0.942***	0.447***	0.457***
	(0.014)	(0.039)	(0.046)	(0.048)	(0.049)
Stay (Y/N)					
Wait Census Mid	-0.413***	-0.677***	-0.658***	-0.273***	-0.505***
	(0.019)	(0.059)	(0.084)	(0.081)	(0.070)
Wait Census High	-0.804***	-1.326***	-1.221***	-0.588***	-0.920***
	(0.023)	(0.070)	(0.103)	(0.081)	(0.081)
ρ	0.264***	0.184**	-0.078	-0.422	0.412***
	(0.037)	(0.084)	(0.125)	(0.237)	(0.123)
Age, Race†, Gender	Yes	Yes	Yes	Yes	Yes
Chief Complaint	Yes	AP only	CP only	LP Only	BP Only
Triage	1-5	1-5	1-5	1-5	1-5
Year, Month, Weekend, Shift	Yes	Yes	Yes	Yes	Yes††
N	107,825	13,802	9,193	10,536	10,099

Standard errors in parentheses; * $p<0.10$, ** $p<0.05$, *** $p<0.01$
†Race included in Triage Test portion of model only
††Month included in Triage Test portion of model only
AP: Abdominal Pain, CP: Chest Pain, LP: Limb Pain, BP: Body Pain

direction we expect; crowding increases triage testing and abandonment. We also see that models 1, 2, and 5 show significant positive correlation in the errors. However, if we repeat the analysis for patients of a single triage level at a time, then the correlation becomes insignificant. Together, these two sets of results suggest that patient abandonment may create a bias in the results, but any bias that does exist makes our findings conservative since the correlations are all positice. Further, these robustness checks suggest that the bias can largely be corrected for with our control variables and by focusing on a single triage level at a time.

To examine the potential endogeneity between triage testing and doctor testing we again use a bivariate probit model. We ignore the middle step of abandonment based on the above results showing that there is not a significant bias. The results of this analysis are mixed in that some models show significant between-equation correlation, and others do not (Table 11). The coefficients in the upper panel are all as expected indicating increased triage testing with increased crowding. With the exception of Column 6, the coefficients in the lower panel are as expected, showing either no change or a decrease in doctor testing with load, controlling for triage testing. Column 6 shows a slight increase in doctor testing when crowding is at the mid level. However, the two load dummy

Table 11 Bivariate Probit of Triage Testing and Doctor Testing

	ED			FastTrack		
	(1)	(2)	(3)	(4)	(5)	(6)
	All ED	Abd. Pain	Chest Pain	All FT	Limb Pain	Body Pain
Triage Test (Y/N)						
Wait Census Mid(a)	0.571***	0.672***	0.713***	0.322***	0.347***	0.303***
	(0.013)	(0.033)	(0.039)	(0.024)	(0.052)	(0.075)
Wait Census High(a)	0.819***	0.891***	1.062***	0.385***	0.407***	0.343***
	(0.016)	(0.044)	(0.049)	(0.029)	(0.060)	(0.085)
Doctor Test (Y/N)						
Wait Census Mid(s)	-0.037***	-0.037	-0.117**	-0.024	-0.087*	0.118*
	(0.013)	(0.046)	(0.057)	(0.020)	(0.051)	(0.062)
Wait Census High(s)	-0.031*	-0.047	-0.211***	-0.038	-0.214***	0.067
	(0.018)	(0.062)	(0.073)	(0.024)	(0.058)	(0.069)
ρ	-0.060 ***	-0.022	-0.172***	-0.136***	-0.409***	0.008
	(0.010)	(0.0932)	(0.034)	(0.019)	(0.041)	(0.068)
Age	Yes	Yes	Yes	Yes	Yes†	Yes†
Race	Yes	Yes	Yes	Yes†	Yes†	Yes†
Gender	Yes	Yes	Yes	Yes	Yes	Yes
Chief Complaint	Yes	AP only	CP only	Yes	LP only	BP only
Triage	1-5	2,3	2,3	3-5	3-5	3-5
Year	Yes	Yes	Yes	Yes	Yes	Yes†
Month	Yes†	Yes†	Yes†	Yes†	Yes†	Yes†
Weekend	Yes	Yes	Yes	Yes	Yes†	No
Shift	Yes	Yes	Yes	Yes	Yes	No
N	98,583	12,482	8,517	35,751	5,113	3,103

Standard errors in parentheses; * $p<0.10$, ** $p<0.05$, *** $p<0.01$

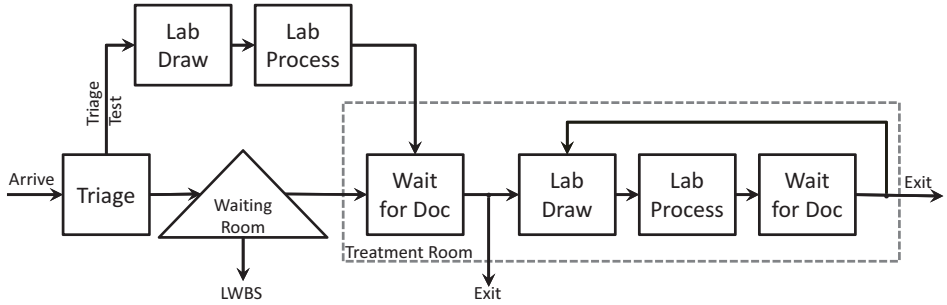
†Variable included in Triage Test portion of model only

Wait Census Low is omitted category

variables (Wait Census Mid & Wait Census High) are jointly insignificant and the fit of the model actually improves if the load variables are removed from the doctor testing equation. Thus, we can safely conclude that across all six columns of Table 11 we see that correcting for potential unobserved correlation only strengthens our conclusion that doctors sometimes reduce testing as crowding increases.

To further check the robustness of our findings regarding the presence of task reduction (Hypothesis 3), we repeat the main study reported in Table 5 with two special subsets of the data. We first test for task reduction for patients that receive no triage tests. Clearly, this is a non-random sample, but it is free of any convoluting effects of doctors responding to triage testing. We find largely the same results as in Table 5 with little evidence of task reduction in the ED while task reduction is present for FT patients in aggregate and for limb pain patients specifically. The second subset we examine is whether abdominal pain and headache patients receive a radiology scan. About 40% of these patients receive a scan, but the scan is ordered by the doctor 99% of the time. Thus, this sample is effectively clear of triage testing treatment bias. We find no evidence of reduced testing

Figure 4 **Patient Flow in Simulation Model**



under crowding. Taken together, all these robustness checks support or strengthen our main findings regarding Hypothesis 3 that doctors make limited use of task reduction under crowding.

8. Simulation

Given our findings of several forms of state-dependent service times in the ED, we are interested in determining what impact these have on performance models. To estimate the impact of the state-dependencies, we build a discrete-event-simulation (DES) model of the ED. Figure 4 diagrams the patient flow in the model. While the model is abstracted from reality, we maintain the essential elements that allow for state-dependent service times, namely the triage testing and doctor testing decisions are state-dependent, and the processing times for Lab Draw and Wait for Doc are state-dependent as well.[6] One additional state-dependency included in the model is the Left Without Being Seen or abandonment rate. While we do not focus on this phenomenon in this paper, our data clearly exhibits a strong positive correlation between LWBS and waiting room census.

We test three configurations of the model (Table 12). In the first configuration (column 1), all state-dependent variables are included and the model is tuned to match the average performance of our study ED. In the second configuration (column 2), the Speedup and Slowdown state-dependencies are deactivated by fixing all variables at their mean values. In the third configuration (column 3), all state-dependencies, including LWBS, are deactivated. The simulation is run for 50,000 simulated hours and standard errors are calculated using the batch-means process with batches of length 200 hours (Law 2007).

Comparing column 2 to column 1 we see that ignoring the Speedup and Slowdown mechanisms leads to a small overestimation of all of the performance measures. Comparing column 3 to column 1 we see that ignoring all state-dependencies leads to a larger overestimation of all performance measures. This potential overestimation is managerially relevant since similar models are commonly used for hospital staffing and planning purposes. These planning models are becoming increasingly

[6] We leave the lab processing time distribution stationary because the lab serves the entire hospital and the ED demand has little impact on lab times.

Table 12 Simulation Results

Outcome (mean)	(1) State-Dependent	(2) State-Independent (except LWBS)	(3) State-Independent (incl. LWBS)
Queue Length	8.3 (0.21)	8.8 (0.17)	9.9 (0.64)
Wait Time (hr.)	1.6 (0.04)	1.7 (0.03)	2.0(0.1)
Length of Stay (hr.)	5.6 (0.05)	5.8 (0.03)	6.1 (0.10)
LWBS %	5.8% (0.002)	6.2% (0.001)	8.6% (0.001)

Standard error in parentheses

important as the Centers of Medicare and Medicaid Services (CMS) begin to phase in new ED reporting guidelines and performance targets. Hospitals will soon be required to report performance measures such as median wait time, median length of stay, and "Left Without Being Seen" percentage (Centers for Medicare & Medicaid Services 2012). Eventually, target values will be established and hospitals will be reimbursed based on their performance relative to the targets. Thus, a hospital that is making planning decisions based on a model which does not include the identified state-dependencies is likely to overinvest in resources and staffing to meet the CMS targets.

9. Discussion & Future Work

Prior research has shown that worker-paced service systems tend to exhibit state-dependent service times. In this paper we explore the mechanisms that lead to state-dependent service times whether from a single resource or between multiple resources. We find evidence of both Speedup and Slowdown mechanisms. In our setting, the Slowdown effects tend to dominate in the emergency department, while in the FastTrack, the effects of Slowdown and Speedup balance out.

We find strong evidence of triage-ordered testing being used to reduce in-room service time during periods of crowding in both the ED and the FT. Triage testing saves time by starting tests sooner and allowing at least some of the lab collection and processing time to occur in parallel with the patient waiting time. The main downside to triage testing is the financial cost of unneeded tests. Since neither an insured patient nor the triage nurse directly incur the financial cost, it likely does not weigh heavily on the testing decision. Given the effectiveness of triage testing as a form of Speedup, it is curious that triage testing is not used more regularly, regardless of crowd level. Our findings suggest that hospitals could potentially benefit from increased use of triage testing. Managers should further explore the true costs of over testing at triage and consider incorporating load-based guidelines into triage nurse protocols.

We find evidence of care providers reducing testing orders in the FT when the system is crowded but only limited evidence of this in the ED. In the healthcare setting, task reduction is clearly a double-edged sword. On the one hand, reducing testing speeds up service, reduces the load on the auxiliary services, and reduces costs. On the other hand, reduced testing may result in decreased

quality of care. (We found no evidence of crowding leading to an increase in 72-hour revisits, a common ED quality metric, in either the ED or the FT.) Determining the "optimal" level of corner cutting is an empirical medical question and is beyond the scope of this paper. Further, it is related to the philosophical question of what should be the role of the ED in the larger health care delivery system? Should the ED be the site of definitive medical care, or should it only serve to stabilize and route to the appropriate resource for full identification and care of the presenting medical condition? This is an ongoing debate in the medical community (Schuur and Venkatesh 2012, Wiler et al. 2012). As Operations Management researchers, we are satisfied to show that task reduction under load does exist in some circumstances and serves to speed up a service system. Thus, again our work suggests that hospital managers should explore the quality trade-offs of task reduction and should potentially include load-based guidelines in care protocols.

Lastly, we find that ignoring state-dependencies leads to inaccurate planning models. In our setting, the error was an overestimation of system busyness. Our results show that it is important to incorporate state-dependent mechanisms into planning models to avoid overinvestment in staffing and physical resources. Our results also show the value of identifying and measuring state-dependencies. While this work focused on server-level state-dependencies, future work should also look at patient-level state-dependencies.

In conclusion, our work expands upon the prior state-dependent service time literature and shows that there can be several server-level mechanisms at work as servers respond to work load. We hope that incorporation of these mechanisms into future normative models will lead to better understanding and management of similar service systems with high server discretion.

Appendix A: Log-Likelihood Function of Zero Inflated Negative Binomial Model

The negative binomial logit hurdle model is estimated by maximization of the log-likelihood function. The function is derived from the combination of a logit model and a negative binomial count model. The function is given below and is based on the function shown in Hilbe (2011, p372). However, the formula in the book contains errors.

$$\mathcal{L}(\beta_1,\beta_2;y,\alpha)=\begin{cases}\ln\left(\frac{1}{1+\exp\left(-x_i'\beta_1\right)}\right)+\left(\frac{1}{1+\exp\left(x_i'\beta_1\right)}\right)\left(\frac{1}{1+\exp\left(x_i'\beta_2\right)}\right)^{1/\alpha} & \text{if } y=0\\ \ln\left(\frac{1}{1+\exp\left(x_i'\beta_1\right)}\right)+\frac{1}{\alpha}\ln\left(\frac{1}{1+\alpha\exp\left(x_i'\beta_2\right)}\right) & \\ +\ln\Gamma\frac{(y_i+1/\alpha)}{(y_i+1)(1/\alpha)}+y_i\ln\left(1-\frac{1}{1+\alpha\exp\left(x_i'\beta_2\right)}\right) & \text{if } y>0\end{cases}$$

References

Aksin, O. Zeynep, Patrick T. Harker. 2001. Modeling a phone center: Analysis of a multichannel, multiresource processor shared loss system. *Management Science* **47**(2) 324–336. doi:10.1287/mnsc.47.2.324.9842.

Alizamir, Saed, Francis deVericourt, Peng Sun. 2011. Diagnostic accuracy under congestion. *Working Paper* .

Allon, Gad, Sarang Deo, Wuqin Lin. 2011. The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Working Paper* .

Armony, Mor, Shlomo Israelit, Avishai Mandelbarum, Yarvin N Marmor, Yulia Tseytlin, Galit B. Yom-Tov. 2012. Patient flow in hospitals: A data-based queuing-science perspective. *Working Paper* .

Bratti, Massimiliano, Alfonso Miranda. 2011. Endogenous treatment effects for count data models with endogenous participation or sample selection. *Health Economics* **20**(9) 1090–1109.

Caldwell, John A. 2001. The impact of fatigue in air medical and other types of operations: A review of fatigue facts and potential countermeasures. *Air Medical Journal* **20**(1) 25 – 32. doi:10.1016/S1067-991X(01)70076-4.

Centers for Medicare & Medicaid Services. 2012. Hospital outpatient prospective and ambulatory surgical center payment systems and quality reporting programs; electronic reporting pilot; inpatient rehabilitation facilities quality reporting program; quality improvement organization regulations. *Federal Register* **77**(146) 45061–45233.

Chan, Carri, Vivek F. Farias, Nicholas Bambos, Gabriel J. Escobar. 2011. Optimizing icu discharge decisions with patient readmissions. *Working Paper* .

Chen, Chao, Zhanfeng Jia, P. Varaiya. 2001. Causes and cures of highway congestion. *Control Systems, IEEE* **21**(6) 26 –32. doi:10.1109/37.969132.

Crabill, Thomas B. 1972. Optimal control of a service facility with variable exponential service times and constant arrival rate. *Management Science* **18**(9) 560–566.

de Ven, Wynand P.M.M. Van, Bernard M.S. Van Praag. 1981. The demand for deductibles in private health insurance: A probit model with sample selection. *Journal of Econometrics* **17**(2) 229 – 252. doi:10.1016/0304-4076(81)90028-2.

deVericourt, Francis, Otis B. Jennings. 2011. Nurse staffing in medical units: A queueing perspective. *Operations Research* **59**(6) 1320–1331. doi:10.1287/opre.1110.0968.

Fee, Christopher, Ellen J. Weber, Carley A. Maak, Peter Bacchetti. 2007. Effect of emergency department crowding on time to antibiotics in patients admitted with community-acquired pneumonia. *Annals of Emergency Medicine* **50**(5) 501–509.e1.

George, Jennifer M., J. Michael Harrison. 2001. Dynamic control of a queue with adjustable service rate. *Operations research* **49**(5) 720–731.

Gerla, M., L. Kleinrock. 1980. Flow control: A comparative survey. *Communications, IEEE Transactions on* **28**(4) 553 – 574. doi:10.1109/TCOM.1980.1094691.

Green, Linda V. 2006. *Patient Flow: Reducing Delay in Healthcare Delivery*, *International Series in Operations Research & Management Science*, vol. 91, chap. Queuing Analysis in Healthcare. Springer.

Green, Linda V., Sergei Savin, Ben Wang. 2006. Managing patient service in a diagnostic medical facility. *Operations Research* **54**(1) 11–25. doi:10.1287/opre.1060.0242.

Greene, William H. 2012. *Econometric Analysis*. 7th ed. Prentice Hall.

Hilbe, Joseph M. 2011. *Negative Binomial Regression*. 2nd ed. Cambridge University Press.

Hopp, Wallace J., Seyed M. R. Iravani, Gigi Y. Yuen. 2007. Operations systems with discretionary task completion. *Management Science* **53**(1) 61–77.

Ibrahim, Rouba, Ward Whitt. 2011. Wait-time predictors for customer service systems with time-varying demand and capacity. *Operations Research* **59**(5) 1106–1118. doi:10.1287/opre.1110.0974.

Imai, K., I Nonaka, H. Takeuchi. 1985. Managing the new product development process: How japanese compaies learn and unlearn. K. B. Clark, R. H. Hayes, C. Lorenz, eds., *The Uneasy Alliance: Managing the Productivity-Technology Dilemma*. Havard Business School Press, Cambridge, MA.

Kc, Diwas S., Christian Terwiesch. 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* **55**(9) 1486–1498.

KC, Diwas Sign. 2011. Does multi-tasking make you more productive? *Working Paper* .

Law, Averill M. 2007. *Simulation Modeling and Analysis*. 4th ed. McGraw Hill.

Lee, Donald K. K., Stefanos A. Zenios. 2009. Optimal capacity overbooking for the regular treatment of chronic conditions. *Operations Research* **57**(4) 852–865. doi:10.1287/opre.1080.0666.

Loch, Christoph H., Christian Terwiesch. 1998. Communication and uncertainty in concurrent engineering. *Management Science* **44**(8) 1032–1048. doi:10.1287/mnsc.44.8.1032.

Lucas, Ray, Heather Farley, Joseph Twanmoh, Andrej Urumov, Nils Olsen, Bruce Evans, Hamed Kabiri. 2009. Emergency department patient flow: The influence of hospital census variables on emergency department length of stay. *Academic Emergency Medicine* **16**(7) 597–602.

McCarthy, Melissa L., Scott L. Zeger, Ru Ding, Scott R. Levin, Jeffrey S. Desmond, Jennifer Lee, Dominik Aronsky. 2009. Crowding delays treatment and lengthens emergency department length of stay, even among high-acuity patients. *Annals of Emergency Medicine* **54**(4) 492–503.e4.

Oliva, Rogelio, John D. Sterman. 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, Harold. 1994. Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin* **116**(2) 220–224. doi:10.1037/0033-2909.116.2.220.

Pines, Jesse M., Judd E. Hollander. 2008. Emergency department crowding is associated with poor care for patients with severe pain. *Annals of Emergency Medicine* **51**(1) 1–5.

Pines, Jesse M., Judd E. Hollander, A. Russell Localio, Joshua P. Metlay. 2006. The association between emergency department crowding and hospital performance on antibiotic timing for pneumonia and percutaneous intervention for myocardial infarction. *Academic Emergency Medicine* **13**(8) 873–878.

Pines, Jesse M., Sanjay Iyer, Maureen Disbot, Judd E. Hollander, Frances S. Shofer, Elizabeth M. Datner. 2008. The effect of emergency department crowding on patient satisfaction for admitted patients. *Academic Emergency Medicine* **15**(9) 825–831.

Pines, Jesse M., Anjeli Prabhu, Joshua A. Hilton, Judd E. Hollander, Elizabeth M. Datner. 2010. The effect of emergency department crowding on length of stay and medication treatment times in discharged patients with acute asthma. *Academic Emergency Medicine* **17**(8) 834–839.

Schultz, Kenneth L., David C. Juran, John W. Boudreau, John O. McClain, L. Joseph Thomas. 1998. Modeling and worker motivation in jit production systems. *Management Science* **44**(12-Part-1) 1595–1607. doi:10.1287/mnsc.44.12.1595.

Schuur, Jeremiah D., Arjun K. Venkatesh. 2012. The growing role of emergency departments in hospital admissions. *New England Journal of Medicine* **367**(5) 391–393. doi:10.1056/NEJMp1204431.

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

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

Stidham, Shaler, Richard R. Weber. 1989. Monotonic and insensitive optimal policies for control of queues with undiscounted costs. *Operations research* **37**(4) 611–625.

Takeuchi, Hirotaka, Ikujiro Nonaka. 1986. The new new product development game. *Harvard Business Review* **64**(1) 137 – 146.

Tan, Tom, Serguei Netessine. 2012. When does the devil make work? an empirical study of the impact of workload on worker productivity. *Working Paper* .

Wiler, Jennifer L., Dennis Beck, Brent R. Asplin, Michael Granovsky, John Moorhead, Randy Pilgrim, Jeremiah D. Schuur. 2012. Episodes of care: Is emergency medicine ready? *Annals of Emergency Medicine* **59**(5) 351 – 357. doi:10.1016/j.annemergmed.2011.08.020.

Wolff, Ronald W. 1989. *Stochastic Modeling and the Theory of Queues*. Industrial and Systems Engineering, Prentice Hall Inc., Upper Saddle River, NJ.

Wooldridge, Jeffery M. 2009. *Introductory Econometrics: A Modern Approach*. 4th ed. South-Western Cengage Learning.

Yamazaki, Genji, Hirotaka Sakasegawa. 1987. An optimal design problem for limited processor sharing systems. *Management Science* **33**(8) 1010–1019. doi:10.1287/mnsc.33.8.1010.

Yankovic, Natalia, Linda V. Green. 2012. A queuing model for nurse staffing. *Working Paper* .