Management Science

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

When Does the Devil Make Work? An Empirical Study of the Impact of Workload on Worker Productivity

Tom Fangyun Tan, Serguei Netessine

To cite this article:
Tom Fangyun Tan, Serguei Netessine (2014) When Does the Devil Make Work? An Empirical Study of the Impact of Workload on Worker Productivity. Management Science 60(6):1574-1593. http://dx.doi.org/10.1287/mnsc.2014.1950

MANAGEMENT SCIENCE
Vol. 60, No. 6, June 2014, pp. 1574–1593
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2014.1950

When Does the Devil Make Work? An Empirical Study of the Impact of Workload on Worker Productivity

Tom Fangyun Tan
Cox Business School, Southern Methodist University, Dallas, Texas 75275, ttan@cox.smu.edu

Serguei Netessine
INSEAD, Singapore 138676, serguei.netessine@insead.edu

We analyze a large, detailed operational data set from a restaurant chain to shed new light on how workload (defined as the number of tables or diners that a server simultaneously handles) affects servers' performance (measured as sales and meal duration). We use an exogenous shock—the implementation of labor scheduling software—and time-lagged instrumental variables to disentangle the endogeneity between demand and supply in this setting. We show that servers strive to maximize sales and speed efforts simultaneously, depending on the relative values of sales and speed. As a result, we find that, when the overall workload is small, servers expend more and more sales efforts with the increase in workload at a cost of slower service speed. However, above a certain workload threshold, servers start to reduce their sales efforts and work more promptly with the further rise in workload. In the focal restaurant chain, we find that this saturation point is currently not reached and, counterintuitively, the chain can reduce the staffing level and achieve both significantly higher sales (an estimated 3% increase) and lower labor costs (an estimated 17% decrease).

Keywords: econometrics; empirical study on staffing; worker productivity; business analytics; restaurant operations; behavioral operations management; quality/speed trade-off
History: Received May 30, 2012; accepted January 25, 2014, by Noah Gans, special issue on business analytics. Published online in *Articles in Advance* April 21, 2014.

1. Introduction

In many service organizations such as retail stores, call centers, and restaurants, labor costs often reach 60%–70% of operating expenses, making it one of the largest cost components. To control these large and stubbornly increasing costs, companies strive to reduce staffing levels, which often risks overwork and leads to high turnover typical in the above industries (e.g., Whitt 2006). At the same time, overstaffing may cause inefficiencies, so identifying the right staffing level is crucial to achieving optimal operational performance. Not surprisingly, many service companies are increasingly utilizing computerized staffing tools (Maher 2007) in order to optimize the deployment of the expensive labor force. As a result, staffing policies have also received wide interest in the academic research, with most of the literature focusing on a now traditional engineering approach to design an optimal staffing policy: the demand for labor is calculated and translated into labor requirements using optimization algorithms.

However, in most of these staffing tools and academic research, employee productivity is typically calculated using "grand averages" over some time intervals (as reflected, for instance, in a call center survey (Gans et al. 2003)), thus overlooking employees' adaptive behavior toward changing work environments. There is, however, growing evidence that behavioral aspects of workers must be taken into account. For instance, Brown et al. (2005) found several anomalies suggesting that some behavioral aspects of labor management may lead to serious staffing errors. Despite these observations, there is generally a lack of empirical research about the specific mechanisms through which staffing affects worker performance (Akşin et al. 2007). Our paper aims to both fill this void in the literature and to aid in practical decision making by demonstrating usefulness of the data analytic approach to labor management.

Recent efforts have bridged operations management (OM) models and human resource management (HRM) to study the impact of external factors on individuals' performance (Boudreau 2004; refer to §2 for more discussion). Closer to the question posed in this study, several researchers have recently turned to understanding the impact of workload, an integral environmental factor, on individual performance. Some of them focused on the impact of workload on service time, and other studies separately examined the impact

of workload on service quality.[1] However, in practice, service time and service quality are both critically important to the firm, and there is often a trade-off between them. For instance, a call center employee can spend more time on the call to increase quality of service or to reduce service time in order to service another client at the expense of service quality. A growing number of papers are starting to use analytical modeling approaches to yield insights into how workers make such trade-off decisions, often termed the quality–speed conundrum (e.g., Hopp et al. 2007, Anand et al. 2011; see §2). Understanding how workers react to workload in terms of both service time and quality in practice is therefore of great significance in the service industry, where service providers aim to simultaneously maximize service quality, which relates to revenues and customer satisfaction, and minimize service time, which is associated with opportunity costs. In reaching these objectives, employees of these service providers are constrained by their limited attention or time available, which we refer to as capacity, leading to the quality/speed trade-off because delivering high service quality takes more time. There are, however, to the best of our knowledge, no empirical studies on this speed/quality trade-off. We demonstrate that understanding this trade-off is essential to the optimal labor management, and we show how a service organization can benefit by understanding this trade-off using large-scale data analysis.

In this paper, we examine how workload affects service speed (as reflected in the service time) and quality (as reflected in the sales amount) decisions, using a set of unique and very detailed transaction-level data from a restaurant chain's point-of-sales system that contains approximately 190,000 check-level observations for five restaurants from August 2010 to June 2011. We demonstrate how staffing capacity can be leveraged to optimize the workload. We show that servers strive to maximize sales and speed efforts simultaneously, depending on the relative values of sales and speed. After disentangling the endogeneity of demand and supply in this setting using a natural experiment (labor management software implementation) and other instruments, in our focal restaurants, we find that servers react nonlinearly to the workload, which is defined as the number of tables or diners that a server simultaneously serves. Surprisingly, when the overall workload is small, servers expend more and more sales efforts with the increase in workload, which consists of both up-selling and cross-selling, but at a cost of slower service. However, above a certain threshold (around 0.46 tables per server above the sample mean of 2.16 tables), servers start to reduce their sales efforts and work more promptly with a further rise in workload due to trading off service quality for service speed. We explain that servers value sales more than speed when the overall workload is low. However, when the overall workload is high, servers may deem speed more valuable than sales and therefore serve the tables faster, which can potentially reduce the sales. Thus, we conclude that our focal restaurants can improve performance by exploiting the inverted-U-shaped relationship between sales and workload so that reducing the number of waiters can *both* significantly increase sales (by about 3%) and reduce costs (by about 17%). We test the robustness of our results using different workload measures, and we discuss the managerial implications of measuring workload differently. In particular, a priori, hourly analysis may seem to be most appropriate because restaurants tend to schedule servers on an hourly basis. We show that the hourly level analysis corroborates our main check-level findings on servers' behavior under various levels of workload, but it may produce different estimates of the magnitudes of workload effects.

2. Related Literature

Our research setting relates to five streams of literature: (1) optimal staffing decisions, (2) OM/HRM interface research, (3) empirical workload studies, (4) speed/quality trade-off, and (5) restaurant revenue management research.

There is an extensive analytical literature on staffing decisions in services (see Gans et al. 2003 and Akşin et al. 2007 for excellent literature reviews). Classical models here tended to assume that workers' productivity is independent from their work environment largely because of modeling tractability issues. Recent studies have started to incorporate some aspects of worker behavior in their models. For example, Arlotto et al. (2014) consider both worker heterogeneity and learning for staffing decisions. Research has also been done to incorporate new work environments. For example, unlike in traditional call centers where one agent handles one call at a time, Luo and Zhang (2013) model how to staff an instant messaging contact center, where one server handles multiple customers simultaneously. Compared with the voluminous analytical literature on optimal staffing decisions, there are only a handful of empirical studies that explicitly examine the impact of staffing on workers' performance. In a retail setting, Fisher et al. (2006) and Perdikaki et al. (2012) found that store staffing levels influences the conversion of traffic into sales. In another retail study, Mani et al. (2011) estimated that an optimal staffing level could improve average store profitability by 3.8% to 5.9%.

[1] We do not restrict quality to service accessibility, which is often assumed in the service operations literature. We broadly regard quality as a standard for service content.

These studies are conducted at the store level, whereas our analysis focuses on an individual employee level. Still, as suggested by Akşin et al. (2007), there is a general lack of empirical research about the impact of staffing on performance, which we study in this paper.

An increasing number of studies has sought to bridge OM models with behavioral literature to relax the often rigid assumptions of the classical OM models and to study the impact of external factors on individuals' performance (see Boudreau et al. 2003 and Bendoly et al. 2006 for comprehensive reviews). For example, Schultz et al. (1998) challenged the traditional assumption that a worker's production rate is independent from the environment. In a production line simulation experiment, they found that individuals' processing times were dependent on the state of the system, such as the buffer size, as well as on the processing speed of coworkers. Unlike what the assumption of independence would predict, the experiment revealed less idle time and higher output because people tended to speed up and avoid idle time. Schultz et al. (1999) explained that a low-inventory system improves productivity because it creates more feedback, stronger group cohesiveness, and better task norms than a high-inventory system. In another lab experiment, Bendoly and Prietula (2008) asked subjects to solve vehicle routing problems. They found a nonmonotonic relationship between pressure (induced by workload) and motivation, which affected performance. This effect can be further moderated by learning. Furthermore, Bendoly (2011) used physiology data about eye dilation and blink rate to measure the arousal and stress levels of subjects, which confirmed that task-state conditions affected emotions and thus task performance. Although this stream of research is experimental, real-world systems are generally more complex; therefore, Boudreau et al. (2003) called for observational studies to validate the behavioral lab findings in real industrial settings. Our research belongs to a very recent stream of observational papers that have answered this call (e.g., Huckman et al. 2009, Staats and Gino 2012) by using archival data to understand the impact of external factors on workers' performance.

Several researchers have recently turned to understanding the impact of workload, an integral environmental factor, on individual performance. They often use healthcare services as a test bed. For example, KC and Terwiesch (2009) conducted a rigorous empirical analysis of the impact of workload on service time using operational data from patient transport services in cardiothoracic surgery. They found that workers speed up as workload increases, and that this positive effect may be diminished after long periods of high workload. KC and Terwiesch (2012) showed further evidence that the occupancy level of a cardiac intensive care unit is negatively associated with patients' length of stay because the hospital, faced with high occupancy, is likely to discharge patients early. Although high workload may stimulate medical workers to accelerate their services, Batt and Terwiesch (2012) discovered that the net effect of high congestion is to actually decrease the service rate. These studies focused on the impact of workload on service time, and other studies separately examined the impact of workload on service quality. For example, Kuntz et al. (2014) suggested a nonlinear relationship between hospital workload and mortality rates. Powell et al. (2012) found that overworked physicians generate less revenue per patient because of a workload-induced reduction in diligence over paperwork. Our study distinguishes itself from this stream of literature in two aspects: (1) our study explicitly stresses the value of empirical studies for making optimal staffing decisions; (2) we study how workers react to workload in terms of both service time and quality, and in particular, we analyze how workers make speed/quality trade-off decisions.

A growing number of papers are starting to use analytical modeling approaches to yield insights into how workers make such trade-off decisions. For example, Hopp et al. (2007) discovered that workers, similar to call center agents, use both time and quality as a buffer for variability. Debo et al. (2008) suggested that service providers may yield higher revenues from "inducing service" at low workloads than at high workloads because "service inducement" may intensify congestion in the case of high workloads. Furthermore, Kostami and Rajagopalan (2014) analyzed this speed/quality trade-off in both single-period and multiperiod settings. In addition, Anand et al. (2011) found that service providers slow down as customer intensity increases, which causes the equilibrium service value to increase. As a result, they suggested that servers may become slower when the number of competing servers increases. Alizamir et al. (2013) studied how to dynamically balance diagnostic accuracy against delays in the process of performing additional diagnoses, considering servers' beliefs about the congestion level, customer type, and the number of tests performed so far. There are, however, to the best of our knowledge, no empirical studies on this speed/quality trade-off.

Finally, whereas papers on restaurant management have analyzed the impact of pricing, table mix, table characteristics, food, atmosphere, fairness of wait, and staff training on financial performance (see Kimes et al. 1998, 1999; Kimes and Robson 2004; Robson 1999; Kimes and Thompson 2004; Sulek and Hensley 2004), we contribute by showing that staff workload has a major impact on revenue generation.

3. Wait Staff Activities and Hypotheses Development

In the United States alone, the restaurant industry employs about 13 million workers, who provide more

than $500 billion in meals per year; yet rigorous empirical studies of restaurant workers are lacking. For our analysis we selected the restaurant setting because (1) workloads in restaurants tend to be highly variable, which provides an opportunity to study how changes in workload affect worker performance; (2) the restaurant industry is labor-intensive, employing approximately 10% of the total workforce in the United States; and (3) its productivity is only half that of manufacturing industries, creating multiple opportunities for productivity improvement (Mill 2006).

3.1. Wait Staff Activities

Waiters and waitresses, also known as servers, serve diners once customers are seated. In a typical work scenario (Fields 2007), they first greet diners shortly after they are seated. They instantaneously fill water glasses, present the menu, and ask diners whether they would like anything from the bar. Then they return to the table to present the specials and take the order. After serving the food, they check on the table during the meal for any special requests or additional drink orders. Finally, they present the check and change, thanking diners on their way out of the restaurant.

In performing these activities, servers expend two types of efforts, i.e., sales (or service quality) and speed, which will influence sales per check and meal duration, respectively. Sales efforts consist of both up-selling more expensive items and cross-selling additional items, which both lead to higher sales per check. As an example of up-selling, servers may testify from their own "tasting experience" that lobster tails are extremely delicious or servers may inform diners that uniquely tender Kobe beef comes from cows that are hand massaged. The lobster tails and Kobe beef are going to cost more than a chicken salad that a diner would probably order otherwise. As an example of cross-selling, a server may make suggestions of certain (somewhat expensive) wines that will go well with the main dish. In addition, servers may check on diners a couple of times to ask them if they would like anything else, such as drinks and desserts. Research shows that diners are more likely to purchase a dessert if a server makes such a suggestion (Fitzsimmons and Maurer 1991). In addition, to expend effort in speed, servers may carry multiple items from the kitchen to save trips and time. They also need to remember cooking times and what stage of the meal the diners are at in arranging the time to drop the entree tickets. By making decisions about how much effort to put into sales and speed, servers aim to simultaneously maximize both. However, servers are constrained by their limited capacities of attention (or time available) to increase sales and speed. Because of these constraints, servers need to make trade-off decisions given current workload level.

How they decide between sales and speed efforts is a very interesting question. According to a study by the National Restaurant Association (Mill 2006), complaints about restaurant service far exceed complaints about food or atmosphere. The majority of complaints are about service speed and inattentive waiters, for example, long waits to settle the bill and a server's impatience with answering menu questions. In addition, sales are of great importance to restaurants which, on average, generate very small pretax profit margins, averaging just 4% of sales revenues. To increase sales, servers are usually instructed and trained to sell more items and to sell more expensive items. Hence, understanding servers' behavior toward sales and speed is critical for improving restaurant service operation.

Because servers' sales and speed efforts are not directly observable, we rely on observable performance metrics, namely, the sales and meal duration of each meal, to infer servers' efforts in sales and speed. In the next subsection, we develop hypotheses about the impacts of workload on sales and meal duration, respectively.

3.2. Hypotheses Development

Conventional wisdom suggests that focusing on one task should ensure a fast completion time. In other words, working on multiple tasks without changing one's total capacity and capability will mechanistically diversify one's attention, thus decelerating the completion time of each task. However, under excessive workload, workers may feel stressed (Bendoly 2011) and decide to rush their work at a cost of quality by cutting corners (Oliva and Sterman 2001) but nevertheless accelerating the completion of tasks. These seemingly conflicting predictions seem to suggest that the impacts of workload on performance is influenced by both mechanics and human factors (Bendoly and Hur 2007). Therefore, we develop our hypotheses about the effects of workload on sales and meal duration through two different types of effects, i.e., mechanistic effects and behavioral effects. We also define and distinguish these effects below.

3.2.1. Mechanistic Effects. Mechanistic effects include the factors that change workers' performance without changing their intrinsic capacity and capability. In a processor sharing system, where an agent distributes his or her limited attention to several customers simultaneously, the service time/quality for one customer depends on the number of customers that the worker handles simultaneously as well as on the service time/quality of other customers (Kleinrock 1976, Akşin and Harker 2001, Luo and Zhang 2013). One of the reasons for this effect is that each of n jobs in a processor-sharing system is likely to get approximately $1/n$th of workers' limited capacity, which may

contribute to lower quality and speed. Another reason is that, mechanistically speaking, an extra customer might require some fixed set-up time. In particular, as the number of other customers increases, the service time might be prolonged, while holding service quality constant. At the same time, while handling multiple customers, a server constrained by the limited capacity for attention will give each customer less attention, which may consequently reduce service quality.

Restaurant servers operate in such a processor sharing system, where one server often waits multiple tables and follows a set-up routine for each new party, such as setting the table and offering water before taking orders. Assuming servers work at the same pace and try to maintain some constant quality, when the workload increases, i.e., the number of parties/tables that a server simultaneously handles increases, diners may have to wait. For example, diners at table i, which have already been seated, may need some assistance from their server who is busy serving other tables or setting up another table. Therefore, they have to wait to get the server's attention, prolonging the meal duration at table i. Furthermore, while serving multiple tables, servers may become so occupied with carrying food that they have no time/capacity to conduct effective suggestive selling, thus lowering the final sales.

3.2.2. Behavioral Effects. Unlike mechanistic effects, behavior effects affect performance via changing workers' intrinsic capacity, capability, and how they perform work. We theorize three behavioral effects that may contribute to the impacts of workload on sales and meal duration.

Effect I: *Motivation*. Motivation unlocks workers' potential capabilities and stimulates them to expend more effort with intrinsic willingness and enthusiasm. Such motivation may be improved by appropriately increasing workload. Goal-setting theory suggests that challenges faced by workers can enhance motivation (Locke 1968, Latham and Locke 1979). Increasing workload can be perceived as a challenge, thus increasing arousal regarding the work (Bendoly 2011) and stimulating motivation to exert more efforts (Deci et al. 1989). Indeed, cognitive psychology also suggests that workload may trigger the cortex to release hormones that improve cognitive performance, which develops workers' potential capacity (Lupien et al. 2007).

For instance, workload can be represented as the number of tables that a restaurant server simultaneously handles. According to the aforementioned goal-setting theory, as workload increases, servers may perceive a challenge, which may stimulate them to expend extra efforts. Furthermore, serving tables is not only a physical job but also an emotional or cognitive job in that servers must constantly anticipate diners' needs and multitask to fulfill them. The extra hormones released from increasing workload should also increase servers' capacity and help them enhance their service performance.

Effect II: *Antiproductive Emotions*. Excessively high workload may induce antiproductive emotions, thus lowering performance. When workload becomes too high, it may function as a constraint that causes frustration and hinders workers from fulfilling their goals (Peters and O'Connor 1980), which may further reduce workers' motivation and commitment (e.g., O'Connor et al. 1984). Moreover, working under high workload will force workers to pursue multiple goals simultaneously in a finite amount of time. These multiple goals may create conflicts and increase the expected difficulty of achieving goals, thus lowering workers' commitment (Donahue et al. 1993, Dalton and Spiller 2012). Furthermore, heavy workload can cause fatigue (Cakir et al. 1980, Setyawati 1995) and stress (Bendoly 2011), which may lead to reduced motivation and effort.

When servers handle too many tables simultaneously, they may also experience aforementioned antiproductive emotions. Servers' goals should be to maximize sales and speed at each table. However, waiting too many tables may hinder them from achieving these goals. For example, servers who serve multiple tables are prone to making errors when taking orders, thus limiting their sales effectiveness. Knowing that they cannot fully achieve their goals, servers may feel frustrated and thus compromise their commitment and effort. For example, they may rush diners by presenting the check without being asked. Servers may also experience fatigue and stress caused by heavy workload, and decrease their efforts.

Effect III: *Discretionary Service*. In contrast to the aforementioned motivation effect that focuses on intrinsic willingness and enthusiasm, workers can simply be incentivized by workload to adjust their service quality and speed at their discretion without necessarily displaying willingness and enthusiasm. Hopp et al. (2007) cite call center agents as an example and theoretically analyze such a setting, where servers can use their discretion in up-selling strategies and thus have great influence over service duration. The authors assume that the expected revenue will be concavely increasing in service duration. Among other findings, they discover that servers may adjust their service time and thus the service quality in response to system congestion. In other words, servers use service time and quality as buffers against congestion variability. More surprisingly, they find that increasing capacity in this discretionary task completion setting may even increase congestion because servers may find it more attractive to prolong service duration in order to achieve higher quality than to speed up their service in order to reduce waiting cost. A similar system is studied by Debo et al. (2008), who theorize that a changing workload together with a variable fee structure may create an incentive for

service providers to extend service time and perform extra service in order to generate higher revenues. In particular, they argue that service providers may yield higher revenues from "inducing service" at low workloads than at high workloads because "service inducement" may intensify congestion in the case of high workload.

Restaurant servers have similar discretion in terms of service quality and speed. Besides the minimum service procedure, such as taking the order and settling the bill, restaurant servers may additionally chat with diners and check if they need to purchase anything else in the middle of the meal. Performing these additional service tasks takes extra time, and should be positively associated with final sales, a measure of service quality. In response to low workloads, servers should have a strong incentive to perform extra service at the cost of longer service time because (1) the waiting cost is relatively low and (2) a sales-maximizing (i.e., tips-maximizing) server should extract more revenues from each of the few tables that they serve. However, in response to high workloads, the incentives of servers may change from seeking extra service quality/sales to faster service because (1) the waiting cost is high and (2) servers wish to turn over the tables to seat new diners, who tend to spend more money per unit of time than lingering diners. In sum, different levels of workload may create different incentives of their discretionary efforts, as reflected in sales and speed.

Competing Hypotheses for Sales: The aforementioned theories seem to show conflicting effects of workload on sales. On one hand, mechanistic effects as well as behavioral effects II and III seem to suggest that increasing workload may reduce sales per check. On the other hand, behavioral effect I seems to imply that workload should boost sales. Hence, we form two competing hypotheses for sales:

Hypothesis 1A (H1A). *As workload increases, sales will increase.*

Hypothesis 1B (H1B). *As workload increases, sales will decrease.*

Competing Hypotheses for Meal Duration: Similarly, the theories seem to imply conflicting effects of workload on meal duration. Whereas mechanistic effects seem to suggest that increasing workload may decelerate meal duration, all three behavioral effects seem to indicate that increasing workload may accelerate meal duration. Therefore, we propose two competing hypotheses for meal duration:

Hypothesis 2A (H2A). *As workload increases, meal duration will increase.*

Hypothesis 2B (H2B). *As workload increases, meal duration will decrease.*

4. Data

4.1. Research Setting and Data Collection

To examine our research hypotheses, we worked closely with a restaurant chain's management to collect point-of-sales (POS) data from five restaurants owned and operated by Alpha (the real name is disguised for confidentiality reasons), a restaurant chain that offers family-style casual dining service in the Boston suburbs. We gained access to their sales data as a part of their implementing a new server scheduling system, the implementation of which is used for identification purposes in §5.2. The restaurants are open from 11:30 A.M. to 10:00 P.M. from Monday to Thursday, and from 11:30 A.M. to 11:00 P.M. from Friday to Sunday. Diners include couples, families, students, and their friends. The restaurants have a full-service bar and they offer internationally inspired fusion food. Our study focuses on the main dining room because the bar and take-out services operate according to a different business model and they would require different operationalization of variables. Our data consist of 11 months of transactions from August 2010 to June 2011. The transaction data include information about servers, sales, gratuities, party size, and service start and end time. To reduce the influence of outliers (e.g., very large parties and private events), we drop the transactions that include the day's top and bottom 7.5% of checks. Our final data set includes approximately 190,000 check-level observations. We believe that our restaurant sample represents an appropriate data set to study the impact of workload on restaurant performance because we possess comprehensive temporal and monetary information for each meal service that occurred during both peak and nonpeak hours, allowing us to systematically quantify the impact of workload on server performance. At the same time, the data set we possess is among the largest and most granular in the existing literature on the impact of workload on performance.

4.2. Measures and Controls

To understand how workload affects servers' behavior of handling each check, we use individual checks as the unit of analysis. In practice, restaurants tend to schedule servers on an hourly basis, so we also aggregate all variables at the hourly level to provide a robustness check and assess staffing implications in §5.5. We are interested in studying servers' performance and therefore we operationalize dependent variables $Sales_i$ and $MealDuration_i$ to reflect the sales and the length of a check i, which is exclusively assigned to one server in our focal restaurants. We infer the meal duration of each check from check opening and closing times recorded in our POS data. This inferred duration could be slightly inaccurate because diners could arrive before the check was opened and they could leave after the check was closed. Nevertheless, our meal duration

measure directly captures the server's involvement with the customer (rather than, say, the host's involvement before the check is open), which is also consistent with previous literature (Kimes 2004).

We define the key independent variable $AvgTables_i$ as the average number of tables (parties) that a server handles simultaneously together with the focal check i being analyzed. For example, suppose check i lasts 40 minutes. During this period, a server overlaps with another table (party) for 20 minutes. Our workload measure $AvgTables_i$ is $(40\ \text{min} + 20\ \text{min})/(40\ \text{min}) = 1.5$ tables. First, weighting the workload by the meal duration reflects the exact amount of load that affects check i because the time spent on other tables either before or after check i should largely not affect check i.[2] Furthermore, we believe that tables are more appropriate than diners as our main unit of analysis for the following reasons. Tables (parties) are likely to be more salient than diners for servers because (1) hosts and hostesses are instructed to distribute tables (parties) evenly among the servers, (2) servers are assigned to sections, which consist of a relatively fixed number of tables.[3] In addition, the marginal workload of an additional table is more significant than the marginal workload of an extra diner in a party because a server needs to perform a fixed set of procedures, such as taking the order, to every table regardless of the party size. Of course, the number of diners is a reasonable alternative workload measure. We use diners per server as an alternative workload measure in the robustness check section and the results are qualitatively the same.

In addition to these main variables of interest, we consider the following control variables. Variable $PartySize_i$ is the number of diners in a particular party i, which should affect both sales and meal duration. Variable $StoreItems_i$ is the arithmetic average of the storewide number of items ordered at the beginning and at the end of check i, which is used to control for the workload on the kitchen. Finally, we also control for the time/date/location of check i. Night hours usually generate more sales than lunch hours, so we include the categorical control variable $Hour_i$ to represent the hour when check i was opened. Weekends are usually busier than weekdays, so we include another categorical control, $DayWeek_i$. Business during the summer in these locations is usually slower than during the winter because many residents go on vacation. In addition, economic trends may affect diners' consumption level. To adjust for these temporal factors, we consider another categorical control variable, $YearWeek_i$, which starts at one from the first week of August of 2010 and ends at 48 in the last week of June of 2011. We choose to have this trend control at the weekly level because our instrumental variables are lagged by one week. (For more on the instrument validity, see §5.4.) We also control for store fixed effects using the variable $Store_i$. To summarize, Table 1 presents a list of variable definitions. These data allow us to test our hypotheses while controlling for factors that can affect servers' performance.

Table 1 Check-Level Analysis Variable Definition

Variable	Definition
$Sales_i$	Sales of check i measured in dollars
$MealDuration_i$	Meal duration of check i measured in minutes
$AvgTables_i$	Average number of tables (parties) that a server handles simultaneously together with check i
$PartySize_i$	Number of diners in a particular party i
$StoreItems_i$	Arithmetic average of storewide number of items ordered at the beginning and at the end of check i
$Hour_i$	Categorical variable indicating the hour when check i was opened
$DayWeek_i$	Categorical variable indicating the day of the week when check i was opened.
$YearWeek_i$	Categorical variable indicating the week order in the study period; e.g., the first week of August 2010 is one, while the last week of June 2011 is 48
$Store_i$	Categorical variable indicating the store where check i happened

4.3. Descriptive Statistics

Table 2 presents the summary statistics of the check-level variables. On average, each check generates $40.38, taking approximately 48 minutes. Each check is, on average, shared by 2.35 diners. In addition, in the course of a meal, there are, on average, close to 80 items ordered in the entire restaurant.

Before testing our hypotheses, we transform *Sales* and *MealDuration* into their natural logarithms in order to linearize the exponential forms of sales and meal duration models (Kleinbaum et al. 2007). These variables have large standard deviations relative to

Table 2 Summary Statistics of Check-Level Variables

	Sales	*MealDuration*	*AvgTables*	*PartySize*	*StoreItems*
N	190,799	190,799	190,799	190,799	190,799
Mean	40.38	47.98	2.16	2.35	79.90
SD	15.69	16.23	0.83	0.87	36.02
Min	7.88	21.84	1	1	2
P5	20.38	28.39	1	1	23
P25	28.16	37.13	1.57	2	53
P50	37.45	43.69	2.05	2	78.5
P75	49.73	56.79	2.63	3	105
P95	70.86	80.82	3.61	4	140.5
Max	131.75	113.59	9.65	5	261.5

[2] We used alternative individual-level workload measures, such as the number of tables either at the beginning of or at the end of check i. (KC and Terwiesch 2009 counted the hospital bed occupancy at the beginning of a patient's admission. KC and Terwiesch 2012 measured the ICU occupancy at the time of a patient's discharge.) These alternative measures yielded qualitatively congruent results.

[3] In practice, some smaller tables can be combined to form a bigger table.

Table 3 Correlation Matrix of Check-Level Variables

	log(*Sales*)	log(*MealDuration*)	*AvgTables*	*PartySize*	*StoreItems*
log(*Sales*)	1.000				
log(*MealDuration*)	0.256*	1.000			
AvgTables	−0.064*	0.098*	1.000		
PartySize	0.536*	0.029*	−0.077*	1.000	
StoreItems	0.214*	0.081*	0.241*	0.113*	1.000

*Significant at the 0.01 level.

their means, so transforming them is recommended to increase normality prior to model estimation (Afifi et al. 2004). Log transformation increases the normality of the errors, which ensures that our hypothesis test statistics follow t-distribution. In addition, transforming the monetary variable normalizes the scale to percentages for easier interpretation.

Table 3 shows the correlations of the check-level variables. We observe that log(*Sales*) is positively associated with log(*MealDuration*) (correlation = 0.256), *PartySize* (correlation = 0.536), and *StoreItems* (correlation = 0.214). The correlations among the predictors are relatively low, suggesting that the predictors should not cause the multicollinearity issue in the model estimation.

5. Estimation and Results

First, we estimate a set of multivariate regression models to provide a preliminary and exploratory analysis. Second, we use an instrumental variable approach to address potential omitted variable bias and simultaneity bias issues. We conduct additional analysis to understand the mechanisms of our empirical results and perform robustness checks of our main results. Finally, we discuss the implications of alternative workload measures.

5.1. Multivariate Regression

We first specify the following linear regression model to provide a preliminary analysis of the relationship between workload and servers' performance:

$$\log(Sales_i) = \alpha_0 + \alpha_1 AvgTables_i + \alpha_2 PartySize_i + \alpha_3 Controls_i + \varepsilon_i, \quad (1)$$

$$\log(MealDuration_i) = \beta_0 + \beta_1 AvgTables_i + \beta_2 PartySize_i + \beta_3 StoreItems_i + \beta_4 Controls_i + \xi_i. \quad (2)$$

To ensure that we correctly specify the model given the competing hypotheses, we allow for nonlinear relationships between dependent and independent variables. To model this nonlinear relationship, we include $AvgTables^2$ in the models in addition to *AvgTables* itself. That is,

$$\log(Sales_i) = \alpha_0 + \alpha_1 AvgTables_i + \alpha_2 AvgTables_i^2 + \alpha_3 PartySize_i + \alpha_4 Controls_i + \varepsilon_i, \quad (3)$$

$$\log(MealDuration_i) = \beta_0 + \beta_1 AvgTables_i + \beta_2 AvgTables_i^2 + \beta_3 PartySize_i + \beta_4 StoreItems_i + \beta_5 Controls_i + \xi_i. \quad (4)$$

In these models, $Controls_i$ include $DayWeek_i$, $Hour_i$, $YearWeek_i$, and $Store_i$ to adjust for the time/date and location factors, which is equivalent to a store fixed-effect model because we include store-specific time-invariant factors among our controls, which help account for unobserved heterogeneity among stores, such as the income level of the neighborhood and other time-invariant omitted variables. We further center *AvgTables* and $AvgTables^2$ around their means for interpretation purposes. Furthermore, we compute Huber–White robust errors to alleviate potential heteroskedasticity issue (White 1980).

Note that the quadratic specification of $AvgTables_i$ allows us to compute the critical points in the regression models. In particular, since the critical point of a quadratic function of the form $f(x) = ax^2 + bx + c$ is $-b/(2a)$, the critical point of, e.g., $\log(Sales_i)$, is expected to be at $-\alpha_1/(2\alpha_2)$.

Although these regression models are useful as a preliminary estimator (Kennedy 2003), they may not address two potential endogeneity issues:

1. *Omitted variable bias*. The omitted variable bias from ordinary least squares (OLS) is given by $r_{Tx}\beta_x s_x/s_T$, where r_{Tx} is the correlation between the omitted variable x and the endogenous variable T (workload), β_x is the relationship between the omitted variable x and the dependent variable (sales), and s_x and s_T are the standard deviations of x and T. In our setting, one major omitted variable is managers' demand forecast, which will enter the error term. The demand forecast should be positively correlated with sales, i.e., $\beta_x > 0$. In addition, sales forecast should be positively correlated with staffing levels, but negatively associated with workload, namely $r_{Tx} < 0$ because managers tend

to match the demand with staffing. The standard deviations of x and T should be positive. Hence, OLS may underestimate the true impact of workload on sales. There are other possible omitted variables in the error, such as consumers' price sensitivity and their intrinsic level of hunger. Nevertheless, this type of consumption behavior-related omitted factors are likely to be uncorrelated with congestion and staffing. To address this potential omitted variable bias issue, we first adopt an instrumental variable two-stage least squares (2SLS) approach (Angrist and Krueger 1994), which is elaborated in §5.2. We also performed Hausman endogeneity tests after 2SLS estimations and rejected the null hypotheses that those workload measures were exogenous in the sales model.

2. *Simultaneity bias* (*reverse causality*). Managers generally do not make demand forecast for meal duration, so the same omitted variable may not apply to meal duration. Nevertheless, the OLS meal duration model may suffer from a simultaneity bias (reverse causality). First, long meal duration may be indicative of insufficient staffing. Second, randomly long meal duration may mechanistically increase workload. For these reasons, this positive correlation between meal duration and workload may overestimate the true effect of workload. Similar to the sales model, we took the instrumental variable approach to address this potential simultaneity bias issue. The Hausman endogeneity test rejected the null hypotheses that workload measures were exogenous in the meal duration model.

5.2. 2SLS Model

We adopt an instrumental variable 2SLS approach (Angrist and Krueger 1994) to address the endogeneity issue for the following reason. First, the 2SLS instrument estimator can provide consistent estimates of the dependent variables using a large sample. It is also quite robust in the presence of other estimation issues such as multicollinearity. For these reasons, the 2SLS instrumental variable approach is widely used to address endogeneity issues (Kennedy 2003). A valid instrumental variable should satisfy relevance and exclusion restriction assumptions (Wooldridge 2002). In particular, it should be uncorrelated with the error (i.e., exclusion restriction) and correlated with the endogenous regressor (i.e., relevance). In other words, the instrument should explain the outcome variable only through the endogenous regressor.

We propose two types of instruments. First, we utilize an exogenous shock in our study period: the implementation of a new staffing system at one of the restaurants. On March 21, 2011, one of the restaurants adopted a new computer-based scheduling system, while the other four restaurants continued to rely on managers to make demand forecasts and staffing-level decisions. In particular, we create the dummy variable *Software*, which equals one for all the observations after the software implementation date at the store that implemented the software and equals zero for all other observations. The management chose this particular restaurant as a pilot project before subsequently implementing the software chainwide. The sales performance of this restaurant is similar to the other four restaurants in that they all show stable sales, thus reducing the concern of selection bias. Using historical sales data, the new software forecasts the need for servers. It is reasonable to assume that the system will prescribe different staffing levels from those that managers might suggest because it uses more historical sales data than a manager can handle. In other words, the system should have affected staffing levels after its implementation. We further control for demand and store fixed effects. Hence, variable *Software* should reflect the impact of staffing levels on workload, satisfying the relevance condition. In addition, we would expect the implementation of the software to affect sales and meal duration only through staffing level because the system simply provides a user-friendly interface to schedule servers, perhaps with a different forecast of demand. Diners do not observe the implementation of this labor scheduling system. For these reasons, the implementation of the system should satisfy the exclusion restriction condition.

Admittedly, both managers and servers in that particular restaurant may have anticipated the implementation of the new software. They may also have had different emotional responses to a computerized scheduling system. For both of these reasons, they might have readjusted their productivity, which could invalidate using the software implementation as an instrument. To address this potential issue, following Bloom and Van Reenen (2007) and Siebert and Zubanov (2010), we supplement our analysis using another type of instrumental variable, the lagged values of the endogenous independent variables. To operationalize these lagged variables, we first compute the hourly workload during the same hour as check i takes place. In particular, this hourly workload is defined as the number of parties who started meals during the same hour divided by the number of servers who processed at least one check in the same hour, i.e., $HRLoad_i = HRTables_i/HRServers_i$. Then we compute *LWHRLoad* and $LWHRLoad^2$, which are the *HRLoad* and $HRLoad^2$ of the same restaurant during the same hour of the previous week to use as instruments for the current week. For example, if check i happened at 8:30 P.M. on August 8, 2010, at restaurant k, its instrument is the hourly load of the 8:00 P.M. slot on August 1, 2010, at restaurant k. We then mean center these instruments for interpretation purposes. We choose the lag to be one week because the restaurants in our study usually consider the load from a week ago to generate staff

schedules for the current week. For this reason, the weekly lagged variables should correlate with the current hourly load. Note that we are unable to construct a check-level lagged workload variable because meals start randomly. Nevertheless, the current hourly load should correlate with check-level workload. Therefore, we anticipate that the weekly lagged hourly workload should satisfy the relevance assumption.

Moreover, we expect these lagged values of the endogenous variables to be exogenous because the staffing decisions from a week ago should not determine the unobserved factors for sales and meal duration during the current week, i.e., contemporaneous shocks. In other words, the lagged variables are not contemporaneously correlated with the disturbance (Kennedy 2003), so they should satisfy the exclusion restriction assumption of a valid instrument. Admittedly, the lagged workload may not be ideal in the event of common demand shocks that are correlated over time. However, these common demand shocks are basically trends (Villas-Boas and Winer 1999), which are controlled for in our models with the categorical control variable *YearWeek*, thus lessening this potential concern. We further provide relevant statistics to show the validity of these instruments in §5.4. With both types of instrumental variables, we employ the following 2SLS estimation procedure:

Stage 1. Estimate endogenous independent variables, namely, *AvgTables* only for the linear models 1 and 2, and *AvgTables* and $AvgTables^2$ for the quadratic models 3 and 4, using OLS and instrumental variables (i.e., *Software* and *LWHRLoad* for the linear models, and *Software*, *LWHRLoad*, and $LWHRLoad^2$ for the quadratic models) and other exogenous controls (specified in models 1–4). We then compute the predicted values of the endogenous independent variables, namely, $\widehat{AvgTables}$ alone in the linear models, and both $\widehat{AvgTables}$ and $\widehat{AvgTables}^2$ in the quadratic models.

Stage 2. Use the predicted endogenous independent variables from Stage 1 to estimate the coefficients of each equation (models 1–4) with OLS regression and robust errors.

5.3. Results

Table 4 shows the results of check-level sales analysis. First, in the linear models, the OLS estimate of *AvgTables* is −0.016 (model 1), in support of H1B. However, after the endogeneity issue is corrected by the instruments, the 2SLS estimate becomes 0.0336 (model 1 estimated by 2SLS), greater than the OLS estimate, as expected, supporting H1A. This result suggests that workload may increase servers' sales performance, controlling for party size and other factors. As expected, a large party size is positively associated with higher sales per check. We then visually check for the potential nonlinear relationship between *AvgTables* and log(*Sales*) and observe

Table 4 **Impact of Check-Level Workload** *AvgTables* **on** log(*Sales*)

	Linear model 1 estimated by OLS	Linear model 1 estimated by 2SLS	Quadratic model 3 estimated by OLS	Quadratic model 3 estimated by 2SLS
AvgTables	−0.0160***	0.0336**	−0.0031**	0.0942***
	(0.0009)	(0.0120)	(0.0010)	(0.0189)
$AvgTables^2$			−0.0134***	−0.1293***
			(0.0005)	(0.0296)
PartySize	0.2240***	0.2266***	0.2226***	0.2116***
	(0.0008)	(0.0012)	(0.0008)	(0.0039)
Controls	Yes	Yes	Yes	Yes
Observations	190,799	185,545	190,799	185,545
Prob > Chi-sq	<0.001	<0.001	<0.001	<0.001

Note. Standard errors are shown in parentheses.
**$p \le 0.01$; ***$p \le 0.001$.

that the relationship between the workload and the sales may not be linear. Furthermore, we analytically check for the potential nonlinearity. The adjusted R^2 of model 1 is 0.3686, while the adjusted R^2 of model 3 is 0.371, which suggests that the quadratic model may provide a better goodness-of-fit than the linear model. The quadratic specification results show that the coefficients of $AvgTables^2$ are consistently negative (−0.0134 from model 3 and −0.1293 from model 3 estimated by 2SLS). Interpreting the coefficients from the 2SLS, we find that the critical workload is about $(0.0942/(2 \times 0.1293) \approx 0.36)$ tables, less than one standard deviation (0.83) above the sample mean, which is 2.16 tables. These results suggest that variable *AvgTables* first concavely increases sales and then concavely decreases sales. By fitting the quadratic function, we essentially obtain approximation of the expected sales function, which takes demand uncertainty into consideration. Thus, we find the "true" optimal workload. As far as we know, the company does not currently use newsvendor logic about staffing underage or overage costs and simply applies the common "rule of thumb" in this industry to calculate the right number of servers based on demand forecast. Our finding that the critical workload is above the sample mean suggests that this rule of thumb is likely to be too generous and the restaurants need fewer servers. Even if the original rule of thumb was developed to buffer against uncertainty, we show that it is not effective, possibly because adaptive behavior of servers is not accounted for. We further calculate that changing the current workload to the optimal value would have generated $(0.0942 \times 0.36 - 0.1497 \times 0.36 \approx 2\%)$ sales lift per check on average, holding party size and other factors constant. Consistent with linear models, a larger party size is positively associated with higher sales per check.

Note that the coefficient of *AvgTables* is negative in the OLS model (−0.0031), but its sign becomes positive in the 2SLS model (0.0942) after the endogeneity issue

is corrected by the instruments, as expected. Without this correction, one would have mistakenly concluded that the optimal workload is smaller than the sample mean, which would lead to erroneous staffing decisions. On the other hand, the OLS-estimated quadratic term is less negative than the 2SLS estimator. The bias direction for a nonlinear relationship is generally complicated to identify because one cannot keep the linear term unchanged while changing the quadratic term.[4]

Restaurants have capacity constraints because of the limited number of tables and diners per table. Of course, a truncation of demand may cause a concave relationship between workload and sales, even regardless of servers' behavior. To address this issue, we control for party size in the check-level analysis. In addition, our dependent variable—sales per check—should be immune to demand truncation caused by storewide capacity constraint. Furthermore, we discover that sales dip after workload reaches a high level. If the alternative explanation about demand truncation were valid here, sales would plateau as workload further increased.

The significant estimates in the quadratic model imply that the difference in mean sales for each extra table load is likely to change, in addition to the better model fit described before. Hence, we use the quadratic model specification as our main results for further interpretation and analysis. Moreover, the quadratic specification results seem to reconcile the two competing hypotheses. When the overall workload is low, behavioral effects, especially behavioral effects I and III, may dominate the mechanistic effects because servers may have excess capacity. In other words, under overall low workload, increasing workload may motivate servers to expend more sales effort and improve performance (as suggested by behavior effect I). Moreover, servers should have a strong incentive to perform extra service because a sales-maximizing (i.e., tips-maximizing) server should extract more revenues from each of the few tables that they serve. However, when the overall workload is excessively high, a further increase in workload may create antiproductive emotions (behavioral effect II) and diversify servers' attention (mechanistic effect), thus reducing sales. In addition, when workload is high, servers' incentives may change from increasing sales to increasing speed (behavioral effect III), which further reduces sales. Note that, when the workload is high, the motivational effect (behavioral effect I) may also be diminished because servers have physical capacity constraints. Hence, as workload increases, sales will first increase and then decrease.

[4] For example, suppose $Y = X + X^2 + e$. We assume the base case is $X = 0$, $e = 0$. If X increases by one, X^2 increases by one, and e decreases by one (omitted variable bias), then Y increases by one. Consider another case. If X increases by two, X^2 increases by four, and e decreases by two (omitted variable bias), then Y increases by four. Using these observed data, we fit a model $Y = aX + bX^2$ and solve for $a + b = 1$; $2a + 4b = 4$, which yields $b = 1$ and $a = 0$. As can be seen, even though the coefficient of X is underestimated, the coefficient of X^2 is not necessarily underestimated.

Table 5 **Impact of Check-Level Workload *AvgTables* on log(*MealDuration*)**

	Linear model 2 estimated by OLS	Linear model 2 estimated by 2SLS	Quadratic model 4 estimated by OLS	Quadratic model 4 estimated by 2SLS
AvgTables	0.0430***	−0.0439	0.0545***	0.0186
	(0.0009)	(0.0241)	(0.0010)	(0.0364)
AvgTables^2			−0.0111***	−0.0846*
			(0.0005)	(0.0333)
PartySize	0.0264***	0.0183***	0.2226***	0.2116***
	(0.0008)	(0.0024)	(0.0008)	(0.0039)
StoreItems	−0.0002***	0.0005**	−0.0002***	0.0002
	(0.0000)	(0.0002)	(0.0000)	(0.0002)
Controls	Yes	Yes	Yes	Yes
Observations	190,799	185,545	190,799	185,545
Prob > Chi-sq	<0.001	<0.001	<0.001	<0.001

Note. Standard errors are shown in parentheses.
*$p \le 0.05$; **$p \le 0.01$; ***$p \le 0.001$.

We now proceed with the check-level meal duration analysis and Table 5 presents the results. In the linear models, the OLS estimate of *AvgTables* is 0.043 (model 2), in support of H2A. However, the 2SLS estimate becomes −0.0439 but only marginally significant at 6.9% level (model 2 estimated by 2SLS) and smaller than the OLS estimate probably because the instruments correct for the expected upward bias of *AvgTables*. This result supports H2B suggesting that increasing workload may decrease meal duration, while holding party size and other factors constant. Following the same procedure as in the sales model, we visually check for the potential nonlinear relationship between *AvgTables* and log(*MealDuration*) and we analytically check for the potential nonlinearity. The adjusted R^2 of model 3 is 0.052, and the adjusted R^2 of model 4 is 0.055, suggesting that the quadratic model provides a better goodness-of-fit than the linear model.

Among the quadratic model results, the coefficients of *AvgTables*^{2} are consistently negative (−0.0111 from model 4 and −0.0846 from 2SLS estimated model 4). The coefficient of *AvgTables* is significant and positive in the OLS model, and the estimated coefficients are statistically undifferentiated from zero in the 2SLS model. The instruments may be correcting the bias of *AvgTables* downward. In addition, we anticipate that the instruments will increase the standard errors of the estimates because they reduce the variation of the $\widehat{AvgTables}_i$. Interpreting the coefficients from the 2SLS estimation, the maximal meal duration seems to happen right at the sample mean of 2.16 tables, suggesting an inverted-U-shaped relationship that *AvgTables* initially

concavely increases the meal duration of each check and then concavely decreases the meal duration.

In addition to the regression models described so far, we conduct a series of duration model analysis of log(*MealDuration*) as a robustness check. We fit a variety of commonly used distributions including Gompertz, Weibull, log-logistic, and log-normal distributions, and include a gamma-distributed error term in the hazard function, i.e., gamma mixture. All these models support that workload has an inverted-U-shaped relationship with meal duration.

We obtain support for the inverted-U-shaped relationship between service duration and workload in the OLS model but weaker support in the 2SLS model. This relationship reconciles the competing hypotheses. When the overall workload is low, as previously argued, in response to increasing workload, servers may spend more time and effort selling more items, which accordingly takes time to consume and consequently increases meal duration (behavioral effect I). Secondary as it is, mechanistic effect may also contribute to prolonging meal duration. In addition, although behavioral effect I suggests that servers may be motivated to increase their effort level, this extra effort is likely to be largely devoted to sales instead of speed because of the unique incentives of a low workload (behavioral effect III). When the overall workload is high, however, servers have stronger incentives to work more promptly (behavioral effect III). Furthermore, because of the antiproductive emotions induced by the excessive workload (behavioral effect II), servers may rush the diners by selling fewer items, which should decrease meal duration. The mechanistic processor-sharing effect alone would suggest that meal duration may keep lengthening as workload increases; nevertheless, this effect may be countered by the increased service rate because of servers' promptness and rushing of diners. Therefore, as workload increases, meal duration first increases and then decreases.

5.3.1. Number of Sold Items. As we discussed earlier, servers often face a speed/quality trade-off: achieving high sales/quality takes time. In particular, on one hand, persuading diners to purchase more items takes extra time and effort. On the other hand, servers may "rush" diners by presenting checks without even being asked to do so, thus reducing meal duration and the number of sold items. In addition to this speed/quality trade-off, servers may simply adjust their promptness to influence the meal duration without affecting the number of sold items. Therefore, it remains unclear whether the effect of workload on meal duration results from this speed/quality trade-off or from servers' promptness or both.

Furthermore, servers may influence sales by either cross-selling or up-selling. The parties that purchase the cross-sold items, such as desserts or wines, usually spend more time on a meal than those parties that only consume entrees. To have a better understanding of these factors. We first control for the impact of the number of sold items during a check, i.e., $Items_i$ on sales. In other words, we insert a control variable $Items_i$, which is the number of items sold during check i, into the sales model and we use the 2SLS estimation with the same set of instruments employed in §5.2. It seems reasonable to assume that controlling for $Items_i$ leads to isolating the cross-selling effect and focusing on the up-selling effect. We further adjust for $Items_i$ in the meal duration model. The additional impact of workload on meal duration therefore should be attributed to servers' promptness and up-selling effort. Finally, we estimate the impact of workload on the number of sold items using the a 2SLS strategy and the same set of instruments employed previously to provide evidence of whether or not servers may affect meal duration through their cross-selling efforts.

Table 6 shows the results of the new 2SLS estimations with *Items* as a control variable and with *Items* as a dependent variable. In estimating log(*Sales*) conditioned on the number of items sold, we notice that the coefficient of $AvgTables^2$ is still negative (−0.0705), whereas the coefficient of *AvgTables* is positive (0.071), suggesting that workload has an inverted-U-shaped relationship with servers' up-selling behavior. In estimating log(*MealDuration*) conditioned on the number of items sold, the coefficient of $AvgTables^2$ is still significant and negative (−0.077), which suggests that servers may decelerate as workload increases below the inflection point, and yet accelerate after workload surpasses the threshold. Finally, in estimating *Items*, the coefficient of $AvgTables^2$ is negative (−0.6365), whereas the coefficient of *AvgTables* is positive (0.2705), which suggests that workload also has an inverted-U-shaped relationship with servers' cross-selling effort. In other

Table 6 Number of Sold Items Analysis

	log(*Sales*)	log(*MealDuration*)	Items
AvgTables	0.0710***	0.0456	0.2705*
	(0.0156)	(0.0342)	(0.1262)
$AvgTables^2$	−0.0705**	−0.0770*	−0.6365**
	(0.0243)	(0.0325)	(0.2012)
PartySize	0.1075***	−0.0409***	1.3728***
	(0.0023)	(0.0028)	(0.0257)
StoreItems		−0.0003	
		(0.0002)	
Items	0.0769***	0.0385***	
	(0.0009)	(0.0008)	
Controls	Yes	Yes	Yes
Observations	185,545	185,545	185,545
Prob > Chi-sq	<0.001	<0.001	<0.001

Note. Standard errors are shown in parentheses.
*$p \le 0.05$; **$p \le 0.01$; ***$p \le 0.001$.

words, as workload increases, servers first sell more items, but then sell fewer items as workload continues increasing.

These results suggest that when overall workload is low, increasing workload stimulates servers to redouble their up-selling and cross-selling efforts at the expense of slower service speed. When overall workload is high, however, further increasing workload spurs servers to accelerate their service at the expense of reduced sales efforts. Furthermore, since consuming more items prolongs the meal duration (note that the coefficient of *Items* is positive in estimating log(*MealDuration*)), the inverted-U-shaped relationship between *Items* and workload provides indirect evidence that servers may reduce meal duration, or "rush," by selling fewer items in addition to simply being more prompt. A similar empirical result is found in Batt and Terwiesch (2012), who find that doctors order fewer diagnostic tests to reduce service time under high workload.

5.4. Validity of Instrumental Variables

To confirm the validity of the instruments and ensure asymptotic consistency of instrumental variable estimators, we check both the relevance condition and the exclusion restriction condition.

Table 7 shows the first-stage regression at the check level. The coefficient of *Software* is significant and negative (−0.0601) when *AvgTables* is regressed in the sales model, which suggests that the implementation of the new scheduling software may have increased staffing level and thus reduced average workload. Specifically, the implementation of the software may have decreased the workload by 6%. However, this variable is not significant in the meal duration model, although the coefficient is also negative (−0.0143). Note that *Software* is positively associated with $AvgTables^2$ (coefficient = 0.0889 and 0.1013) in both models because some values of *AvgTables* are negative after mean centering. In addition, as expected, the one-week lagged workload is positively associated with workload in the current week (coefficient = 0.1135 and 0.0468). The quadratic term of the last week is also positively correlated with the quadratic term of the current week (coefficient = 0.0232 and 0.0295).

Table 7 **First-Stage Regressions of *AvgTables* and $AvgTables^2$**

	Sales model		Meal duration model	
	AvgTables	$AvgTables^2$	*AvgTables*	$AvgTables^2$
Software	−0.0601***	0.0889***	−0.0143	0.1013***
	(0.0100)	(0.0234)	(0.0097)	(0.0235)
LWHRTableLoad	0.1135***	0.0418***	0.0468***	0.0238**
	(0.0043)	(0.0077)	(0.0042)	(0.0077)
$LWHRTableLoad^2$	−0.0113**	0.0232**	0.0120***	0.0295***
	(0.0037)	(0.0078)	(0.0036)	(0.0078)
PartySize	−0.0739***	−0.1584***	−0.0954***	−0.1642***
	(0.0023)	(0.0054)	(0.0022)	(0.0055)
StoreItems			0.0076***	0.0021***
			(0.0001)	(0.0001)
Controls	Yes	Yes	Yes	Yes
Observations	186,357	186,357	186,357	186,357
Prob > Chi-sq	<0.001	<0.001	<0.001	<0.001

Note. Standard errors are shown in parentheses.
**$p \le 0.01$; ***$p \le 0.001$.

Although *Software* is not significant when estimating *AvgTables* in the meal duration model, we still choose to keep it in our instrumental variable estimations because (1) the *F*-statistics for the joint significance of the first-stage estimations are all over 10, namely, the suggested rule of thumb of weak instruments (Staiger and Stock 1997), which indicates that our instrumental variables combined are not weak and they should satisfy the relevance condition; and (2) three instruments make the two endogenous variables overidentified, which allows us to use Sargan overidentifying restriction tests to ensure that our instruments satisfy the exclusion restriction assumption (Kennedy 2003).

Unfortunately, there is no generally accepted statistical test for the exclusion restriction assumption. Nevertheless, we conduct Sargan tests of overidentifying restrictions, which are often used to test exogenous instruments. We find that the *p*-values are over 0.5 for both models, and therefore we fail to reject the null hypothesis that the error terms of the structural models are uncorrelated with the instrumental variables. We would also argue that the implementation of the software should affect restaurant performance only through staffing levels, without affecting demand factors or the service quality of individual servers. Moreover, from our industry knowledge and our interviews with restaurant managers, we believe that hourly staffing levels from one week ago should be independent of the contemporaneous shock to meal duration and sales of the current week after controlling for both time-varying and time-invariant effects.

5.5. Robustness Checks

5.5.1. Correlated Errors. The regression models in §5.1 assume that the error terms in the sales and meal duration models are independent from each other. Nevertheless, sales and meal duration may be simultaneously affected by common unobserved exogenous shocks, such as a baseball game in town. These correlated shocks should be largely positive because longer meals are usually associated with larger sales (the correlation is about 0.256).

To allow the errors to be correlated with each other, in addition to addressing the potential endogeneity issues, we propose a system of simultaneous equations using a three-stage least squares (3SLS) estimation method (Zellner and Theil 1962) for the following reasons. First, the 3SLS instrument estimation can provide consistent estimates of the endogenous variables. It is also quite

Table 8 3SLS Estimations on log(*Sales*) **and** log(*MealDuration*)

	log(*Sales*)	log(*MealDuration*)
AvgTables	0.1291***	0.0444
	(0.0090)	(0.0343)
AvgTables^2	−0.1497***	−0.0987**
	(0.0291)	(0.0326)
PartySize	0.2109***	0.0103**
	(0.0040)	(0.0037)
StoreItems		−0.0000
		(0.0002)
Controls	Yes	Yes
Observations	185,545	185,545
Prob > Chi-sq	<0.001	<0.001

Note. Standard errors are shown in parentheses.
**$p \le 0.01$; ***$p \le 0.001$.

robust in the presence of other estimating issues such as multicollinearity. Furthermore, the system of the simultaneous-equations approach utilizes all available information in the estimates and is therefore more efficient than a single-equation approach (Kennedy 2003). We use the same instruments as described in §5.2 and propose the following estimation procedure:

Stage 1. Same as the first stage in the 2SLS approach.

Stage 2. After using the predicted values from Stage 1 to estimate the coefficients of each equation, we use these 2SLS estimates to predict errors in the system of simultaneous equations, i.e., structural equation errors. These predicted errors are further used to compute the contemporaneous variance-covariance matrix of the structural equation's errors.

Stage 3. Compute the general least squares (GLS) estimators of the system of equations.

Table 8 presents the results of the 3SLS estimations. Consistent with the results in Tables 4 and 5, the coefficients of *AvgTables* are significant in the sales model (0.1291), but insignificant in the meal duration model (0.0444). In addition, the coefficients of *AvgTables*^{2} are significant and equal to −0.1497 and −0.0987 in the sales and meal duration models, respectively. These results support the previous findings about the inverted-U-shaped relationships between workload and both sales and meal duration. Interpreting the coefficients of the sales model, we calculate that the optimal workload is about $(0.1291/(2 \times 0.1497) \approx 0.43)$ tables above the sample mean. In addition, changing the current workload to the optimal value would have generated $(0.1291 \times 0.43 - 0.1497 \times 0.43^2 \approx 3\%)$ sales lift per check on average, controlling for party size and other factors.

5.5.2. Hourly Level Analysis and Discussion of Workload Measures. Our main analysis (§5.3) is conducted at the check level because the check-level data provide a granular sample to understand the nuances of servers' behavior. Nevertheless, a priori, an hourly level analysis seems to be most appropriate because restaurants tend to schedule servers on an hourly basis. In this subsection, we aggregate all variables at the hourly level to provide a robustness check of the check-level results and to examine the practical implications of staffing decisions. We show that the hourly level analysis corroborates our check-level findings on servers' behavior under various levels of workload, but it may produce different estimates of the magnitudes of workload effects.

To be comparable to the check-level analysis, we define the hourly level dependent variables in terms of hourly average sales per check and hourly average meal duration. In other words, $HRAvgSales_{tk} = (\sum_{i \in tk} Sales_i)/HRChecks_{tk}$, and $HRAvgMealDuration_{tk} = (\sum_{i \in tk} MealDuration_i)/HRChecks_{tk}$, where i is a check that started in hour t at restaurant k, and $HRChecks_{tk}$ is the total number of checks that started in hour t at restaurant k. Unlike the total sales per hour, hourly average sales per check should be immune to demand truncation because of constrained capacity.

We define the independent variable $HRTableLoad_{tk}$ as the workload during hour t at restaurant k. It is computed as the number of parties who started meals during hour t divided by the number of servers who processed at least one check in the same hour. We provide an alternative definition of workload in terms of the number of diners, namely, $HRDinerLoad_{tk}$. As with the check-level analysis, we center these workload variables and their quadratic terms for interpretation purposes. These measures are commonly used among restaurant managers to decide on staffing levels. In addition, we consider the following control variables. Variable $HRCheck_{tk}$ is used to adjust for demand and to account for the load on the kitchen and other functions in the restaurants. We also include the one-hour lagged workload in terms of tables/diners per server, namely, $LagHRTableLoad_{tk}$ or $LagHRDinerLoad_{tk}$ because high traffic in the previous hour could generate some congestion over the next hour. Finally, we use the same set of time/date/location control variables as in models 3 and 4.

Table 9 shows the summary statistics of hourly variables. On average, each meal lasts approximately 47 minutes, generating sales of $39.13 per check on average. About 11.13 parties start their meals during an average hour. In addition, each restaurant staffs, on average, close to six servers per hour, which results in an hourly workload of 1.85 tables or 4.33 diners per server.

We first specify our models as follows using hourly tables per server, *HRTableLoad*, as a workload measure:

$$\begin{aligned}\log(HRAvgSales_{tk}) &= \alpha_0 + \alpha_1 HRTableLoad_{tk} + \alpha_2 HRTableLoad_{tk}^2 \\ &\quad + \alpha_3 HRChecks_{tk} + \alpha_4 LagHRTableLoad_{tk} \\ &\quad + \alpha_5 Controls_{tk} + \varepsilon_{tk},\end{aligned}$$

Table 9 Summary Statistics of Hourly Variables

	HRAvgMealDuration	*HRAvgSales*	*HRChecks*	Number of servers per hour	*HRTableLoad*	*HRDinerLoad*
N	16,874	16,874	16,874	16,874	16,874	16,874
Mean	47.05	39.13	11.13	5.71	1.85	4.33
SD	8.00	8.26	7.69	3.18	0.64	1.66
Min	21.85	9.98	1	1	0.17	1
P5	34.95	26.59	1	1	1	2
P25	42.01	33.53	4	3	1.33	3
P50	46.72	38.69	10	6	1.80	4.18
P75	51.55	44.32	17	8	2.22	5.38
P95	59.81	52.77	25	11	3	7.22
Max	109.23	96.12	45	18	7	15.50

$$\log(HRAvgMealDuration_{tk}) = \beta_0 + \beta_1 HRTableLoad_{tk} + \beta_2 HRTableLoad_{tk}^2 + \beta_3 HRChecks_{tk} + \beta_4 LagHRTableLoad_{tk} + \beta_5 Controls_{tk} + \xi_{tk},$$

where $Controls_{tk}$ include $DayWeek_{tk}$, $Hour_{tk}$, $YearWeek_{tk}$, and $Store_{tk}$ to adjust for the time/date and location factors. We conduct 3SLS estimation using the same instruments as those used in the check-level analysis, namely, the software implementation, one-week lagged hourly workload in terms of tables per server and its quadratic terms. As an alternative workload measure, we then use hourly diners per server, *HRDinerLoad*, in the following models and follow the same 3SLS estimation using the instruments in terms of diners per server:

$$\log(HRAvgSales_{tk}) = a_0 + a_1 HRDinerLoad_{tk} + a_2 HRDinerLoad_{tk}^2 + a_3 HRChecks_{tk} + a_4 LagHRDinerLoad_{tk} + a_5 Controls_{tk} + \eta_{tk},$$

$$\log(HRAvgMealDuration_{tk}) = \beta_0 + \beta_1 HRDinerLoad_{tk} + \beta_2 HRDinerLoad_{tk}^2 + \beta_3 HRChecks_{tk} + \beta_4 LagHRDinerLoad_{tk} + \beta_5 Controls_{tk} + \vartheta_{tk}.$$

Table 10 shows the hourly analysis results using alternative workload definitions. In estimating

Table 10 Impacts of Hourly Level Workload on log(*HRAvgSales*) **and** log(*HRAvgMealDuration*)

	Table load		Diner load	
	log(*HRAvgSales*)	log(*HRAvgMealDuration*)	log(*HRAvgSales*)	log(*HRAvgMealDuration*)
HRTableLoad	0.5561*	0.2125		
	(0.2781)	(0.1728)		
HRTableLoad^2	−0.3906*	−0.2066*		
	(0.1638)	(0.1018)		
HRChecks	−0.0216	−0.0052	−0.0137*	0.0006
	(0.0133)	(0.0083)	(0.0068)	(0.0052)
LagHRTableLoad	−0.0092	−0.0200***		
	(0.0072)	(0.0045)		
HRDinerLoad			0.1498**	0.0353
			(0.0555)	(0.0426)
HRDinerLoad^2			−0.0412**	−0.0214*
			(0.0139)	(0.0107)
LagHRDinerLoad			−0.0013	−0.0067***
			(0.0025)	(0.0019)
Controls	Yes	Yes	Yes	Yes
Hypothesis supported	H1	H2	H1	H2
Observations	14,768	14,774	14,768	14,774
Prob > Chi-sq	<0.001	<0.001	<0.001	<0.001

Note. Standard errors are shown in parentheses.
*$p \le 0.05$; **$p \le 0.01$; ***$p \le 0.001$.

log(*HRAvgSales*), the coefficients of $HRTableLoad^2$ and $HRDinerLoad^2$ are both significant and negative $(-0.3906, -0.0412)$. The coefficients of *HRTableLoad* and *HRDinerLoad* are both significant and positive $(0.5561, 0.1498)$. These are qualitatively consistent with our check-level results—workload may have an inverted-U-shaped relationship with sales per check, and the optimal workload to maximize sales is greater than the sample mean. Using these estimated coefficients, we compute that the optimal *HRTableLoad* is about 0.71 tables/server above the sample mean (1.84 tables/server), and the optimal *HRDinerLoad* is about 1.81 diners/server above the sample mean (4.3 diners/server). These two optimal points seem to be consistent with each other because 2.6 diners, on average, sit at one table in our sample. In addition, in interpreting the estimated coefficients, we find that the optimal *HRTableLoad* would have increased *HRAvgSales* by $(0.5561 \times 0.71 - 0.3906 \times 0.71^2) \approx 20\%$, and the optimal *HRDinerLoad* would have increased *HRAvgSales* by $(0.1498 \times 1.8 - 0.0412 \times 1.8^2) \approx 13\%$. In estimating log(*HRAvgMealDuration*), the coefficients of $HRTableLoad^2$ and $HRDinerLoad^2$ are both significant and negative $(-0.2066, -0.0214)$, suggesting that workload initially concavely increases and then concavely decreases the average meal duration of each check. Similar to the check-level results, the linear terms of both workload measures are statistically insignificant at the 0.05 level.

We acknowledge that the sales-lift results from hourly sales analysis results are quantitatively different from the check-level results. Above, we estimated that optimal check-level workload in terms of tables/server was 0.43 tables above the sample mean, which would have generated about 3% extra sales. Nevertheless, the optimal hourly level workload is 0.71 tables/server, which would have generated approximately 20% additional sales. We provide three possible explanations.

First, check-level workload mechanically has a higher sample mean than hourly workload because those servers who handle more tables contribute a higher weight to the average check-level workload. For example, suppose we have six checks in an hour and two servers. One of the servers handles four tables, while the other handles only two. The check-level sample mean is $(4 \times 4 + 2 \times 2)/6 \approx 3.33$ tables/server. In contrast, the hourly sample mean is $6/2 = 3$ tables/server. If we assume that the intrinsic optimal workload is approximately the same regardless of the level of analysis (both hour-level and check-level workload measures essentially reflect how many tables one server handles simultaneously), and that it is greater than the sample mean, then the check-level sample mean is closer to the intrinsic optimal workload than the hour-level sample mean. Our empirical results show that the optimal workload is 0.43 tables above the check-level sample mean, and 0.71 tables above the hourly sample mean.

Second, by analyzing hourly average workload, such as *HRTableLoad*, we implicitly assume that all the servers receive the same number of tables in an hour, thus neglecting the workload variation across each check in that table. In other words, the variance of *HRTableLoad* should be smaller than the variance of check-level workload, which includes an extra variability from work assignment across servers. In fact, $\mathrm{Var}(HRTableLoad) \approx 0.42 < \mathrm{Var}(AvgTables) \approx 0.7$. This difference in workload variances may contribute to the fact that the estimated hourly coefficients are *greater* in absolute values than the estimated check-level coefficients, which contributes to a smaller magnitude of sales lift.

Third, servers should have heterogeneous capabilities to handle different levels of workload. As mentioned above, hourly aggregation implicitly assigns the same number of diners to all servers, which is suboptimal for the restaurant. In reality, however, more capable servers may serve more tables than less capable ones, which may self-optimize the sales impact of workload. Therefore, we find a larger sales lift in the hourly analysis than in the check-level analysis.

Although check-level and hourly level results are quantitatively different, they are qualitatively consistent in that (1) as workload increases, both sales and meal duration will first increase and then decrease, and (2) the optimal workload to maximize sales is larger than the sample mean, suggesting that reducing staffing level may contribute to not only a labor cost reduction but also a sales lift.

5.5.3. Alternative Inverted-U-Shaped Hypothesis Testing. The commonly used criterion for identifying an inverted-U-shaped relationship, i.e., the significance of the quadratic term that we used in the main analysis, has been questioned in some recent literature (Lind and Mehlum 2010). This literature argues that the quadratic specification may erroneously create an extreme point even though the true relationship is concave and monotone. We believe that this concern does not necessarily apply to our analysis because our extreme points are close to the sample means. Another concern would be that the quadratic term is limited to the "nonlocal" assumption that implies that the fitted dependent variables, i.e., $\log(\widehat{Sales})$ and $\log(\widehat{MealDuration})$, at a given $AvgTables = AvgTables_0$ depend heavily on *AvgTables* values far from $AvgTables_0$. To provide robustness checks, we test whether or not the slope of the curve is positive at the start and negative at the end of a reasonably chosen interval of the main variable $[X_l, X_h]$, which is often chosen to be $[X_{\min}, X_{\max}]$ (Lind and Mehlum 2010).

Table 11 Alternative Inverted-U-Shaped Hypothesis Testing

	Sales model (model 3)		Meal duration model (model 4)	
	Lower bound	Upper bound	Lower bound	Upper bound
Interval	−1.160	7.490	−1.160	7.490
Slope	0.369	−1.698	0.217	−1.245
t-value	4.329	−3.839	1.985	−2.533
$P > \|t\|$	0.000	0.000	0.024	0.006

To implement this alternative hypothesis testing method, take model 3 for example. We test the following two standard one-sided *t*-tests:

$$H_0^L\colon \alpha_1 + 2\alpha_2 AvgTables_l \le 0 \quad \text{versus}$$
$$H_1^L\colon \alpha_1 + 2\alpha_2 AvgTables_l > 0,$$
$$H_0^H\colon \alpha_1 + 2\alpha_2 AvgTables_h \ge 0 \quad \text{versus}$$
$$H_1^H\colon \alpha_1 + 2\alpha_2 AvgTables_h < 0.$$

The rejection area is therefore

$$R_\alpha = \Big\{(\alpha_1, \alpha_2)\colon (\alpha_1 + 2\alpha_2 AvgTables_l) \cdot \Big(\sqrt{s_{11} + 2 \times 2 AvgTables_l s_{12} + (2AvgTables_l)^2 s_{22}}\Big)^{-1} > t_\alpha \text{ and } (\alpha_1 + 2\alpha_2 AvgTables_l) \cdot \Big(\sqrt{s_{11} + 2 \times 2 AvgTables_h s_{12} + (2AvgTables_h)^2 s_{22}}\Big)^{-1} < -t_\alpha\Big\},$$

where s_{11}, s_{12}, and s_{22} are the 2SLS estimated variances of α_1 and α_2 and the covariance between them, and t_α is the α-level tail probability of the *t*-distribution.

Table 11 shows the hypothesis testing results. In the sales model (model 3), the slope is positive (0.369) at the lower bound and negative (−1.698) at the upper bound. The *p*-values of the *t*-values are both less than 0.001, so we reject the null hypothesis that the relationship between *AvgTables* and *Sales* is either monotone or U-shaped at the 0.001 confidence level. Similarly, in the meal duration model (model 4), the slope is positive (0.217) at the lower bound and negative (−1.245) at the upper bound. The *p*-values of the *t*-values are both less than 0.05, which rejects the null hypothesis that the relationship between *AvgTables* and *MealDuration* is monotone or U-shaped at the 0.05 confidence level.

6. Managerial Insights and Concluding Remarks

6.1. Managerial Insights

Making optimal staffing decisions is critical for restaurants to achieve better performance. Our study underscores several insights for restaurant managers facing the increasing challenges and pressures of managing a complex workforce in a highly demanding work environment. Perhaps the most counterintuitive finding of our study is that *reducing* the staffing level may improve sales and save labor costs—having one's cake and eating it too. We find that the optimal workload using the average sales per check as performance metric is approximately 0.43 tables per server above the current sample mean, controlling for demand. The average tables/server ratio in our sample is currently on average equal to 2.16 tables per server during one check. Our findings indicate that an optimal staffing of 2.59 tables per server would simultaneously increase sales and reduce labor costs. Using the estimates in the 3SLS estimation of log(*Sales*), we project that optimal staffing will directly increase the average sales per check by approximately 3%. In our robustness checks, the hour-level analyses suggest that the optimal workload is 0.71 tables/server or 1.81 diners/server above the sample means, which may increase average hourly sales per check by 20%.

To stay on the conservative side, we advocate the check-level workload measure to estimate the economic impacts of workload. The commonly used hourly workload measure implicitly assumes that workload is distributed evenly across servers, which is rather simplistic and unrealistic. In addition, although the estimated sales lift in check-level analysis is about 3%, much less than the 20% of the hourly analysis, it is still very significant in a high-fixed-cost industry like restaurants. In this type of industry, a 3% increase in sales at no additional cost has a substantial impact on profits, even without accounting for the labor cost reduction resulting from the optimal workload adjustment. Our estimated sales lift is in line with Mani et al. (2011), who estimated that an optimal staffing level could improve average store profitability by 3.8% to 5.9% in a retail setting.

Although the hourly workload measure does not accurately reflect the economic impact of optimal workload, its simplicity is relatively practical for restaurant managers to implement optimal staffing levels. After forecasting demand in terms of tables or diners, managers can update their demand/server ratio to generate new staffing decisions. Using hourly level analysis, we find that over 75% of the time, our focal restaurants tend to over staff by, on average, one server per hour. Reducing the staffing level by one server each hour can save about 17% of current labor costs (the current average hourly staffing level is 5.71 servers). Of course, our model does not allow us to make an entirely accurate estimate of the potential improvement from optimal staffing (e.g., further labor-related nonwage costs), nor can the restaurants perfectly forecast demand. We nevertheless anticipate a significant sales lift and cost saving from optimal staffing because of the benefits

from correcting both understaffing and overstaffing errors.

Firms nowadays have access to big data, such as new human resource management software, which allows them to analyze the impact of workload at a more granular level. The new software is also capable of monitoring the workload of servers in real time, which facilitates the acceptance of more detailed managerial implication. Our check-level workload measure provides a first step in utilizing big operational data to understand the impact of workload.

6.2. Concluding Remarks

Most studies on staffing decisions in services tend to overlook employees' adaptive behavior to work environments. There is a general lack of empirical research about the impact of staffing on workers' performance. A growing stream of literature has documented that workers adjust their performance in response to work environments. In particular, prior research has focused on the impact of workload, an integral environmental factor, on either service time or quality, separately. Little observational research has (1) explicitly emphasized the value of empirical research for staffing decisions and (2) examined how workload affects service workers, who make joint speed/quality decisions.

In this paper, we utilize detailed operational data gathered from a restaurant chain to study the effects of workload on servers' performance in terms of both sales and meal duration, taking endogeneity into consideration. We find that, when the overall workload is low, increasing the workload may motivate servers to generate more sales. When the workload is high, increasing the workload may reduce servers' effective sales. We also find that, as workload increases, meal duration first increases and then decreases. Because of this inverted-U-shaped relationship between workload and sales, we demonstrate that reducing the number of waiters in those restaurants whose current average workload is below the optimum may *both* significantly increase sales and reduce labor costs, which is against the traditional workforce management presumption.

Our empirical findings contribute to the existing analytical models on staffing in three ways. First, our research shows the value of empirical research for making staffing decisions. Second, the nonlinearity of the impact of workload on meal duration enriches the analytical research on staffing that considers workload-dependent productivity. Hasija et al. (2010) have written an important and timely paper on the linear speeding-up behavior induced by workload to estimate a call center's capacity. De Véricourt and Jennings (2011) also explicitly model the workload of nurses to determine efficient nurse staffing policies. Future research may further assume nonlinear productivity induced by workload and use our estimates. Third, our finding further provides empirical evidence to support existing research studying the effect of workload on service speed and quality (see, e.g., Hopp et al. 2007, Debo et al. 2008, Anand et al. 2011). Higher sales not only benefit the restaurant's bottom line but also may arguably reflect higher service quality. Understanding the trade-off between productivity and quality induced by workload may strengthen the analytical models on staffing.

The drivers of workload effects are initially unclear. On one hand, a high workload may indicate high demand, which will increase hourly performance. On the other hand, a high workload may indicate understaffing, which may result in overloaded servers and diminished performance. Through instrumental variables, we show that optimal staffing decisions, i.e., supply factors, mainly drive the results of our analysis. In particular, optimal staffing can improve sales generation and save labor costs. Moreover, we explain that, when overall workload is low, increasing workload stimulates servers to redouble both their up-selling and cross-selling efforts at the expense of slower service speed. When overall workload is high, however, further increasing workload spurs servers to accelerate their service at the expense of reduced sales efforts. Since consuming more items prolongs the meal duration, our results also provide indirect evidence that a server may reduce meal duration, or "rush," by selling fewer items in addition to simply being more prompt. A similar empirical result is found in Batt and Terwiesch (2012), who find that doctors order fewer diagnostic tests to reduce service time.

It is important to take into account the limitations of our findings. Although our data set is among the largest in the existing literature on worker performance response to external factors, it misses a few interesting variables. For example, we do not observe the exact duration of each service procedure, such as taking the order and settling the bill. An interesting avenue for future research would be to examine the impact of workload on each specific service procedure and how servers switch their service from table to table (see Bendoly et al. 2013 for some initial work in this direction). In addition, we lacked data about complete tipping information because we only observed tips paid through credit cards. We analyzed tip data that was available to us and found that tips showed very little variation (as a percentage of the check); therefore, we did not find a robust impact of workload on tips. However, other types of customer satisfaction data, such as customer surveys, would be desirable to study the impact of workload on guest satisfaction. Furthermore, because of data limitations, our study does not examine the impact of other factors, such as kitchen capacity and diner heterogeneity. Although we employed instrumental variables to address this omitted variable issue, these factors would be worth

studying in future research. Additionally, our data only shows the number of servers who handled checks, which should cause a downward bias relative to actual staffing decisions. Nevertheless, as we find that the restaurant is already overstaffed, including more precise information in this case would only strengthen our findings. Further research opportunities in this setting include studying other OM/human resources interface issues, such as the "chemistry" among team members and team composition. Using our findings about servers' adaptive behavior to environmental constrains to design new workforce scheduling algorithms would also offer an interesting and fruitful direction. Finally, in our models, to separate the supply-side driver of workload effect, we assume exogenous demand, namely, the number of diners starting service every hour. In practice, arriving diners may choose to enter the restaurant or leave depending on its occupancy. For example, when a restaurant is too empty, diners may interpret it as a sign of low restaurant quality, thus deciding to leave. However, when the restaurant is too full, diners may anticipate a long wait, thus balking at the door. It would be interesting to empirically test how occupancy affects demand.

Acknowledgments

The authors sincerely acknowledge the entire review team for their most diligent and constructive comments. In particular, the authors thank one reviewer for encouraging a granular analysis of the "speed/quality" conundrum. The authors are also indebted to the reviewers for encouraging the strengthening of the theoretical development. Furthermore, the authors are grateful to the faculty members and the Ph.D. students at the Wharton School; the seminar participants at the University of Notre Dame, Boston College, the University of South Carolina, Southern Methodist University, and the National University of Singapore; and to the INFORMS Annual Conference participants for their valuable suggestions. In addition, the authors thank the INSEAD–Wharton Alliance for their financial support throughout this project. Finally, the research would not be possible without the generous data provider, Objective Logistics, and their cofounders, Philip H. Beauregard and Matthew Grace.

References

Afifi AA, Clark V, May S (2004) *Computer-Aided Multivariate Analysis* (CRC Press, Boca Raton, FL).

Akşin OZ, Harker PT (2001) Modeling a phone center: Analysis of a multichannel, multiresource processor shared loss system. *Management Sci.* 47(2):324–336.

Akşin Z, Armony M, Mehrotra V (2007) The modern call center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* 16(6):665–688.

Alizamir S, de Véricourt F, Sun P (2013) Diagnostic accuracy under congestion. *Management Sci.* 59(1):157–171.

Anand KS, Paç MF, Veeraraghavan S (2011) Quality–speed conundrum: Trade-offs in customer-intensive services. *Management Sci.* 57(1):40–56.

Angrist J, Krueger AB (1994) Why do World War II veterans earn more than nonveterans? *J. Labor Econom.* 12(1):74–97.

Arlotto A, Chick SE, Gans N (2014) Optimal hiring and retention policies for heterogeneous workers who learn. *Management Sci.* 60(1):110–129.

Batt RJ, Terwiesch C (2012) Doctors under load: An empirical study of state-dependent service times in emergency care. Working paper, University of Wisconsin–Madison, Madison.

Bendoly E (2011) Linking task conditions to physiology and judgment errors in RM systems. *Production Oper. Management* 20(6):860–876.

Bendoly E, Hur D (2007) Bipolarity in reactions to operational "constraints": OM bugs under an OB lens. *J. Oper. Management* 25(1):1–13.

Bendoly E, Prietula M (2008) In the "zone": The role of evolving skill and transitional workload on motivation and realized performance in operational tasks. *Internat. J. Oper. Production Management* 28(12):1130–1152.

Bendoly E, Donohue K, Schultz KL (2006) Behavior in operations management: Assessing recent findings and revisiting old assumptions. *J. Oper. Management* 24(6):737–752.

Bendoly E, Swink M, Simpson WP III (2013) Prioritizing and monitoring concurrent project work: Effects on switching behavior. *Production Oper. Management*, ePub ahead of print September 24, http://onlinelibrary.wiley.com/doi/10.1111/poms.12083/abstract.

Bloom N, Van Reenen J (2007) Measuring and explaining management practices across firms and countries. *Quart. J. Econom.* 122(4): 1351–1408.

Boudreau JW (2004) Organizational behavior, strategy, performance, and design in management science. *Management Sci.* 50(11): 1463–1476.

Boudreau JW, Hopp W, McClain JO, Thomas LJ (2003) On the interface between operations and human resources management. *Manufacturing Service Oper. Management* 5(3):179–202.

Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L (2005) Statistical analysis of a telephone call center: A queuing-science perspective. *J. Amer. Statist. Assoc.* 100(469):36–50.

Cakir A, Hart DJ, Stewart TFM (1980) *Visual Display Terminals: A Manual Covering Ergonomics, Workplace Design, Health and Safety, Task Organization* (John Wiley & Sons, Hoboken, NJ).

Dalton AN, Spiller SA (2012) Too much of a good thing: The benefits of implementation intentions depend on the number of goals. *J. Consumer Res.* 39(3):600–614.

de Véricourt F, Jennings OB (2011) Nurse staffing in medical units: A queueing perspective. *Oper. Res.* 59(6):1320–1331.

Debo LG, Toktay LB, Van Wassenhove LN (2008) Queuing for expert services. *Management Sci.* 54(8):1497–1512.

Deci EL, Connell JP, Ryan RM (1989) Self-determination in a work organization. *J. Appl. Psych.* 74(4):580–590.

Donahue EM, Robins RW, Roberts BW, John OP (1993) The divided self: Concurrent and longitudinal effects of psychological adjustment and social roles on self-concept differentiation. *J. Personality Soc. Psych.* 64(5):834–846.

Fields R (2007) *Restaurant Success by the Numbers* (Ten Speed Press, Berkeley, CA).

Fisher ML, Krishnan J, Netessine S (2006) Retail store execution: An empirical study. Working paper, The Wharton School, University of Pennsylvania, Philadelphia.

Fitzsimmons JA, Maurer GB (1991) A walk-through audit to improve restaurant performance. *The Cornell Hotel and Restaurant Admin. Quart.* 31(4):94–99.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* 5(2):79–141.

Hasija S, Pinker E, Shumsky RA (2010) OM Practice—Work expands to fill the time available: Capacity estimation and staffing under Parkinson's Law. *Manufacturing Service Oper. Management* 12(1):1–18.

Hopp WJ, Iravani SMR, Yuen GY (2007) Operations systems with discretionary task completion. *Management Sci.* 53(1):61–77.

Huckman RS, Staats BR, Upton DM (2009) Team familiarity, role experience, and performance: Evidence from Indian software services. *Management Sci.* 55(1):85–100.

KC DS, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Sci.* 55(9):1486–1498.

KC DS, Terwiesch C (2012) An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing Service Oper. Management* 14(1):50–65.

Kennedy P (2003) *A Guide to Econometrics* (MIT Press, Cambridge, MA).

Kimes SE (2004) Restaurant revenue management: Implementation at Chevys Arrowhead. *Cornell Hotel and Restaurant Admin. Quart.* 45(1):52–67.

Kimes SE, Robson SKA (2004) The impact of restaurant table characteristics on meal duration and spending. *Cornell Hotel and Restaurant Admin. Quart.* 45(4):333–346.

Kimes SE, Thompson GM (2004) Restaurant revenue management at Chevys: Determining the best table mix. *Decision Sci.* 35(3): 371–392.

Kimes SE, Barrash DI, Alexander JE (1999) Developing a restaurant revenue-management strategy. *Cornell Hotel and Restaurant Admin. Quart.* 40(5):18–29.

Kimes SE, Chase RB, Choi S, Lee PY, Ngonzi EN (1998) Restaurant revenue management: Applying yield management to the restaurant industry. *The Cornell Hotel and Restaurant Admin. Quart.* 39(3):32–39.

Kleinbaum DG, Kupper LL, Muller KE (2007) *Applied Regression Analysis and Other Multivariable Methods* (Duxbury Press, Pacific Grove, CA).

Kleinrock L (1976) *Queueing Systems: Volume 2: Computer Applications*, Vol. 82 (John Wiley & Sons, Hoboken, NJ).

Kostami V, Rajagopalan S (2014) Speed–quality trade-offs in a dynamic model. *Manufacturing Service Oper. Management* 16(1):104–118.

Kuntz L, Mennicken R, Scholtes S (2014) Stress on the ward: Evidence of safety tipping points in hospitals. *Management Sci.* Forthcoming.

Latham GP, Locke EA (1979) Goal setting: A motivational technique that works. *Organ. Dynam.* 8(2):68–80.

Lind JT, Mehlum H (2010) With or without U? The appropriate test for a U-shaped relationship. *Oxford Bull. Econom. Statist.* 72(1):109–118.

Locke EA (1968) Toward a theory of task motivation and incentives. *Organ. Behav. Human Performance* 3(2):157–189.

Luo J, Zhang J (2013) Staffing and control of instant messaging contact centers. *Oper. Res.* 61(2):328–343.

Lupien SJ, Maheu F, Tu M, Fiocco A, Schramek TE (2007) The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition. *Brain and Cognition* 65(3):209–237.

Maher K (2007) Wal-mart seeks flexibility in worker shifts. *Wall Street Journal* (January 3), http://online.wsj.com/news/articles/SB116779472314165646.

Mani V, Kesavan S, Swaminathan JM (2011) Understaffing in retail stores: Drivers and consequences. Working paper, Pennsylvania State University, University Park.

Mill RC (2006) *Restaurant Management: Customers, Operations, and Employees*, 3rd ed. (Prentice Hall, Upper Saddle River, NJ).

O'Connor EJ, Peters LH, Pooyan A, Weekley J, Frank B, Erenkrantz B (1984) Situational constraint effects on performance, affective reactions, and turnover: A field replication and extension. *J. Appl. Psych.* 69(4):663–672.

Oliva R, Sterman JD (2001) Cutting corners and working overtime: Quality erosion in the service industry. *Management Sci.* 47(7): 894–914.

Perdikaki O, Kesavan S, Swaminathan JM (2012) Effect of traffic on sales and conversion rates of retail stores. *Manufacturing Service Oper. Management* 14(1):145–162.

Peters LH, O'Connor EJ (1980) Situational constraints and work outcomes: The influences of a frequently overlooked construct. *Acad. Management Rev.* 5(3):391–397.

Powell A, Savin S, Savva N (2012) Physician workload and hospital reimbursement: Overworked physicians generate less revenue per patient. *Manufacturing Service Oper. Management* 14(4): 512–528.

Robson SKA (1999) Turning the tables: The psychology of high volume restaurant design. *Cornell Hotel and Restaurant Admin. Quart.* 40(3):56–63.

Schultz KL, Juran DC, Boudreau JW (1999) The effects of low inventory on the development of productivity norms. *Management Sci.* 45(12):1664–1678.

Schultz KL, Juran DC, Boudreau JW, McClain JO, Thomas LJ (1998) Modeling and worker motivation in JIT production systems. *Management Sci.* 44(12):1595–1607.

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

Siebert WS, Zubanov N (2010) Management economics in a large retail company. *Management Sci.* 56(8):1398–1414.

Staats BR, Gino F (2012) Specialization and variety in repetitive tasks: Evidence from a Japanese bank. *Management Sci.* 58(6): 1141–1159.

Staiger D, Stock JH (1997) Instrumental variables regression with weak instruments. *Econometrica* 65(3):557–586.

Sulek JM, Hensley RL (2004) The relative importance of food, atmosphere, and fairness of wait: The case of a full-service restaurant. *Cornell Hotel and Restaurant Admin. Quart.* 45(3): 235–247.

Villas-Boas JM, Winer RS (1999) Endogeneity in brand choice models. *Management Sci.* 45(10):1324–1338.

White H (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48(4):817–838.

Whitt W (2006) The impact of increased employee retention on performance in a customer contact center. *Manufacturing Service Oper. Management* 8(3):235–252.

Wooldridge JM (2002) *Econometric Analysis of Cross Section and Panel Data* (The MIT Press, Cambridge, MA).

Zellner A, Theil H (1962) Three-stage least squares: Simultaneous estimation of simultaneous equations. *Econometrica* 30(1): 54–78.