2013

The Implications of Worker Behavior for Staffing Decisions: Empirical Evidence and Best Practices

Tom Tan
Southern Methodist University, ttan@cox.smu.edu

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

The Implications of Worker Behavior for Staffing Decisions: Empirical Evidence and Best Practices

Tom Fangyun Tan

Cox Business School, Southern Methodist University, Dallas, Texas, U.S.A.

ttan@cox.smu.edu

Serguei Netessine

INSEAD, Singapore

serguei.netessine@insead.edu

Abstract

Employees in service organizations such as restaurants are humans, and their productivity is often heterogeneous due to innate abilities, motivation, and environmental factors. However, none of these considerations are taken into account by typical workforce planning models and software packages. In this paper we describe the results of a study on how workers respond to workload, an integral work environment factor, and we provide operational insights to improve efficiency and strengthen restaurant financials. We study a comprehensive data set from a family-style restaurant chain to clarify how workload, which is defined as the number of tables or diners that a server concurrently handles, affects his/her performance (measured in terms of sales and meal duration). Among our findings, servers's sales increase with the workload at a cost of longer meal duration, when the overall workload is low. However, their sales performance decreases and they work more expeditiously after a certain workload threshold. We find that this workload threshold is currently not reached in our focal restaurants and, counter-intuitively, the chain can reduce its staffing level to achieve both significantly higher sales (an estimated 3% increase) and lower labor costs (an estimated 17% decrease). We further discuss how companies can benefit by explicitly considering other aspects of employee behavior when scheduling workforce.

Keywords: restaurant operations; quality/speed trade-off; server performance; employee productivity

Introduction and Related Literature

Labor is typically one of the largest cost components of service organizations such as call centers and restaurants. For example, in the USA alone, the restaurant industry employs about 13 million workers pear year, approximately 10% of the total workforce (Mill, 2004). As managers know, labor decisions, such as staffing, are critical in driving operational performance and strengthening financials, and computerized staffing tools (Maher, 2007) have been implemented by many service companies to maximize their performance. However, most of these systems calculate employee productivity using "grand averages" of historical data, overlooking employees' adaptive behavior towards changing work environments (Brown et al., 2005). This can cause serious staffing errors.

The simplified view of constant worker productivity is inherited from classical operations management (OM) models that often assume that worker performance is independent from the state of the system, or at best that performance has random variations (see Boudreau et al., 2003 and Bendoly et al., 2006 for comprehensive reviews). Recent studies have started to bridge OM models and human resource management in order to more accurately understand the impact of external factors on individuals' performance (e.g., Schultz et al., 1998; Boudreau et al., 2003; Bendoly and Prietula, 2008; Bendoly, 2011). Indeed, some researchers have recently started to study the impact of workload, an critical environmental factor, on individual performance, and often use healthcare services as a test bed (Kc and Terwiesch, 2009, 2011; Powell et al., 2012).

Although examining the impact of workload on worker performance has generated considerable recent research interest, these studies predominantly focus on either service time or service quality, separately. However, delineating the effects of workload on both service time and quality is of great significance in the service industry because service providers strive to concurrently maximize service quality, which is related to sales and customer satisfaction, and minimize service time, which is associated with opportunity costs. These service providers are nevertheless constrained by capacity and often encounter a quality/speed trade-off because delivering high service quality takes more time. Therefore, there remains a need to understand how workers make trade-offs under various workplace pressures.

A growing number of papers are making use of analytical modeling approaches to better understand how workers make such trade-off decisions (Hopp et al., 2007; Debo et al., 2008; Anand

et al., 2011; Alizamir et al., 2013). There are, however, to the best of our knowledge, no empirical studies on this speed/quality trade-off.

In this paper, we describe the results of Tan and Netessine (2013), where we examine how workload affects service speed (as reflected in the length of service) and quality (as reflected in the sales amount) decisions, using a set of unique and very comprehensive transaction-level data from a sit-down casual restaurant chain's point-of-sales system that contains approximately 190,000 check-level observations for five restaurants from August 2010 to June 2011. In this setting, servers have discretionary powers to control sales and speed of service. For example, to expend sales effort, servers may chat with diners and anticipate their needs by suggesting dishes and drinks without appearing aggressive (Fitzsimmons and Maurer, 1991). Servers' suggestions for appetizers, soup and wine are known to stimulate demand that would otherwise be unexpressed. To devote effort in speed, servers may carry multiple items from the kitchen to save trips and time. They also need to be aware of 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, because of their capacity constraints, they have to make a trade-off between speed and sales.

How servers decide between sales and speed efforts is an interesting and complicated question. According to a study by the National Restaurant Association (Mill, 2004), 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 or a server's impatience with answering menu questions. In addition, sales are of great importance to restaurants which, on average, generate very small pre-tax profit margins, averaging just 4%. In order to increase sales, servers are usually instructed and trained to sell more items and to sell more expensive items. Therefore, learning servers' behavior towards sales and speed is critical for improving restaurant service operation. While 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), little research has directly shown that staff workload has a major impact on revenue generation.

We demonstrate how staffing capacity can be leveraged to optimize the workload. We find

that servers react non-linearly 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 sales effort with the increase in workload at a cost of slower service speed. However, above a certain threshold (around 2.59 tables per server) servers start to reduce their sales efforts and work more promptly with a further rise in workload. On average, the restaurant chain in our study allocates 2.16 tables per server. Thus, we conclude that our focal restaurants are largely overstaffed and 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. We further describe how this and other behavioral aspects of employee productivity can be used to increase restaurant revenues.

Hypotheses Development

Effect on Sales

While the diner's preferences for items on the menu is the key factor that determines the sales per check, these preferences can be influenced by servers via suggestive selling efforts which increase restaurant sales (Fitzsimmons and Maurer, 1991). Such efforts include up-selling low-price menu items and cross-selling items that diners would otherwise not order. Research shows that diners are more likely to purchase a dessert or after-meal drinks if a server makes such a suggestion when clearing the plates of the main course. In addition, servers' suggestions for appetizers, soup, wine and high-margin items are also known to stimulate demand that would otherwise be unexpressed. Thus, when the initial workload is low, an increasing workload may motivate servers to exert a more suggestive selling effort. Indeed, cognitive psychology suggests that workload may trigger the cerebral cortex to release hormones that improve cognitive performance (Lupien et al., 2007). According to Parkinson's Law (Parkinson, 1958), employees tend to fill idle time with irrelevant activities, such as smoking outside, chatting with each other and folding napkins, creating inefficiency as opposed to selling more items. Hence, increasing the workload may reduce servers' idle time, thus increasing their selling effort. However, when workload surpasses a critical level and becomes too high, it is likely to limit sales per check: servers may become so occupied with carry-

ing food that they have no time to conduct suggestive selling. In addition, they may be distracted by other diners, reducing their sales service effectiveness. Fatigue caused by heavy workload may also lead to reduced effort (Cakir et al., 1980; Setyawati, 1995). Servers may even suffer from "contact overload", the emotional drain from handling too many customers over a prolonged period of time (Mill, 2004). The result of contact overload can be emotional burnout, which reduces a server's sales efforts and effectiveness. For these reasons we hypothesize that:

HYPOTHESIS 1 (H1): *As workload increases, hourly sales will first increase and then decrease.*

Effect of Servers on Meal Duration

Naturally, diners' speed of eating primarily determines the meal duration. Nevertheless, one would expect a server's effort and attitude to significantly affect meal duration too: for example, an efficient server will quickly present menu and later the bill to expedite the order and check-settlement procedures. Occasionally, a server may implicitly rush diners by presenting the check without being asked for it. She/he may choose to be quick when transporting food from the kitchen to the table. A diligent server will be swifter to answer diners' requests. Furthermore, a server may prolong meal duration by offering more menu items such as wine.

We argue that, when the workload is low, a higher level of workload will prolong meal duration. Operationally, when a server serves more diners, his/her attention is divided into smaller portions because of process sharing. Consequently, he/she may not address diners' needs promptly, thus extending meal duration. For example, diner i may need some assistance from his/her server, who is busy serving other diners. Therefore, diner i has to wait to get the server's attention. Furthermore, workload can be seen as a challenge and therefore a motivation stimulus (Deci et al., 1989). As workload increases, motivation also increases, which is shown to improve effort (Locke et al., 1978; Yeo and Neal, 2004). The server may be more motivated to make recommendations and suggest additional menu items. As a result, diners order extra food, which extends the meal.

However, when workload becomes too high, a higher level of workload may encourage servers to speed up. One reason is that servers may want to reduce the costs of customer waiting (e.g., waiting to settle the check) by accelerating service. Kc and Terwiesch(2009; 2011) find empirical evidence in the hospital setting that the higher workload reduces the service time. Moreover,

when servers are overworked, they may cut corners, thus reducing service time (Oliva and Sterman, 2001). From a psychological perspective, a too high workload may cause servers to become frustrated. Consequently, they may rush diners by presenting the check without being asked. Similarly, Brown et al. (2005) found that call center agents intentionally hung up on callers to reduce their workload and obtain extra rest time. Based on the arguments above, we propose an inverted U-shaped relationship between workload and meal duration.

HYPOTHESIS 2 (H2): *As workload increases, meal duration first increases and then decreases.*

Data

Research Setting and Data Collection

To test our research hypotheses, we collaborated with Objective Logistics, a restaurant-focused software company that provides an artificial-intelligence based labor management platform called MUSE. The software creates forecast of demand and schedules servers to meet the demand. As a part of the software implementation process, we collected 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. The restaurants are open from 11:30 am to 10:00 pm from Monday to Thursday, and from 11:30 am to 11:00 pm from Friday to Sunday. Diners include couples, families, students and their friends. The restaurants have a full-service bar and 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 would require different operationalization of variables. Our data contains 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. In order to avoid spurious extrapolations due to outliers (e.g., very large parties and private events or unobservable discounted meals), we drop the transactions which include the day's top and bottom 7.5% of checks. This threshold is relatively small and is in line with standard econometric approaches (Kennedy, 2003), which should not significantly influence our analysis. Our final data set includes approximately 190,000 check-level observations.

Measures and Controls

In order to carefully understand how workload affects servers' behavior in 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 aggregate all variables at the hourly level to provide a robustness check and to consider implications for staffing decisions. 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. We operationalize dependent variables $Sales_i$ and $MealDuration_i$ to reflect the sales and the length of a check i, which is only assigned to one server in our focal restaurants. Note that we infer the meal duration of each check from check opening and closing times recorded in our POS data, which is 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 concurrently 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 min + 20 min)/(40 min) = 1.5 tables. Weighting the workload by 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 not substantially affect check i[1]. We use diners per server as an alternative workload measure and the results are qualitatively the same.

In addition, we consider the following control variables to account for confounding factors. Variable $PartySize_i$ controls for 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 store-wide number of items ordered at the beginning and at the end of check i, which is used to adjust for the workload on the kitchen, which can potentially affect the meal duration. Finally, we also control for the time/date/location of check i. Night hours usually generate more sales than lunch hours, so we include a 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, $DayWeek_i$.

[1]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 2011 measured the ICU occupancy at the time of a patient's discharge). These alternative measures yielded qualitatively congruent results.

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. In order 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 also control for effects specific to the store using the categorical variable $Store_i$. These time and store fixed effects also indirectly adjust for the capacity of the kitchen, whose staff tends to be salaried and have fixed work schedules.

Table 1 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 their means, so transforming them is recommended to increase normality prior to model estimation (Afifi et al., 2004). We further center *AvgTables* and $AvgTables^2$ around their means for interpretation purposes.

Table 2 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 low, suggesting that the predictors should not cause the multicollinearity issue in the model estimation.

Results[2]

We adopt a system of simultaneous equations using a three-stage least squares (3SLS) estimation method, which allows us to find the causal impact of workload on servers' perofrmance (Zellner and Theil, 1962). This method is appropriate for our study for two reasons. First, 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). Secondly, and more importantly, the 3SLS instrument estimation can provide unbiased and consistent estimates of

[2]This section is based on Subsection 5.3 of Tan and Netessine (2013).

AvgTables and $AvgTables^2$, which can correct the simultaneity bias resulted from managers' adjusting the server staffing level to match projected demand, and the omitted variable bias caused by not observing other drivers of sales and meal duration. As an integral part of the 3SLS estimation technique, we propose two types of instrumental variables, namely the implementation of a new staffing system and one-week lagged independent variables, which we find to be valid from statistical validity tests (more details about our instruments are available upon request).

In addition, we use the quadratic specification of $AvgTables_i$, which allows us to compute the critical points in the regression models.

Table 3 shows the results of check-level analyses. First, in the log(*Sales*) model, the coefficient of $AvgTables^2$ is negative (-0.1497) and significant at a 0.001 level, supporting our Hypothesis 1. This result suggests that variable *AvgTables* first concavely increases sales and then concavely decreases sales. In the same column, the coefficient of *AvgTables* is 0.1291 and significant at a 0.001 level. Interpreting the coefficients from the log(*Sales*) model, we find that the optimal workload is about $(0.1291/(2 \times 0.1497) \approx 0.43)$ tables above the sample mean, which is 2.16 tables. In addition, changing the current workload to the optimal value would have generated ($0.1291 \times 0.43 - 0.1497 \times 0.43^2$ ≈3%) sales lift per check on average, controlling for party size and other factors. Furthermore, as expected, a larger party size is positively associated with higher sales per check (coefficient is 0.2109).

In the log(*MealDuration*) model, the coefficient of $AvgTables^2$ is also negative (-0.0987) and significant at a 0.01 level, suggesting that *AvgTables* initially concavely increases the meal duration of each check and then concavely decreases the meal duration, consistent with Hypothesis 2.

In addition to the 3SLS model described so far, we conduct a series of robustness checks. First, we conduct duration model analysis of log(*MealDuration*) with 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. Furthermore, we include server fixed effects to control for the heterogeneous skills of servers. All these models support that workload has an inverted-U-shaped relationship with meal duration.

Factors of the Inverted-U Shaped Relationships

As discussed before, servers intend to maximize sales/quality in the shortest amount of time. Nevertheless, they face a speed/quality trade-off: achieving high sales quality takes time. 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. To have a better understanding of these factors of the inverted-U shaped relationships, we first control the impact of the number of sold items during a check, namely $Items_i$, on meal duration. The additional impact of workload on meal duration therefore should be attributed to servers' promptness. We further examine the impact of the number of sold items on sales to provide insights about the marginal effects of cross-selling and up-selling activities. It seems reasonable to assume that controlling for $Items_i$ leads to isolating the cross-selling effect. Finally, we estimate the impact of workload on the number of items sold using a 3SLS strategy to provide evidence of whether or not servers may affect meal duration through their cross-selling efforts.

Table 4 presents the results of factors of the inverted-U shaped relationships. In estimating $\log(MealDuration)$ conditioned on the number of items sold, the coefficient of $AvgTables^2$ is still significant and negative (-0.0718), which suggests that servers may become slower as workload increases below the inflection point, and yet become faster after workload surpasses the threshold. In estimating $\log(Sales)$ conditioned on the number of items sold, we notice that the coefficient of $AvgTables^2$ is still significant and negative (-0.067), while the coefficient of $AvgTables$ is significant and positive (0.049), suggesting that workload has an inverted-U shaped relationship with servers' up-selling behavior. Interpreting the coefficients, we find that the inflection point is about 0.36 tables above the sample mean, which is slightly below 0.43 tables (Table 3), the inflection point of the combined sales effect, which consists of both up-selling and cross-selling efforts. We further compute that the up-selling effort contributes to about $((0.049 \times 0.43 - 0.067 \times 0.43^2)/3\% \approx 29\%)$ of the total sales lift from the optimal workload (0.43 tables). Finally, in estimating $Items$, the coefficients of $AvgTables^2$ is significant and negative (-1.089), while the coefficient of $AvgTables$ is significant and positive (1.041), which suggests that workload also has an inverted-U shaped relationship with servers' cross-selling effort. In other words, as workload increases, servers first sell more items, but then, as workload continues increasing, they sell fewer items.

These results suggest that under the light workload, increasing the workload stimulates servers to increase their up-selling and cross-selling efforts at the expense of slower service speed. Under the heavy workload, however, further intensifying workload prompt servers to accelerate their service at the expense of reduced sales efforts. Furthermore, since consuming more items prolongs 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 described in Batt and Terwiesch (2012), who find that doctors order fewer diagnostic tests to reduce service time.

Hour-level Analysis

Restaurants generally 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. In order to be comparable to the check-level analyses, we define the hour-level dependent variables in terms of hourly average sales per check and hourly average meal duration. In other words, $HRAvgSales_{tk} = \frac{\sum_{i \in tk} Sales_i}{HRChecks_{tk}}$, and $HRAvgMealDuration_{tk} = \frac{\sum_{i \in tk} MealDuration_i}{HRChecks_{tk}}$, 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.

We define the independent variable $HRTableLoad_{tk}$ as the workload during hour t at restaurant k. We calculate this workload 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. An alternative definition of workload is provided in terms of the number of diners, namely $HRDinerLoad_{tk}$. As with the check-level analysis, these workload variables and their quadratic terms are centered for interpretation purposes. These measures are commonly used by among restaurant managers to decide on staffing levels. In addition, we consider the following control variables. We use variable $HRCheck_{tk}$ to control for demand and to account for the load on the kitchen and other functions in the restaurants. The one-hour lagged workload in terms of tables/diners per server, namely $LagHRTableLoad_{tk}$ or $LagHRDinerLoad_{tk}$ is also included because high traffic in the previous hour could generate some congestion over the next hour.

Table 5 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 apply 3SLS estimation using the same instruments as those used in the check-level analysis, i.e., the software implementation, one-week lagged hourly workload in terms of tables per server and its quadratic terms. We then analyze the impact of hourly diners per server, *HRDinerLoad* to provide an alternative workload measure.

Table 6 shows the hourly analysis results using alternative workload definitions. In estimating 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 hour-level results qualitatively coincide 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 higher 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 match each other because 2.6 diners on average sit at one table in our sample. Furthermore, we interpret that the optimal *HRTableLoad* would have increased *HRAvgSales* by $(0.5561 \times 0.71 - 0.3906 \times 0.71^2) \approx 20\%$, while the optimal *HRDinerLoad* would have increased *HRAvgSales* by $(0.1498 \times 1.8 - 0.0412 \times 1.8^2) \approx 13\%$. For log(*HRAvgMealDuration*) estimation, the coefficients of $HRTableLoad^2$ and $HRDinerLoad^2$ are both significant and negative (-0.2066, -0.0214), suggesting that workload first concavely increases and then concavely decreases the average meal duration of each check. Consistent with the check-level results, the linear terms of both workload measures are statistically insignificant at the 0.05 level.

Hour-level and Check-level Workload Measures

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 hour-level workload is 0.71 tables/server, which would have generated approximately 20% additional sales.

We provide two possible explanations. First, the implicit assumption of analyzing hourly average workload, such as *HRTableLoad*, is that all the servers receive the same number of tables in an hour, which neglects the workload variation across each check in that table. Particularly, 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, Var(*HRTableLoad*) $\approx$ 0.42 $<$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.

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

While check-level and hour-level results are quantitatively different, they qualitatively coincide 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.

Discussion

Managerial Insights

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. Making optimal staffing decisions is critical for restaurants to achieve better performance. Perhaps the most counter-intuitive 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.

To stay on the conservative side, we advocate the check-level workload measure to estimate the economic impacts of workload. For comparison, 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 an industry with high fixed costs 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 hour-level analysis, we find that over 75% of the time, our focal restaurants tend to be over-staffed 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 non-wage 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 under-staffing and over-staffing errors.

Concluding Remarks

Nowadays firms have access to big data, which enables them to analyze the impact of workload at a more granular level. The new software that collects the data is also capable of monitoring the workload of servers in real time, which facilitates the acceptance of more comprehensive managerial implication. Our check-level workload measure provides a first step in utilizing big operational data to understand the impact of workload. We utilize detailed operational data collected from a restaurant chain to study the effects of workload on servers' performance in terms of both

sales and meal duration and we explain the value of empirical analysis on staffing decisions. The key take-away is graphically shown in Figure 1, with the predicted values of sales on the left and the predicted values of meal duration on the right. 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. Due to this inverted-U shaped relationship between workload and sales, we demonstrate that reducing the number of servers in those restaurants whose current average workload is below the optimum may *both* significantly increase sales and reduce labor costs.

The drivers of workload effects are initially unclear but 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.

Further research opportunities in this setting include studying other OM/Human Resources interface issues. In particular, outside this study we observed that servers have widely heterogeneous skill levels. We fitted a fixed-effect model of servers' sales abilities and predicted servers' intrinsic sales values, which are shown in Figure 2. Based on such observations, companies like Objective Logistics experiment with a restaurant scheduling system that ranks workers by their abilities (e.g., based on sales, customer satisfaction or any other attribute that is deemed important by the restaurant chain) and assigns shifts according to this ranking so that best workers are first to pick their preferred shifts to work. Figure 3 demonstrates a snapshot of such a ranking system. The system also allows restaurant management to set corresponding weights for different performance indicators. Such a system makes it possible, through data analytics, to schedule employees more accurately while understanding the individual capabilities of each server. Understanding how to take advantage of the heterogeneity of servers and designing a scheduling algorithm based on the servers' abilities to maximize their potentials would offer an interesting and fruitful direction. We are in the process of data analysis based on experiments with imple-

mentation of MUSE software. Clearly, usability of such software is not limited to the restaurant industry but it can be applied in any service-intensive environment where the individual capabilities of the employees play a major role including in call centers or retail, see Netessine and Yakubovich (2012).

Table 1: 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
Stdev	15.69	16.23	0.83	0.87	36.02
Min	7.88	21.84	1	1	2
Median	37.45	43.69	2.05	2	78.5
Max	131.75	113.59	9.65	5	261.5

Table 2: 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.

Table 3: Impact of Check-level Workload *AvgTables* on log(*Sales*)

	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
Hypothesis Supported	H1	H2
Observations	185,545	185,545
Prob>Chi-sq	<0.001	<0.001

1. Standard errors are shown in parentheses.

2. *: p-value$\leq$ 0.05, **: p-value$\leq$ 0.01, ***: p-value$\leq$ 0.001

Table 4: Factors of the Inverted-U Shaped Relationships

	log(*MealDuration*)	log(*Sales*)	*Items*
AvgTables	0.0317	0.0490***	1.0416***
	(0.0336)	(0.0073)	(0.0636)
$AvgTables^2$	-0.0718*	-0.0670**	-1.0888***
	(0.0314)	(0.0224)	(0.2061)
PartySize	-0.0418***	0.1060***	1.3549***
	(0.0027)	(0.0023)	(0.0284)
StoreItems	-0.0002		
	(0.0002)		
Items	0.0385***	0.0773***	
	(0.0008)	(0.0007)	
Controls	Yes	Yes	Yes
Observations	185,545	185,545	185,545
Prob>Chi-sq	<0.001	<0.001	<0.001

1. Standard errors are shown in parentheses.

2. *: p-value$\leq$ 0.05, **: p-value$\leq$ 0.01, ***: p-value$\leq$ 0.001

Table 5: Summary Statistics of Hourly Variables

	HRAvgMealDuration	*HRAvgSales*	*HRChecks*	Number of Servers per Hour	*HRTableLoad*	*HRDineroad*
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
Stdev	8.00	8.26	7.69	3.18	0.64	1.66
Min	21.85	9.98	1	1	0.17	1
Median	46.72	38.69	10	6	1.80	4.18
Max	109.23	96.12	45	18	7	15.50

Table 6: Impact of hour-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	14768	14774	14768	14774
Prob>Chi-sq	<0.001	<0.001	<0.001	<0.001

1. Standard errors are shown in parentheses.

2. *: p-value$\leq$ 0.05, **: p-value$\leq$ 0.01, ***: p-value$\leq$ 0.001

Figure 1: Summary Plots of Predicted Check-level 3SLS Model Results

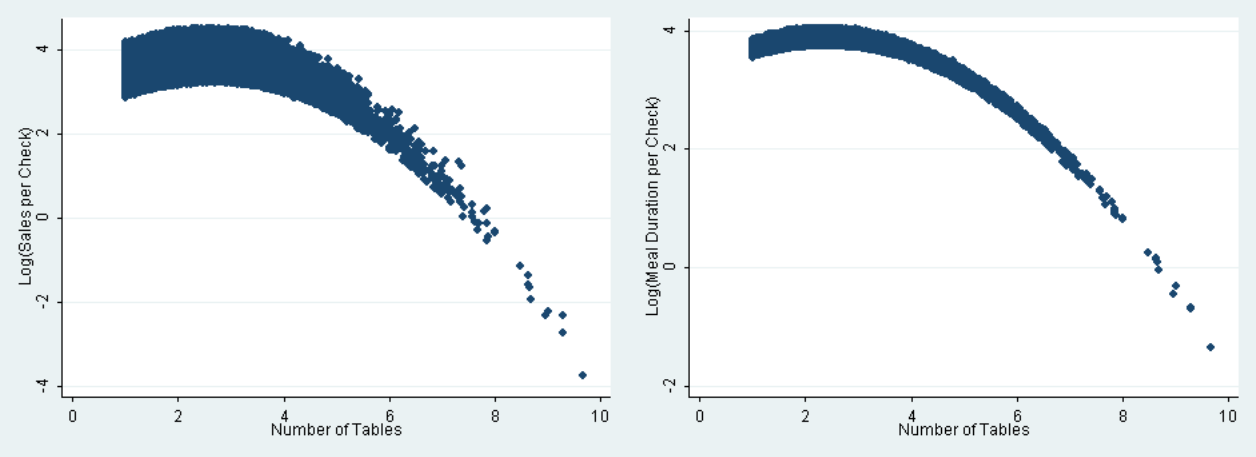


Figure 2: Heterogeneous Sales Skills

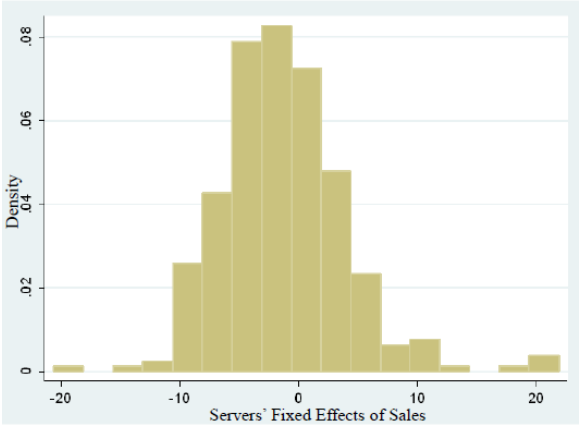
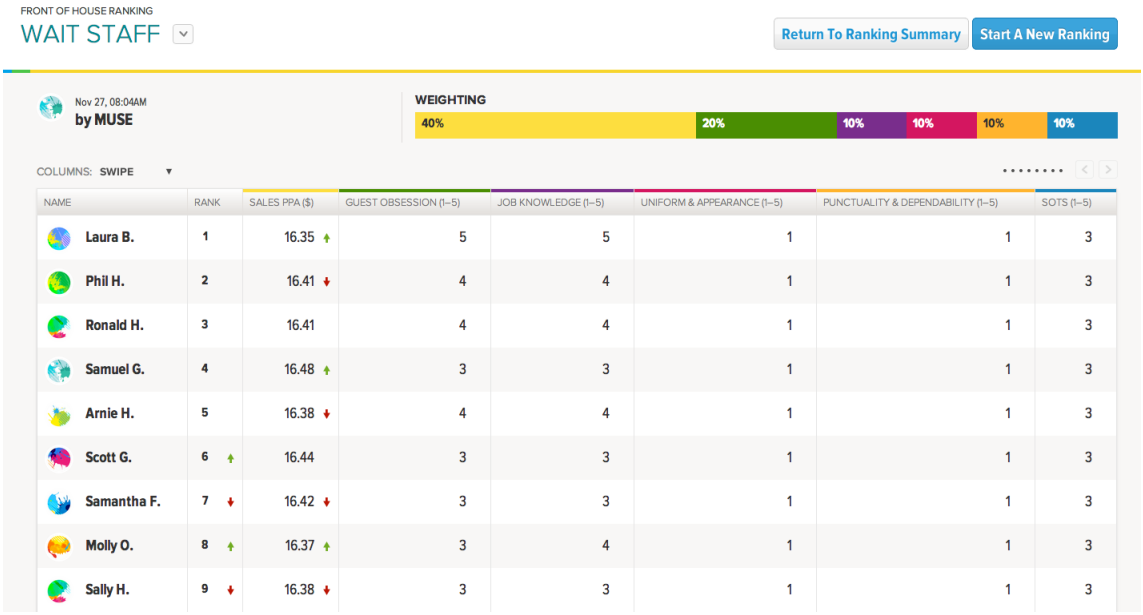


Figure 3: Ranking Servers' Heterogeneous Abilities



References

Afifi, A.A., V. Clark, S. May. 2004. *Computer-aided multivariate analysis*. CRC Press.

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

Anand, K.S., M.F. Paç, S. Veeraraghavan. 2011. Quality-speed conundrum: Tradeoffs in customer-intensive services. *Management Science* **57**(1) 40–56.

Batt, R.J., C. Terwiesch. 2012. Doctors under load: An empirical study of state-dependent service times in emergency care. Wharton Working Paper.

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

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

Bendoly, E., M. Prietula. 2008. In the 'Zone': The role of evolving skill and transitional workload on motivation and realized performance in operational tasks. *International Journal of Operations & Production Management* **28**(12) 1130–1152.

Boudreau, J.W., W. Hopp, J.O. McClain, L.J. Thomas. 2003. On the interface between operations and human resources management. *Manufacturing & Service Operations Management* **5**(3) 179–202.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: A Queuing-science perspective. *Journal of the American Statistical Association* **100**(469) 36–50.

Cakir, A., DJ Hart, TFM Stewart. 1980. *Visual display terminals: A manual covering ergonomics, workplace design, health and safety, task organization*. Wiley.

Debo, L. G., L.B. Toktay, L.N. Van Wassenhove. 2008. Queuing for expert services. *Management Science* **54**(8) 1497–1512.

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

Fitzsimmons, J.A., G.B. Maurer. 1991. A walk-through audit to improve restaurant performance. *The Cornell Hotel and Restaurant Administration Quarterly* **31**(4) 94–99.

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

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

Kc, D.S., C. Terwiesch. 2011. An econometric analysis of patient flows in the cardiac ICU. *Manufacturing & Service Operations Management* **14**(1) 50–65.

Kennedy, P. 2003. *A guide to econometrics*. The MIT Press.

Kimes, S.E. 2004. Restaurant revenue management: Implementation at Chevys Arrowhead. *Cornell Hotel and Restaurant Administration Quarterly* **45**(1) 52–67.

Kimes, S.E., D.I. Barrash, J.E. Alexander. 1999. Developing a restaurant revenue-management strategy. *Cornell Hotel and Restaurant Administration Quarterly* **40**(5) 18–29.

Kimes, S.E., R.B. Chase, S. Choi, P.Y. Lee, E.N. Ngonzi. 1998. Restaurant revenue management: Applying yield management to the restaurant industry. *The Cornell Hotel and Restaurant Administration Quarterly* **39**(3) 32–39.

Kimes, S.E., S.K.A. Robson. 2004. The impact of restaurant table characteristics on meal duration and spending. *Cornell Hotel and Restaurant Administration Quarterly* **45**(4) 333–346.

Kimes, S.E., G.M. Thompson. 2004. Restaurant revenue management at Chevys: Determining the best table mix. *Decision Sciences* **35**(3) 371–392.

Kleinbaum, D.G., L.L. Kupper, K.E. Muller. 2007. *Applied regression analysis and other multivariable methods*. Duxbury Pr.

Locke, E.A., A.J. Mento, B.L. Katcher. 1978. The interaction of ability and motivation in performance: An exploration of the meaning of moderators1. *Personnel Psychology* **31**(2) 269–280.

Lupien, S.J., F. Maheu, M. Tu, A. Fiocco, T.E. Schramek. 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. *The Wall Street Journal* January 3rd.

Mani, V., S. Kesavan, J.M. Swaminathan. 2011. Understaffing in retail stores: Drivers and consequences. Pennsylvania State University Working Paper.

Mill, R.C. 2004. Restaurant management: Customers, operations, and employees .

Netessine, S., V. Yakubovich. 2012. The darwinian workplace. *Harvard Business Review* (May) 25–28.

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

Parkinson, C. 1958. Parkinsons Law: The pursuit of progress.

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

Robson, S.K.A. 1999. Turning the tables: The psychology of high volume restaurant design. *Cornell Hotel and Restaurant Administration Quarterly* **40**(3) 56–63.

Schultz, K.L., D.C. Juran, J.W. Boudreau, J.O. McClain, L.J. Thomas. 1998. Modeling and worker motivation in JIT production systems. *Management Science* **44**(12) 1595–1607.

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

Sulek, J.M., R.L. Hensley. 2004. The relative importance of food, atmosphere, and fairness of wait: The case of a full-service restaurant. *Cornell Hotel and Restaurant Administration Quarterly* **45**(3) 235–247.

Tan, F., S. Netessine. 2013. When does the devil make work? an empirical study of the impact of workload on worker productivity. *Southern Methodist University Working Paper* .

Yeo, G.B., A. Neal. 2004. A multilevel analysis of effort, practice, and performance: Effects; of ability, conscientiousness, and goal orientation. *Journal of Applied Psychology* **89**(2) 231.

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