The Impact of Pooling on Throughput Time in Discretionary Work Settings: An Empirical Investigation of Emergency Department Length of Stay

HARVARD | BUSINESS | SCHOOL

The Impact of Pooling on Throughput Time in Discretionary Work Settings: An Empirical Investigation of Emergency Department Length of Stay

Hummy Song
Anita L. Tucker
Karen L. Murrell

Working Paper

13-079

March 19, 2013

The Impact of Pooling on Throughput Time in Discretionary Work Settings:

An Empirical Investigation of Emergency Department Length of Stay

Hummy Song
Harvard University
Boston, MA 02163
hsong@hbs.edu

Anita L. Tucker
Harvard Business School
Boston, MA 02163
atucker@hbs.edu

Karen L. Murrell
Kaiser Permanente South Sacramento Medical Center
South Sacramento, CA 95823
karen.l.murrell@kp.org

Revised: March 19, 2013

Acknowledgments

This research would not have been possible without the collaboration of the Kaiser Permanente South Sacramento Medical Center's Emergency Department (KP SSC ED). In particular, the authors are deeply appreciative of the support provided by Mark B. Kauffman, Director of Emergency Systems and Delivery Systems Optimization at KP SSC ED. The authors are also grateful to Brent E. Soon, Senior Financial Analyst at KP SSC, for providing us access to the dataset used in this paper. The authors thank Alexandra A. Killewald and participants in the Longitudinal Data Analysis course at Harvard University for their insightful comments, and William B. Simpson, Simo Goshev, and Andrew Marder for their advice regarding data analysis methods.

Abstract

We conduct an empirical investigation on the impact of pooling tasks and resources on throughput times in a discretionary work setting. We use an Emergency Department's (ED) patient-level data (N = 234,334) from 2007 to 2010 to test our hypotheses. We find that when the ED's work system had pooled tasks and resources, patients' lengths of stay were longer than when the ED converted to having dedicated tasks and resources. More specifically, we find that dedicated systems resulted in a 9 percent overall decrease in length of stay, which corresponds to a 25-minute reduction in length of stay for an average patient of medium severity in this ED. We propose that the improved performance comes from a reduction in social loafing and a more distributed utilization of shared resources. These benefits outweigh the expected efficiency gains from pooling, which are commonly predicted by queuing theory.

Key words: pooling, discretionary work, social loafing, shared resources, empirical operations

1. Introduction

Improving the productivity of their organizations' operating systems is an important objective for managers. Pooling—an operations management technique—has been proposed as a way to improve performance by reducing the negative impact of demand variability (Eppen, 1979). This technique enables incoming work to be processed by any one of a bank of servers, which deceases the odds that an incoming unit of work will arrive to the processing step and have to wait, as compared to the situation where the unit of work can only be processed by a single dedicated server.

Using queuing theory and associated analytical models, operations management scholars have shown that pooling multiple demand streams across multiple servers can enable shorter waiting times and higher average utilization, thereby increasing productivity (Alfaro & Corbett, 2009; Alptekinoglu, Banerjee, Paul, & Jain, 2012; Benjaafar, Cooper, & Kim, 2005; Corbett & Rajaram, 2006; Eppen, 1979; Mandelbaum & Reiman, 1998). This prediction that pooling should reduce waiting time is an important result for service organizations (Jouini, Dallery, & Nait-Abdallah, 2008). In this context, customers are the "inventory" that benefit from shorter waiting times, which in turn reduces throughput times. Service organization workers who operate in a pooled system have an increased level of utilization, because they can work on any type of incoming demand, as opposed to just one special type, and thus have less idle time. This increased level of utilization makes the entire operating system more productive.

There exists a substantial body of work that examines the benefits of pooling different types of inventory into one system (Alfaro & Corbett, 2009; Alptekinoglu et al., 2012; Benjaafar et al., 2005; Corbett & Rajaram, 2006; Graves & Tomlin, 2003) and pooling multiple queues into one stream (Ata & Van Mieghem, 2008; Gans, Koole, & Mandelbaum, 2003; Smith & Whitt, 1981; Stidham, 1970; van Dijk, 2002; Wallace & Whitt, 2005; Whitt, 1992). This body of research can be characterized by two main factors. First, it is primarily focused on non-discretionary work settings, or "blue-collar" work, which involves highly specified and routine tasks (Hopp, Iravani, & Liu, 2009). This is in contrast to "white-collar" work, which involves professional workers who can exercise discretion over their intellectual and creative tasks. Second, most research on this topic has developed analytical models to examine the benefits of pooling.

Yet, over the past several decades, there has been a steady shift toward service and professional jobs that engage workers in discretionary work as opposed to non-discretionary work. For example, a doctor has discretion in the plan of care she designs for her patient, a lawyer can determine how to best prepare a case, and a master chef has creative license in creating new dishes. However, we know very little about the impact of pooling in discretionary work settings. This is because it is difficult to apply many of the results from "blue-collar" research to "white-collar" systems, due to differences between non-discretionary and discretionary work. In particular, the former assumes that workers are either inflexible

or are given very limited flexibility, which is in fact a defining characteristic of discretionary work systems. In addition, given the focus of prior research on developing analytical models as opposed to conducting empirical studies, we know little about how human behavior affects the operational performance of pooled systems where workers have a high degree of discretion over their tasks. As the call for more empirical research in the field of operations management suggests, this a significant limitation because empirical research may be better equipped to explore the impact of employee discretion on the operational performance of pooled operations (Boudreau, Hopp, McClain, & Thomas, 2003; Flynn, Sakakibara, Schroeder, Bates, & Flynn, 1990; Meredith, 1998; Roth & Menor, 2009; Scudder & Hill, 1998).

This paper makes several contributions to research. We contribute to the operations management literature on pooling by examining the effect of pooling on throughput time in a *discretionary* work setting—an area that to our knowledge has been understudied. In addition, we contribute to the operations management literature more broadly by using empirical data to test theory related to the impact of pooling on worker productivity. We find that in a hospital emergency department (ED), where workers have discretionary control over work content, pace, and resources, pooling is associated with *longer* throughput times. This is in contrast to results predicted by queuing theory's analytical models, which suggest that pooling should result in *shorter* throughput times.

In particular, we examine the impact of pooling by distinguishing two types of pooling: task pooling and resource pooling. We define task pooling as the situation in which each worker of a group pulls his or her next job from a shared set of tasks that are waiting to be processed by any one of the workers. Similarly, resource pooling occurs when a group of workers draw from a shared set of resources that are available for use by any one of the workers. This distinction is made to better examine the interaction of human behavior and operations in our analyses of the effects of pooling. We find that task pooling and resource pooling are each associated with longer throughput times, which together contribute to the overall effect of pooling on throughput times.

The findings of our paper have important implications for the organization of workflow in discretionary work settings, such as the ED. This is an especially timely finding with regard to health care delivery, as EDs across the country contemplate ways to handle the large increases in their patient volume. Our results suggest that organizing EDs so that (1) a physician is assigned to care for the patient upon the patient's arrival to the ED (non-pooled tasks) and (2) each physician has a dedicated team of nurses working for her during the shift (non-pooled resources) may engender physician behaviors that result in shorter lengths of stay and increased capacity to handle larger patient volumes than if EDs were organized with pooled tasks and resources.

2. Pooling in Non-discretionary Work

Much of the research on pooling has been conducted with reference to non-discretionary work (Hopp et al., 2009). In such settings, tasks are more precisely defined and controlled than in "white collar" settings. For example, in a manufacturing setting, an incoming unit of work is automatically processed by a server—which is usually a machine—as soon as one becomes available. The machine does not have discretion to process the work at a faster or slower speed. Each step of the process is precisely specified beforehand, and variation in the work does not arise as a function of personal discretion.

Under this general framework of non-discretionary work, operations scholars have distinguished two types of settings in which pooling may occur: production-inventory systems and queuing networks. Inventory pooling, defined as "the practice of using a common pool of inventory to satisfy two or more sources of random demand," has been studied in contexts such as manufacturing firms and supply chains (Alptekinoglu et al., 2012, p. 33). Since the seminal paper by Eppen (1979) on the benefits of consolidating demand, an extensive body of work developing analytical models has documented the effect of pooling on production-inventory systems (Alfaro & Corbett, 2009; Gerchak & He, 2003). These models account for factors such as utilization, demand and process variability, service levels, and structure of the production process to determine the impact of pooling. They overwhelmingly conclude that the benefits of pooling are robust to constraints such as perfectly correlated demands (Alptekinoglu et al., 2012), high utilization (Benjaafar et al., 2005), suboptimal inventory policies (Alfaro & Corbett, 2009), non-normal dependent demand (Corbett & Rajaram, 2006), and long lead times (Tagaras & Cohen, 1992).

In queuing networks, operations researchers have focused on the effect of pooling queues, resources, and, to a lesser degree, tasks (Mandelbaum & Reiman, 1998). Much of this work has been conducted with reference to call centers, and has found that the benefits of having flexible servers and pooled queues outweighs its potential drawbacks (Anupindi, Chopra, Deshmukh, Van Mieghem, & Zemel, 2005; Ata & Van Mieghem, 2008; Gans et al., 2003; Jouini et al., 2008). Researchers have reached similar conclusions in other settings, such as mail delivery, and have found that pooling improves service quality while concurrently reducing costs (Ata & Van Mieghem, 2008). Much less work has been carried out on task pooling, despite Loch's (1998) argument for its importance in thinking about alignment of incentives and process design.

3. Pooling in Discretionary Work

Discretionary work is distinguished from non-discretionary work based on a comparison of the types of *tasks* that workers conduct. In contrast to non-discretionary work, as was described above, workers who

engage in discretionary work function in the absence of highly prescribed and detailed operational rules. According to a survey of the literature by Hopp et al. (2009), discretionary work is different from non-discretionary work in seven key ways: (i) discretionary tasks are intellectual and/or creative in nature, (ii) discretionary work relies more heavily on knowledge-based resources, (iii) learning is slower and more central in discretionary work systems, (iv) output measurement is more difficult in discretionary work systems, (v) discretionary work is likely to involve self-generated work in addition to externally generated work, (vi) workers have more discretion over processing times in discretionary work systems, and (vii) incentives play a more important role in discretionary work. In short, discretionary workers have the flexibility to make discretionary decisions to spend extra time and effort on their work, which ultimately affects throughput time, cost, and other operational outcome measures. Because of their capacity to make discretionary decisions, they are also more likely to be influenced by learning, performance measures, incentives, and technology.

To our knowledge, only a handful of papers examine the effect of pooling in discretionary work settings on worker productivity or performance. First, Hopp et al. (2007) examine systems with discretionary task completion, which they define as systems under which the completion criteria for tasks are determined by a worker's subjective standards. They formulate a model that predicts that, though smaller than in nondiscretionary task completion systems, pooling in discretionary task completion systems still exhibits operational benefits. This reduction in the benefit of pooling is due to the introduction of quality as an additional factor for buffering variability. In essence, workers with discretion are likely to choose to spend more time with a customer when tasks (customers) are pooled because by providing higher quality work to their customer, workers will reap benefits from high customer satisfaction while coworkers will be forced to tend to the other customers waiting for service.

Debo and his colleagues (2008) also find that workers with discretion over their tasks can engage in behaviors that slow down the service rate. They examine credence services, which are discretionary work settings where service providers have an incentive and the ability to "skim" additional fees from an unknowing customer by providing more services than are actually needed. The classic example is of taxi cab drivers, who can earn a higher fare on a slow night by driving a customer on a longer route than necessary because the customer pays by distance driven rather than a flat rate. They develop a model that shows that under certain price structures and workload dynamics, the provider will skim money from a customer by slowing down the rate of service. In our setting, when patients awaiting service are pooled among a set of physicians, the physician will not earn additional revenue by taking more time to treat a current patient. However, she could benefit from a reduction in the total number of patients she is responsible for seeing during her shift, since any patients still waiting to be seen at the conclusion of her

shift will be seen by the next physician coming on to the shift. This will lead to longer patient stays on average.

Tan and Netessine (2012) examine the impact of workload on worker productivity in a discretionary work setting, although they do not explicitly compare the effect of pooling to that of not pooling. Specifically, they explore wait staff productivity in restaurants with dedicated (non-pooled) tasks, where servers are assigned to a section of tables, to which the host assigns diners as they arrive. Using empirical data from a restaurant chain, they find that employees perform more tasks when the workload is low, and conversely increase their productivity when the workload is higher. This finding suggests that when workers have discretion and are assigned dedicated tasks, they are capable of working at a slower or faster pace when it is beneficial to do so.

We conduct an empirical investigation to further this line of research on the impact of pooling on throughput times in discretionary systems. We extend the argument presented by Hopp et al. (2007) on discretionary task completion systems by examining systems in which it is not only the task completion criteria that is discretionary but also the pace at which workers carry out the work and the resources they utilize. We argue that in discretionary work settings, where multiple elements of the operational process allow for worker discretion, pooling may be associated with an *increase* in throughput times.

> *Hypothesis 1: In discretionary work settings, pooling is associated with longer throughput times compared to when there is no pooling of work.*

Next we distinguish between task pooling and resource pooling to better understand how pooling impacts discretionary work. First, in our examination of task pooling in discretionary work settings, we consider two scenarios: a dedicated task scenario in which customers waiting to enter the organization's production system are assigned to a specific employee as soon as they enter the queue, and a pooled task scenario in which customers are assigned to a specific employee only once they enter the production system. In the dedicated task scenario, employees are effectively designated responsibility for a customer "early", whereas in the pooled task scenario the next available employee is assigned responsibility for a customer when the customer enters the production system. Thus, we examine the impact of shifting the task responsibility boundary upstream such that incoming customers are assigned to specific employees rather than being pooled and able to be processed by anyone.

In contrast to what pooling models in non-discretionary work settings predict, task pooling may have negative implications in discretionary work settings. An important element that has not been previously considered in such models is the aspect of non-random human behavior. Because human workers are influenced by and able to influence the nature and speed of the work in discretionary work settings, this is

an important and predictable factor to consider. In particular, we argue that pooling tasks may negatively impact throughput times in discretionary work settings due to the tendency of individuals to engage in social loafing when tasks are pooled.

The literature on social loafing suggests that having shared responsibility for work (e.g., when tasks are pooled) will result in lower individual effort levels as compared to when individuals have sole responsibility for their work (i.e., are working alone on non-pooled tasks) (Chidambaram & Tung, 2005; Karau & Williams, 1993; Latané, Williams, & Harkins, 1979). This tendency for social loafing comes from multiple individuals having shared responsibility over a common set of tasks. This is distinct from—though similar to—moral hazard, which refers to the tendency for an individual to take risks when the cost of taking the risk is shared among multiple individuals (Arrow, 1963, 1965; Pauly, 1968, 1974; Spence & Zeckhauser, 1971; Zeckhauser, 1970). When applied to settings with discretionary tasks, social loafing theory suggests that a non-pooled task assignment system that designates task responsibilities early on will produce shorter throughput times compared to a pooled task assignment system that does not proactively assign task responsibilities to individuals. This is because workers in the non-pooled task assignment system will be less likely to loaf in such a way that another employee will be forced to pick up the next unit of work that needs to be completed. Thus, in contrast to queuing theory's prediction of a beneficial effect of pooling tasks on throughput times, we hypothesize the following:

> *Hypothesis 2: In discretionary work settings, task pooling is associated with longer throughput times compared to when there are dedicated tasks.*

In contrast to task pooling, resource pooling considers how equipment or other resources necessary to complete a given task are assigned to employees. We again consider two scenarios: a dedicated resource scenario in which a subset of resources is assigned to a specific employee, and a pooled resource scenario in which any employee can use any item in a set of resources that are made jointly available to all employees. In the pooled resource scenario, any one item in a set of resources may be used by any employee, and an employee may draw on as many resources as desired. In contrast, in the dedicated resource scenario, resources are assigned to a specific employee, and employees can only draw from the set of resources that has been assigned to them. Thus, if a resource is idle but the employee to whom the resource has been assigned is busy and cannot utilize it, the resource remains idle, even if another worker wishes to use it. In both scenarios, we relax the assumption of exclusivity and assume that one employee is able to service multiple customers simultaneously. In other words, an employee is not obligated to complete the service of one customer in order to begin servicing the next customer.

In discretionary work settings with multiple workers working alongside one another, resource pooling may result in lower system performance compared to when there are dedicated resources. Given a limited yet shared set of resources, fast-working employees may obtain their speed by engaging in self-serving practices that are suboptimal at the system-level, such as parallel processing work by using multiple resources or over-utilizing resources on non-essential tasks (Shapiro, 1998). While such tactics may speed up their own work times and enhance their customers' perception of their work, fast workers can ultimately reduce coworkers' productivity, either by causing rework or by creating equipment shortages for other workers. This suggests that, with pooled resources, fast-working employees may exhibit a negative effect on their coworkers' productivity by extensively utilizing shared resources and forcing others to wait until the resources become free. This results in longer overall throughput times because fast workers use more than their fair share of resources and create waste by using resources on non-essential tasks. We note that pooling resources may reduce individual throughput times for the fast workers, but we focus on the system-level effect on average throughput times because systems seek to maximize overall productivity. Thus, in contrast to predictions made by models used in non-discretionary work settings, we hypothesize the following:

> *Hypothesis 3: In discretionary work settings, resource pooling is associated with longer throughput times compared to when there are dedicated resources.*

4. Setting, Data, and Empirical Methods

We conducted our study in a hospital's Emergency Department (ED). An ED is a discretionary work setting where tasks and resources can be pooled or dedicated. We refer to the care of a patient from arrival to discharge as a single *task*, which is comprised of numerous sub-tasks that are carried out over the course of a patient's ED visit. The comprehensive set of staff and equipment that supports the physician's work of caring for a patient are considered the available *resources*. For example, the x-ray machines, x-ray technicians, lab technicians, and nurses—all of which are available in limited supply—are considered *resources* that are utilized by physicians to carry out their *task* of caring for patients presenting to the ED.

4.1. Research Setting

Our data come from the ED of a 162-bed community hospital in northern California. This hospital was part of a larger network of 37 hospitals. We selected this ED for study because it experienced an intervention—described in more detail below—that transformed part of the ED from a pooled system to a dedicated system. This intervention enabled us to test hypotheses related to the impact of pooling on

physician productivity. We measured physician performance by their patients' length of stay in the ED, from March 2007 to July 2010. Depending on the time of day, the ED had two to five physicians staffing 41 ED beds and up to nine hallway gurneys. This ED saw approximately 68,937 patients annually, with an average 5 percent increase in census each year from 2007 to 2010. This was a relatively large patient census in comparison to other EDs in the surrounding areas.

There was a standardized patient flow process. Upon a patient's arrival, a registration clerk conducted a brief registration process. A triage nurse then obtained vital signs, collected the chief complaint, and assigned an Emergency Severity Index (ESI) triage category, which is a standard ranking of ED patient severity that ranges from levels 1 (highest acuity) through 5 (lowest acuity). Higher acuity patients (ESI levels 1, 2, or 3) were treated in the main area (main ED). Lower acuity patients (ESI levels 4 or 5) were treated in the Rapid Care Area (RCA), unless they arrived after 11pm or before 7am, in which case they were treated in the main ED because the RCA was closed. Physicians arrived at staggered times throughout the day, such that there was not a certain time at which all physicians changed shifts. Physicians could change shifts on the hour between 5am and 11am, between 2pm and 5pm, and at 11pm or midnight. There was usually one physician working in the RCA and four physicians working in the main ED. Physicians were assigned to a particular location in the ED for the duration of their shift by the physician shift scheduler. Physicians were paid a flat rate for their shift without any additional compensation for the amount of services provided to patients or the number of hours worked. Thus, there were no incentives to stretch out treatment times by providing additional services.

4.2. Intervention: Switching from Pooled to Dedicated Tasks and Resources

In the main ED, an intervention—called the Team Assignment System (TAS)—was implemented in August 2008. TAS effectively restructured the main ED from having pooled tasks and pooled resources to dedicated tasks and dedicated resources. Prior to the TAS intervention, after being triaged, all main ED patients returned to the waiting room until a main ED bed became available. A specific physician was assigned to care for the patient once a patient had been placed in a bed, and not before. At this point, the triage nurse assigned the patient to a physician in a round-robin fashion. Hence, patient assignment was random rather than due to a physician's speed of discharging patients, and physicians were not able to select which patient they wanted to take next. Thus, before the intervention, the tasks (i.e., patients waiting in the waiting room) were pooled as they could end up being served by any of the physicians currently working in the main ED. Before the TAS intervention, physicians also utilized a set of pooled resources (e.g. nurses), which they shared with all other physicians currently working in the main ED. For example, a physician chose one nurse to work with concerning the care of a given patient, but the same

nurse likely was also caring for other patients, who might be assigned to different physicians. Thus, resources were also pooled in the pre-TAS period.

With the implementation of TAS, patients were assigned during triage to a team of one ED physician and two ED nurses who exclusively worked together throughout the duration of their shifts. Thus, after TAS implementation, physicians had dedicated nursing resources and dedicated tasks (i.e., patients). Patients were still assigned to physicians in a round-robin fashion, so patient assignment remained independent of a physician's speed of discharging patients. When a physician logged into the patient management system on one of the ED computers, his or her display clearly showed which patients were assigned to his or her team. Physicians were expected to complete their care for the cohort of patients assigned to them prior to leaving the shift. In other words, they were expected to continue working in the ED until care was complete for every one of their assigned patients, even if their assigned patients were still in the waiting room at the scheduled end of the physician's shift. The change in patient assignment removed the incentive to engage in social loafing because physicians were unable off-load their work onto the oncoming shift of physicians, and because they were paid a flat rate for the shift rather than a variable rate by the number of hours worked.

In the RCA, a single physician worked with pooled tasks and dedicated resources for the entire duration of our study. Patients were assigned to the physician on shift when they were called to be seen in the RCA's examination room, not when they were in the waiting room. In other words, a physician on a shift in the RCA was not responsible for any patient who was still waiting in the waiting room at the conclusion of her shift; any patient still waiting was now the responsibility of the next physician coming on to the shift. In effect, patients were "pooled tasks" waiting to be seen by a physician.

A physician working in the RCA was assigned one nurse (which could be considered the key *resource* for a physician). The physician-nurse team staffing the RCA had access to a designated supply of equipment and medicine set aside for RCA use, such that they did not need to share most of these resources with other physicians concurrently working in the main ED. Figure 1 provides a visual representation of the pooled and dedicated resources and tasks in the main ED and RCA before and after TAS implementation.

---------- Insert Figure 1 About Here ----------

4.3. Data

This study used four years of de-identified electronic medical record (EMR) data of all 243,248 patients treated in the ED from March 1, 2007 to July 31, 2010. The data extracted from the EMR contained the following patient-level information: the patient's time of arrival and departure, length of stay, acuity

level, attending physician, and disposition. We excluded patients with no attending physician or acuity level listed on their record, and patients who had a length of stay of 0 minutes or less. In addition, we excluded patients whose length of stay was greater than 48 hours; most of these patients presented with a psychological condition and were waiting to be discharged to an appropriate facility. We excluded these observations from our dataset because their extended length of stay was driven by placement logistics rather than a physician's level of productivity. Altogether, this resulted in an exclusion of 4,208 patients, which constituted 1.72 percent of the overall sample.

Using the final sample of 239,040 patients, we created a panel dataset that treats the physician as the panel variable. For the regression analysis, we limited our sample to the 234,334 patients who were seen by physicians who were full-time employees of this ED. We chose to limit our sample in this way to be able to construct reliable measures of a physician's permanent productivity, which was necessary to categorize fast workers and, in turn, capture the presence of a fast worker on a given shift. Physicians who worked in this ED but were not full-time employees tended to be employees of other hospitals in the network who were brought in to cover small portions of shifts when the full-time ED physicians were not able to staff the ED (e.g., during physician staff meetings).

4.4. Dependent Variable

To measure throughput time, we used the patient's length of stay in the ED, which was measured in minutes and defined as the time from a patient's arrival to the ED to his or her discharge from the ED. These data were obtained directly from the hospital's EMR system. We log-transformed this variable because the distribution was otherwise right-skewed.

4.5. Independent Variables

4.5.1. Non-pooling intervention in main ED. The implementation of the Team Assignment System (TAS) marked the time at which the main ED transitioned from having pooled tasks and pooled resources to dedicated (or non-pooled) tasks and dedicated resources. We captured this transition with a binary variable that was equal to 1 after the implementation of TAS and 0 before the implementation of TAS. Because it was unknown on exactly what date of August 2008 the TAS system had been implemented, and in order to account for an acclimation period, we omitted data from August 2008 in constructing the variable for TAS. We designated the pre-TAS period to include up to July 31, 2008 and the post-TAS period to begin with September 1, 2008.

4.5.2. Presence of a fast worker. To examine the impact of resource pooling on throughput times, we assessed whether the presence of a fast physician resulted in shorter or longer overall throughput times. We measured the presence of a fast physician as a binary variable. As has been done in previous

studies of fast workers (Mas & Moretti, 2009), we first determined each physician's permanent productivity level by calculating the average length of stay for patients treated by the physician, adjusting for a full set of control variables that include the patient's acuity level, the date and time of admission, and the number of other physicians on that shift, among others (a discussion of each of the control variables is presented in the next subsection). We then categorized physicians whose permanent productivity level was greater than the 50^{th} percentile as "fast physicians". Because we were interested in the effect of the presence of any *other* fast worker, regardless of focal worker's productivity level, we constructed the binary variable such that it equals 1 if there was at least one above-average productivity physician on the shift and 0 if there was no above-average productivity physician on the shift, not accounting for the productivity level of the focal physician. We also considered an alternate specification, in which we categorized physicians with permanent productivity levels above the 25^{th} percentile as "fast physicians". As an additional robustness check, we conducted similar analyses using a continuous measure of having an additional fast physician on a shift.

4.6. Control Variables

We accounted for several other factors that may have affected a patient's length of stay and may be correlated with our independent variables of interest. To account for the variation in length of stay due to the severity of a patient's condition, we controlled for the patient's acuity level (ESI level 1, 2, 3, 4, or 5) using a series of dummy variables. We controlled for the number of patients currently in the ED and the total number of other physicians working at the same time during a physician's shift to capture ED congestion. We also controlled for the general time frame of the physician's shift (AM, PM, or overnight) and the location of the shift (main ED or RCA) to account for systematic differences in patients' length of stay that would arise from differences in structural elements of the ED. To account for a physician's experience, both within and beyond the ED, we controlled for the number of years since graduation from medical school and the number of shifts the physician had worked in this particular ED since the beginning of the dataset up until the point of each patient encounter. Lastly, we accounted for time trends and related influences by adding dummy variables for each day of week, each month, and each calendar year.

4.7. Empirical Models

To test our hypotheses, we used linear regression models with physician fixed effects and clustered standard errors. Standard errors were clustered at the physician level to account for within-physician correlations of the error terms, both within and across shifts, rather than imposing the usual assumption that all error terms are independently and identically distributed. The fixed effects model allowed us to

control for unobservable individual physician effects that do not vary over time, such as level of motivation, innate ability, and practice routines. This is important to account for because they may significantly influence a physician's productivity level in ways that cannot be measured otherwise.

Specifically, we estimated the following models:

$$\ln LOS_{ij} = \beta_0 + \beta_1 fastothers_{ij} + \beta_2 TAS_{ij} + \delta' \mathbf{X}_{ij} + \gamma MD_i + \varepsilon_{ij} \quad (1)$$

$$\ln LOS_{ij} = \phi_0 + \phi_1 fastothers_{ij} + \phi_2 TAS_{ij} + \phi_3 TAS_{ij} \times main_{ij} + \delta' \mathbf{X}_{ij} + \gamma MD_i + \varepsilon_{ij} \quad (2)$$

$$\ln LOS_{ij} = \eta_0 + \eta_1 fastothers_{ij} + \eta_2 TAS_{ij} + \eta_3 TAS_{ij} \times fastothers_{ij} + \delta' \mathbf{X}_{ij} + \gamma MD_i + \varepsilon_{ij} \quad (3)$$

In these models, $\ln LOS_{ij}$ represents the logged number of minutes that patient *i* of physician *j* stayed in the ED; *fastothers* represents the presence of a fast physician on the shift besides the focal physician; *TAS* indicates whether the Team Assignment System (TAS) has been implemented; *TAS* × *main* is an interaction term of whether the TAS system has been implemented and whether the shift was located in the main ED; *TAS* × *fastothers* is an interaction term of whether the TAS system has been implemented and whether a fast physician was present on the shift; **X** is a column vector of covariates; prime (′) denotes transpose; *MD* represents each physician; β's, ϕ's, η's, and δ's represent vectors of coefficients; γ represents physician fixed effects; and ε is the time-varying error term not already captured by γ. The column vector of covariates, **X**, includes all control variables described in the previous subsection, which includes year fixed effects to control for time trends as well as the main effect for the shift location indicator variable, *main*. Table 1 provides summary definitions for all variables included in these models.

---------- Insert Table 1 About Here ----------

To test Hypothesis 1, we estimate Model 1 using the data from the main ED. We exclude data from the RCA because there was no change in work structure in the ED throughout the study period, whereas the main ED moved from having a pooled system to a dedicated system. We estimate the overall impact of this transition from having a pooled system to a dedicated system in the main ED by examining β_2, the coefficient on *TAS*. We predict that β_2 is negative and significant.

To test Hypothesis 2, we estimate Model 2 using data from both the main ED and the RCA. We employ a difference-in-differences estimator to compare the difference in average throughput times in the two locations before TAS implementation to the difference after TAS implementation. Because the RCA remained under a system in which tasks were pooled, whereas the main ED moved from having pooled

tasks to dedicated tasks, we can consider the shifts worked in the RCA as comprising the untreated comparison group and those worked in the main ED as comprising the treatment group. To apply the difference-in-differences method, we first establish the parallel trend assumption and calculate autocorrelation-consistent standard errors (Abadie, 2005; Duflo, 2001). We then estimate the effect of task pooling on throughput times by examining the coefficient on the interaction term, *TAS* × *main*. We predict that this coefficient, ϕ_3, will be negative and significant, suggesting that having dedicated tasks is associated with shorter throughput times and having pooled tasks is associated with longer throughput times.

Lastly, to test Hypothesis 3, we estimate Model 3 using data from the main ED. To investigate the effect of pooling resources on throughput times, we leverage the presence of other workers who are fast, who may increase overall throughput times under a system with pooled resources. For this analysis, we exclude data from the RCA because there is only one physician on a given shift, and therefore no possibility of being influenced by another worker in the RCA who is fast. Using the effect of fast coworkers as a proxy, we estimate the effect of moving from having pooled resources to dedicated resources on throughput times by examining η_3, the coefficient on *TAS* × *fastothers*. We predict that η_3 is negative and significant.

In addition to the standard assumptions of linear regression models, fixed effects models make two key assumptions, both of which are satisfied in our study. First is the assumption of strict exogeneity, which means that the error term of the model is uncorrelated with each of the covariates in all time periods (Wooldridge, 2010). This is a plausible assumption to make in our context, because the patient error term is unlikely to be correlated with the covariates for other patients. In addition, the random traits of patients that affect their length of stay are not likely to be associated with the key independent variables of interest. Specifically, the round-robin assignment of patients to physicians makes it unlikely for the fastest physicians to receive the most complicated cases. In other words, patient assignment to physicians is random and is not driven by physician speed or physician preference. Furthermore, because we are within the context of an emergency room, in which incoming patient complexity is hard to predict ahead of time, it is unlikely that the best performing physicians would be assigned to shifts with the most challenging cases. In addition, we do not expect that there were changes in patient traits over time that were differentially affecting patients seen in the main ED and the RCA.

We choose to use fixed effects models rather than random effects models because we do not believe that the random effects assumption of zero correlation between the physician effect and the covariates (such as the number of shifts worked by the physician, or the number of years since graduating from medical school) would necessarily hold. By using fixed effects models, we are able to account for the unobserved traits of each physician that are associated with a patient's length of stay that are also

correlated with the independent variables of interest. Accordingly, we conducted the Durbin-Wu-Hausman test, which rejected the random effects models in favor of the fixed effects models ($\chi^2 > 68.52$, $p < 0.001$ for all three models).

5. Results

5.1. Descriptive Statistics

Table 2 presents means, standard deviations, and correlations for all continuous variables and percentages for all categorical or binary variables included in the models. In particular, we find that the average length of stay for a patient seen in this ED was 220 minutes (*s.d.* = 243). Specifically, the average length of stay was 277 minutes (*s.d.* = 271) for a patient seen in the main ED and 106 minutes (*s.d.* = 105) for a patient seen in the RCA. An average ED physician had almost 13 years of experience working as a physician, and worked on average 276 shifts (*s.d.* = 166) in this ED between March 1, 2007 and July 31, 2010. On average, 35 patients (*s.d.* = 12) were in this ED at any given time.

---------- Insert Table 2 About Here ----------

Less than 1 percent of patients were of acuity level 1, whereas 51 percent were of acuity level 3 and 39 percent were of acuity level 4. Patients were approximately uniformly distributed across days of the week and months of the year. An increasing number of patients presented to this ED in recent years, with approximately half arriving during the PM shift and another 13 percent arriving during the overnight shift. Approximately 67 percent of patients were seen in the main ED and 86 percent of patients were seen while his or her attending physician had an above-average productivity physician on his or her shift.

As expected, patients' average length of stay differed significantly by their level of acuity. Table 3 presents the means and standard deviations of patients' average length of stay by acuity level, as well as the frequencies of each acuity level. We note that the relationship between the length of stay and acuity level was not a completely monotonic one. While there was a general monotonic trend in which patients of a lower acuity level (e.g., acuity level 5) had a shorter length of stay, patients of acuity level 2 had a significantly longer length of stay than those of acuity level 1. Based on observations at our study site and conversations with ED personnel, this appeared to be attributable to three reasons. First, a large fraction of acuity level 1 patients were too severely ill to be treated or resuscitated. This resulted in a truncating effect in which several acuity level 1 patients had relatively short lengths of stay due to death. Second, due to the severity of their illness, many acuity level 1 patients were quickly admitted to other departments (e.g., the Intensive Care Unit) or were taken to the Operating Room. This also had a truncating effect on these patients' lengths of stay. Third, most patients presenting with a psychological

condition were acuity level 2, and these patients often had longer stays in the ED due to difficulties in finding an appropriate facility to which they could be discharged. These qualitative statements seemed to be supported by our data, given the large variation in length of stay of acuity level 1 patients (*s.d.* = 211) and the fact that only 0.33 percent of all patients (879 patients over a period of five years) fell into this category. Because we adjust for patient acuity using a dummy variable for each acuity level, the non-linearity of the relationship between acuity level and length of stay does not meaningfully affect our analysis.

---------- Insert Table 3 About Here ----------

5.2. Tests of Hypotheses

We estimated the three models specified in the previous section to test each of our hypotheses. The results of our analyses are summarized in Table 4.

---------- Insert Table 4 About Here ----------

We start by examining the overall effect of pooling on throughput times. Model 1 of Table 4 presents a fixed effects model that includes all control variables and captures the effect of moving from a pooled system to a dedicated system (i.e., the negative of the pooling effect). The negative coefficient on *TAS* indicates that the transition from a pooled system to a dedicated system is associated with a highly significant reduction on patients' lengths of stay ($\beta_2 = -0.09$, $p < 0.001$). This 8.86 percent decrease in a patient's length of stay after the implementation of TAS corresponds to a decrease of 25.1 minutes for an average patient of acuity level 3 seen in the main ED. In other words, the average patient's length of stay was significantly longer in the pooled system than in the dedicated system. This offers strong support for H1, which predicted that pooling is associated with longer throughput times in discretionary work settings.

Model 2 of Table 4 adds an interaction term of TAS implementation and the shift's location. We find that the difference in throughput times between the main ED and the RCA is greater prior to TAS implementation when tasks were pooled in both the main ED and the RCA. Once the main ED adopted a dedicated task system, this difference in throughput times reduced. This difference-in-differences is captured by the coefficient on $TAS \times main$ ($\phi_3 = -0.10$, $p < 0.001$). This 9.90 percent decrease in a patient's length of stay in the main ED after TAS implementation corresponds to a decrease of 28 minutes for an average patient of acuity level 3 presenting to the main ED. This finding offers strong support for

H2, which predicted that, in discretionary work settings, task pooling is associated with longer throughput times compared to when tasks are dedicated.

Model 3 of Table 4 substitutes the interaction term of TAS implementation and the shift's location with the interaction term of TAS implementation and the presence of other fast physicians. We find that when resources are pooled in the main ED, the presence of fast coworkers on one's shift is associated with a significantly longer length of stay for the average patient. This suggests that, under a system with pooled resources, fast physicians (who tend to draw on a greater amount of shared resources) negatively impact the productivity of their coworkers. This effect is captured by the coefficient on *TAS × fastothers* ($\eta_3 = -0.08$, $p < 0.001$), which shows that the transition from having pooled resources to having dedicated resources in the main ED was associated with a 7.75 percent decrease in an average patient's length of stay. This corresponds to a 21.9-minute decrease that occurred when" the main ED moved to a dedicated resource system. This finding supports H3, which predicted that, in discretionary work settings, resource pooling is associated with longer throughput times compared to when resources are dedicated.

5.3. Specification Tests

To examine the robustness of our models, we tested a variety of other specifications in addition to the reported models. Specifically, we examined an alternative specification of the *fastothers* variable that was used in our models to indicate the presence of a fast physician on a focal physician's shift. We used an indicator variable for an above-25th percentile productivity physician as opposed to an above-average (50th percentile) productivity physician. With this alternative specification, we found that the transition from a pooled system to a dedicated system resulted in an 8.92 percent decrease in a patient's length of stay ($\beta_2 = -0.09$, $p < 0.001$). We found that the transition from pooled to dedicated tasks was associated with a 9.92 percent decrease in length of stay ($\phi_3 = -0.10$, $p < 0.001$), and that the transition from pooled to dedicated resources was associated with a 3.53 percent decrease in the same measure ($\eta_3 = -0.04$, $p < 0.001$). Specifically regarding the last measure, on the effect of transitioning from pooled to dedicated resources, we note that the magnitude of the effect is smaller than it was in the original specification. This is not surprising because the alternate specification resulted in a stricter definition of who is considered a "fast physician", and it is to be expected that the overall negative effect of fast workers under a pooled resource system is dampened when there are fewer fast workers. Thus, our findings for all three hypotheses remained robust to this alternate specification.

Our results do not appear to be due to the different types of patients that are typically cared for in the two areas of the ED. To examine this, we assessed whether the transition from a pooled system to a dedicated system differentially affected length of stay depending on a patient's acuity level. To conduct this analysis, we used the same models as those specified above, but limited to patients of acuity levels 4

and 5, and with each independent variable of interest interacted with acuity level 5. We limited the sample to patients of these two acuity levels because they constituted the group of patients who were potentially seen in both areas of the ED (because all acuity 4 and 5 patients were seen in the main ED after 11pm). This analysis suggests that there are no differential effects by patient acuity level (the *p*-values of all independent variables of interest interacted with acuity level 5 ranged from 0.67 to 0.70).

We also included all observations that had previously been excluded as outliers (i.e., observations with a length of stay greater than 48 hours). All coefficients of interest and their corresponding significance levels remained robust to this specification.

Lastly, we used hierarchical linear models, which specify random rather than fixed effects at the physician level. This specification test was conducted in order to test each of our hypotheses with greater efficiency gains. Here, we used three levels for our multilevel analyses: patient-level, physician shift-level, and physician-level. The effect of transitioning from a pooled to a dedicated system remained robust ($\beta_2 = -0.09$, $p < 0.001$), as did the effect of moving from pooled to dedicated tasks ($\phi_3 = -0.10$, $p < 0.001$) and resources ($\eta_3 = -0.07$, $p < 0.001$). Thus, the magnitude and significance of all coefficients of interest remained robust to this alternative specification.

6. Discussion and Conclusions

Using EMR data from a hospital's ED over four years, we find that counter to what queuing theory would predict, pooling may increase throughput times in discretionary work settings. More specifically, we find that patients have longer lengths of stay when ED physicians work in systems with pooled tasks and resources, as opposed to when they work in systems with dedicated tasks and resources. We assert that the improved performance with dedicated tasks comes from a reduction in social loafing, which is largely left unchecked when tasks are pooled for processing by any physician rather than assigned to an individual physician. The improved performance with dedicated resources comes from a more distributed utilization of shared resources, in which fast-working physicians are prevented from over-utilizing shared resources that are necessary to conduct process tasks.

In the context of our study setting, we find it particularly important to consider the significance of the effect sizes found in our analyses. For example, in Model 1 of Table 4, we find that moving from a pooled system to a dedicated system is associated with an 8.86 percent decrease in a patient's length of stay. For an average patient of acuity level 3 seen in the main ED, this corresponds to a decrease of 25.1 minutes. This is a particularly meaningful difference in the context of a hospital's emergency room. With approximately 200 patients in the ED every day, this is roughly equivalent to an additional 83.7 patient-

hours per day that were saved with the dedicated system. Once we take into account the large costs associated with emergency room care, it becomes clear that the time and cost implications are substantial.

We note that, unlike much prior work on the impact of pooling, we focus on throughput times as opposed to specifically waiting times. This was not possible with our empirical data, because it was not possible to determine at what stages and how much time a patient spent waiting once entering a RCA examination room or being placed in a main ED bed, thus making it difficult to accurately distinguish processing time from waiting time. Future research should consider separately measuring the effect of pooling on processing times versus waiting times.

6.1. Theoretical Contributions

This paper contributes to the operations management literature on pooling in several ways. We use empirical data to examine the effect of pooling on throughput time in a *discretionary* work setting. We find that when workers have discretionary control over the pace and resources involved in carrying out their work and no incentive for speeding up their work, pooling results in *longer* throughput times. Queuing theory about non-discretionary work settings would suggest instead that pooling would result in *shorter* throughput times (Jouini et al., 2008). Thus, our paper provides empirical support for prior mathematical models that predicted that human behaviors could reduce the positive benefit of pooling (Hopp et al., 2007; Jouini et al., 2008).

Our result is related to Debo et al.'s (2008) finding that service providers who have discretion over the set of tasks performed for customers may increase the time required to deliver service if they have an incentive to do so. We find empirical evidence to support their claim, although in our setting it is not a financial incentive, but rather a social loafing incentive for slowing down the service rate that triggers this behavior. In Debo et al.'s (2008) study, the customer's lack of visibility into the true level of services that should be provided enables workers to increase throughput time, whereas in our setting—similar to Jouini et al. (2008)—the lack of management visibility into work pace enables longer throughput times to exist. Our study suggests that managers of discretionary work systems should design control mechanisms to mitigate behaviors that benefit the employee to the detriment of the customers or the organization. We find that one mechanism is to make the workload constant regardless of work pace, which removes the benefit of slowing down. Future research could examine whether this arrangement creates other problems, such as a speeding up that results in lower quality.

Unlike prior research in this domain (Corbett & Rajaram, 2006; Debo et al., 2008; Jouini et al., 2008), we use empirical data to study how task pooling and resource pooling affect the productivity of workers who engage in discretionary work. In doing so, we are able to quantify the effects of pooling in an actual working environment. In our context of a hospital ED, we find that having a pooled system, as opposed to

a dedicated system, is associated with a 9 percent increase in patients' length of stay. In addition, our paper answers the call for additional research on the interaction of human behavior and operations management (Boudreau et al., 2003; Jouini et al., 2008). Specifically, we consider how individual tendencies to engage in social loafing and over-utilization of shared resources may undermine the potential benefits of pooling. We explore this by distinguishing the effects of task pooling and resource pooling. Our analyses suggest that each of these types of pooling has an effect. As hypothesized, task pooling results in longer throughput times because physicians engage in social loafing when they perceive that slowing down the task completion process will increase the likelihood that some other physician will assume responsibility for the next patient. Similarly, resource pooling lengthens throughput times because fast-working physicians utilize a disproportionately large share of resources to the detriment of their colleagues working on the same shift.

6.2. Limitations

This study has limitations, and its results should be interpreted accordingly. First, we note the threat of omitted variable bias that is common to many empirical models. While it would have been helpful to be able to include more patient characteristics in our specification, such as patient diagnosis or medical comorbidities, these data were protected information and not available for use. However, this is not an important threat because patients were randomly assigned to physicians rather than selected by them. This is supported by the fact that the average acuity level of patients seen by each physician was less than one standard deviation away from the average acuity level of all patients seen in the ED (*mean* = 3.32, *s.d.* = 0.66).

Second, our study is limited to one hospital's ED and its response to one intervention that transitioned the work system from having pooled tasks and resources to dedicated tasks and resources. This may limit the generalizability of our findings, though we believe our findings have strong theoretical underpinnings. Nevertheless, we welcome future research on these effects and mechanisms in different empirical contexts for further substantiation.

Third, while it is beyond the scope of this paper to examine the effects of increased productivity, decreased throughput times, or dedicated work systems on quality, we acknowledge that this is an important element to consider in future research. In this paper, we were not able to extend our analysis to include the effects on quality because we did not have data on patients' clinical outcomes. Though this may be a particularly important consideration in the context of health care delivery—decreased lengths of stay at the expense of health care quality would not be desirable—we argue that this consideration should be extended to other contexts as well. We note that prior research has begun to explore these relationships between productivity and product quality (Krishnan, Kriebel, Kekre, & Mukhopadhyay, 2000).

6.3. Practical Implications and Conclusions

Our research also offers valuable practical insights for workplace managers and health care policy makers. First, our findings suggest that, in workplaces where workers have discretionary control, the potential negative effects of designing pooled systems must be carefully considered. This has implications for designing and managing staffing structures and workflows, particularly in the context of service delivery organizations.

Second, our findings suggest that managers should consider implementing group incentives rather than individual incentives to motivate workers. This may encourage fast workers to reduce their speed just enough so that they will not negatively affect the productivity of others by over-utilizing shared resources. While workplaces often seek to incentivize workers through pay-for-performance programs that focus on individual productivity, we refer to prior research that compares individual versus collective incentives to suggest that a group-level approach may help counteract the negative effects that fast workers exhibit on overall productivity levels (Arya & Mittendorf, 2011; Bandiera, Barankay, & Rasul, 2011).

Specifically in the context of health care, our findings suggest that EDs may benefit from implementing non-pooled work systems in which patients are assigned to a doctor-nurse team immediately upon arrival. While results may differ across different settings and various information technology systems, the mechanisms through which changes in throughput time occurred may help shed light on cost savings predictions in other contexts. This could have significant implications for health care delivery, especially given the expected increase in ED patient volume as a result of the recent health reform legislation (*Patient Protection and Affordable Care Act*, 2010).

References

Abadie, A. (2005). Semiparametric Difference-in-Differences Estimators. *Review of Economic Studies*, *72*(1), 1–19. doi:10.1111/0034-6527.00321

Alfaro, J. A., & Corbett, C. J. (2009). The Value of SKU Rationalization in Practice (The Pooling Effect Under Suboptimal Inventory Policies and Nonnormal Demand). *Production and Operations Management*, *12*(1), 12–29. doi:10.1111/j.1937-5956.2003.tb00195.x

Alptekinoglu, A., Banerjee, A., Paul, A., & Jain, N. (2012). Inventory Pooling to Deliver Differentiated Service. *Manufacturing & Service Operations Management*, *15*(1), 33–44. doi:10.1287/msom.1120.0399

Anupindi, R., Chopra, S., Deshmukh, S. D., Van Mieghem, J. A., & Zemel, E. (2005). *Managing Business Process Flows: Principles of Operations Management* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review*, *53*(5), 941–973.

Arrow, K. J. (1965). *Aspects of the theory of risk-bearing*. Helsinki: Yrjö Jahnssonin Säätiö.

Arya, A., & Mittendorf, B. (2011). The Benefits of Aggregate Performance Metrics in the Presence of Career Concerns. *Management Science*, *57*(8), 1424–1437. doi:10.1287/mnsc.1110.1363

Ata, B., & Van Mieghem, J. A. (2008). The Value of Partial Resource Pooling: Should a Service Network Be Integrated or Product-Focused? *Management Science*, *55*(1), 115–131. doi:10.1287/mnsc.1080.0918

Bandiera, O., Barankay, I., & Rasul, I. (2011). Field Experiments with Firms. *Journal of Economic Perspectives*, *25*(3), 63–82. doi:10.1257/jep.25.3.63

Benjaafar, S., Cooper, W. L., & Kim, J.-S. (2005). On the Benefits of Pooling in Production-Inventory Systems. *Management Science*, *51*(4), 548–565. doi:10.1287/mnsc.1040.0303

Boudreau, J., Hopp, W. J., McClain, J. O., & Thomas, L. J. (2003). On the Interface Between Operations and Human Resources Management. *Manufacturing & Service Operations Management*, *5*(3), 179–202. doi:10.1287/msom.5.3.179.16032

Chidambaram, L., & Tung, L. L. (2005). Is Out of Sight, Out of Mind? An Empirical Study of Social Loafing in Technology-Supported Groups. *Information Systems Research*, *16*(2), 149–168. doi:10.1287/isre.1050.0051

Corbett, C. J., & Rajaram, K. (2006). A Generalization of the Inventory Pooling Effect to Nonnormal Dependent Demand. *Manufacturing & Service Operations Management*, *8*(4), 351–358. doi:10.1287/msom.1060.0117

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

Duflo, E. (2001). Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment. *American Economic Review*, *91*(4), 795–813. doi:10.1257/aer.91.4.795

Eppen, G. D. (1979). Note - Effects of Centralization on Expected Costs in a Multi-Location Newsboy Problem. *Management Science*, *25*(5), 498–501. doi:10.1287/mnsc.25.5.498

Flynn, B. B., Sakakibara, S., Schroeder, R. G., Bates, K. A., & Flynn, E. J. (1990). Empirical research methods in operations management. *Journal of Operations Management*, *9*(2), 250–284. doi:10.1016/0272-6963(90)90098-X

Gans, N., Koole, G., & Mandelbaum, A. (2003). Commissioned Paper: Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing & Service Operations Management*, *5*(2), 79–141. doi:10.1287/msom.5.2.79.16071

Gerchak, Y., & He, Q. (2003). On the Relation Between the Benefits of Risk Pooling and the Variability of Demand. *IIE Transactions*, *35*(11), 1027–1031. doi:10.1080/07408170304399

Graves, S. C., & Tomlin, B. T. (2003). Process Flexibility in Supply Chains. *Management Science*, *49*(7), 907–919. doi:10.1287/mnsc.49.7.907.16381

Gupta, S., Verma, R., & Victorino, L. (2009). 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

Hopp, W. J., Iravani, S. M. R., & Liu, F. (2009). Managing White-Collar Work: An Operations-Oriented Survey. *Production and Operations Management*, *18*(1), 1–32. doi:10.1111/j.1937-5956.2009.01002.x

Hopp, W. J., Iravani, S. M. R., & Yuen, G. Y. (2007). Operations Systems with Discretionary Task Completion. *Management Science*, *53*(1), 61–77. doi:10.1287/mnsc.1060.0598

Jouini, O., Dallery, Y., & Nait-Abdallah, R. (2008). Analysis of the Impact of Team-Based Organizations in Call Center Management. *Management Science*, *54*(2), 400–414. doi:10.1287/mnsc.1070.0822

Karau, S., & Williams, K. (1993). Social Loafing : A Meta-Analytic Review and Theoretical Integration. *Journal of Personality and Social Psychology*, *65*(4), 681–706.

Krishnan, M. S., Kriebel, C. H., Kekre, S., & Mukhopadhyay, T. (2000). An Empirical Analysis of Productivity and Quality in Software Products. *Management Science*, *46*(6), 745–759. doi:10.1287/mnsc.46.6.745.11941

Latané, 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), 822–832. doi:10.1037/0022-3514.37.6.822

Loch, C. (1998). Operations management and reengineering. *European Management Journal*, *16*(3), 306–317. doi:10.1016/S0263-2373(98)00007-3

Mandelbaum, A., & Reiman, M. I. (1998). On Pooling in Queueing Networks. *Management Science*, *44*(7), 971–981. doi:10.1287/mnsc.44.7.971

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

Meredith, J. (1998). Building operations management theory through case and field research. *Journal of Operations Management*, *16*(4), 441–454. doi:10.1016/S0272-6963(98)00023-0

Patient Protection and Affordable Care Act, Pub. L. No. 124 Stat. 119 through 124 Stat. 1025 (2010).

Pauly, M. V. (1968). The Economics of Moral Hazard: Comment. *American Economic Review*, *58*, 531–536. doi:10.1017/CBO9780511528248.010

Pauly, M. V. (1974). Overinsurance and Public Provision of Insurance: The Roles of Moral Hazard and Adverse Selection. *The Quarterly Journal of Economics*, *88*(1), 44. doi:10.2307/1881793

Roth, A. V., & Menor, L. J. (2009). Insights into Service Operations Management: A Research Agenda. *Production and Operations Management*, *12*(2), 145–164. doi:10.1111/j.1937-5956.2003.tb00498.x

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. doi:10.1016/S0272-6963(97)00008-9

Shapiro, R. D. (1998). Donner Co. *Harvard Business School Case*, *689-030*.

Smith, D. R., & Whitt, W. (1981). Resource Sharing for Efficiency in Traffic Systems. *The Bell System Technical Journal*, *60*(1), 39–56.

Spence, M., & Zeckhauser, R. J. (1971). Insurance, Information, and Individual Action. *American Economic Review*, *61*(2), 380–387.

Stidham, S. (1970). On the Optimality of Single-Server Queuing Systems. *Operations Research*, *18*(4), 708–732. doi:10.1287/opre.18.4.708

Tagaras, G., & Cohen, M. A. (1992). Pooling in Two-Location Inventory Systems with Non-Negligible Replenishment Lead Times. *Management Science*, *38*(8), 1067–1083. doi:10.1287/mnsc.38.8.1067

Tan, T., & Netessine, S. (2012). When Does the Devil Make Work? An Empirical Study of the Impact of Workload on Worker Productivity. *SSRN Electronic Journal*. doi:10.2139/ssrn.2071113

Van Dijk, N. M. (2002). To pool or not to pool? The benefits of combining queuing and simulation. *Proceedings of the Winter Simulation Conference* (Vol. 2, pp. 1469–1472). IEEE. doi:10.1109/WSC.2002.1166421

Wallace, R. B., & Whitt, W. (2005). A Staffing Algorithm for Call Centers with Skill-Based Routing. *Manufacturing & Service Operations Management*, *7*(4), 276–294. doi:10.1287/msom.1050.0086

Whitt, W. (1992). Understanding the Efficiency of Multi-Server Service Systems. *Management Science*, *38*(5), 708–723. doi:10.1287/mnsc.38.5.708

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data: Second Edition* (2nd ed.). Cambridge, Massachusetts: MIT Press.

Zeckhauser, R. J. (1970). Medical Insurance: A Case Study of the Tradeoff between Risk Spreading and Appropriate Incentives. *Journal of Economic Theory*, *2*, 10–26.

Figure 1. Diagram of the Pre- and Post-TAS Task and Resource Pooling Structure

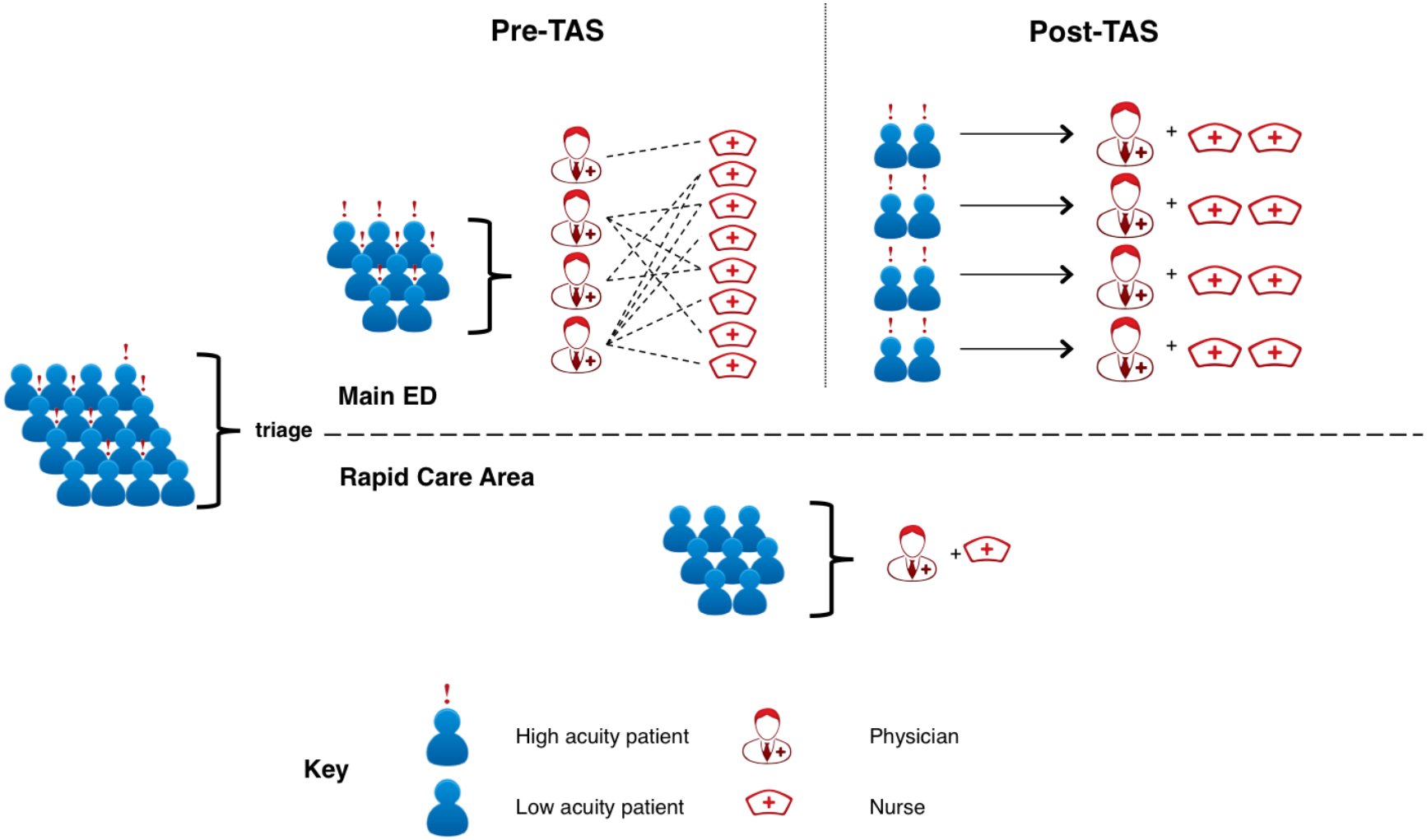

Table 1. Summary definition of variables included in fixed effects models

Variable / Category	Description
Length of stay	Logged number of minutes for which patients stayed in ED.
Acuity level	5 indicators for patient's acuity level (1, 2, 3, 4, 5).
No. of other MDs on shift	Number of other physicians working at the same time as this shift.
Shift number	Indicator for what number shift this is for this physician in this dataset (since January 1, 2007).
Years since graduation	Number of years since graduation from medical school.
Day of week	7 indicators for day of week of shift.
Month	12 indicators for month of shift.
Year	4 indicators for year of shift.
Shift type	3 indicators for shift type (AM, PM, overnight).
Current patient count	Number of total patients currently in the ED.
Main ED	Shift location (= 1 for Main ED, = 0 for RCA).
Fast others	Indicator for presence of at least one above-average productivity worker on shift (not including oneself) (= 1 for present, = 0 for absent).
TAS implemented	Indicator for whether TAS had been implemented (= 1 for post-implementation, = 0 for pre-implementation).
Interactions	TAS × Main ED. TAS × Fast others.

Table 2. Summary statistics of variables included in fixed effects models[a]

Variable	Mean (SD)	Min.	Max.	1	2	3	4	5	6
1. Length of stay	220.03 (243.34)	1	2879	1.00					
2. Logged length of stay	5.02 (0.83)	0	7.97	0.83*	1.00				
3. Other MDs on shift	7.88 (2.51)	1	21	-0.01*	-0.02*	1.00			
4. Shift number	276.07 (165.94)	1	710	-0.08*	-0.10*	0.23*	1.00		
5. Years since graduation	12.65 (7.92)	2	37	0.01*	0.01*	0.05*	0.11*	1.00	
6. Current patient count	34.52 (11.72)	1	84	0.06*	0.12*	0.27*	0.09*	-0.01*	1.10

Variable	Percent	Variable	Percent
Acuity level 1	0.30	January	7.40
Acuity level 2	8.48	February	7.36[b]
Acuity level 3	50.96	March	10.34[b]
Acuity level 4	38.85	April	9.82[b]
Acuity level 5	1.41	May	10.47[b]
Sunday	15.00	June	9.67[b]
Monday	15.03	July	9.99[b]
Tuesday	14.04	August	7.27
Wednesday	13.60	September	7.13
Thursday	13.76	October	7.06
Friday	13.82	November	6.74
Saturday	14.75	December	6.75
AM shift	39.57	2007	20.23
PM shift	47.67	2008	24.83
Overnight shift	12.76	2009	27.39
Main ED	66.74	2010	27.55
Other fast MDs on shift	85.84		
TAS implemented	59.31		

* $p < 0.05$

[a] N = 234,334. Excludes observations earlier than March 1, 2007 and after July 31, 2010.

[b] Because all observations earlier than March 1, 2007 and after July 31, 2010 have been excluded, it is not surprising that a larger percentage of patients in our dataset presented to the ED in the months between March and July, inclusive. When these summary statistics are produced with the inclusion of observations all from January 1, 2007 to December 31, 2010, we obtain an approximately uniform distribution of patients across all months of the year.

Table 3. Average length of stay in minutes by patient acuity level[a]

Acuity level	Mean	s.d.	Frequency
1 (most severe)	287.17	216.73	710
2	391.47	352.94	19861
3	279.69	262.83	119424
4	108.76	98.77	91029
5 (least severe)	84.29	95.26	3310

[a] N = 234,334.

Table 4. Fixed effects models for patients' logged length of stay

Variables	Model 1	Model 2	Model 3
Acuity level 2	0.257***	0.239***	0.257***
	(0.0346)	(0.0356)	(0.0347)
Acuity level 3	-0.0712	-0.0937*	-0.0709
	(0.0402)	(0.0410)	(0.0402)
Acuity level 4	-0.692***	-0.706***	-0.692***
	(0.0397)	(0.0388)	(0.0397)
Acuity level 5	-1.050***	-0.944***	-1.050***
	(0.0388)	(0.0393)	(0.0389)
No. of other MDs on shift	0.00448***	-0.00275*	0.00476***
	(0.000919)	(0.00136)	(0.000913)
Shift number	-0.000620***	-0.000274	-0.000599***
	(0.000167)	(0.000182)	(0.000164)
Years since graduation	0.0352	-0.00542	0.0324
	(0.0263)	(0.0287)	(0.0256)
PM shift	-0.128***	-0.0862***	-0.129***
	(0.00863)	(0.0157)	(0.00857)
Overnight shift	-0.0356**	0.00272	-0.0356**
	(0.0119)	(0.0208)	(0.0119)
Current patient count	0.00934***	0.0136***	0.00932***
	(0.000362)	(0.000942)	(0.000361)
Main ED		0.370***	
		(0.0226)	
Other fast MDs on shift	0.0212*	0.0180*	0.0724***
	(0.0105)	(0.00878)	(0.0151)
TAS	-0.0886***	-0.0157	-0.0228
	(0.0149)	(0.0165)	(0.0177)
TAS × Main ED		-0.0990***	
		(0.0150)	
TAS × Other fast MDs on shift			-0.0775***
			(0.0165)
Day of week controls	Yes	Yes	Yes
Month controls	Yes	Yes	Yes
Year controls	Yes	Yes	Yes
Constant	4.910***	4.891***	4.895***
	(0.324)	(0.348)	(0.317)
Observations	152,841	228,935	152,841
Adjusted R^2	0.154	0.327	0.155
Number of ED physicians	40	40	40

Robust standard errors in parentheses
*** p <0.001, ** p <0.01, * p <0.05